UCLA
UCLA Electronic Theses and Dissertations

Title
Studying Nucleosome Positioning, DNA Methylation, and Other Epigenetic Factors Using Genomic Approaches

Permalink
https://escholarship.org/uc/item/5qg1h83p

Author
Chodavarapu, Ramakrishna Kant

Publication Date
2012

Supplemental Material
https://escholarship.org/uc/item/5qg1h83p#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Studying Nucleosome Positioning, DNA Methylation, and Other Epigenetic Factors Using Genomic Approaches

A dissertation submitted in satisfaction of the requirements for the degree Doctor of Philosophy in Molecular, Cell and Developmental Biology

by

Ramakrishna Kant Chodavarapu

2012

ABSTRACT OF THE DISSERTATION

Studying Nucleosome Positioning, DNA Methylation, and Other Epigenetic Factors Using Genomic Approaches

By

Ramakrishna Kant Chodavarapu

Doctor of Philosophy in Molecular, Cell and Developmental Biology

University of California, Los Angeles, 2012

Professor Matteo Pellegrini, Chair

Epigenomics is the genome-wide study of the factors that cause changes to gene expression, but are not directly related to the primary DNA sequence. Among these factors are DNA methylation, nucleosome positioning, and histone modifications. Two epigenomic projects are described here. The first describes a ten base periodicity of DNA methylation and an enrichment of DNA methylation over nucleosomal DNA. It also describes nucleosomes being enriched over exons and being placed preferentially at intron-exon boundaries. The other describes an integrative study of BS-seq, RNA-seq, and siRNA-seq data in rice hybrids. SNPs were found between two closely related subspeicies of rice at a rate of 1/253 bp and the frequency of DNA methylation differences between the ecotypes was much higher at nearly 8%. DNA methylation differences between the parents and the hybrids were much more muted, at nearly 1%. It also describes the impact of the DNA methylation differences on differential expression of genes between the parents and the hybrids and the two ecotypes. The supplementary tables referenced here have been included separately. The dissertation includes a brief discussion on the main methods employed in these two studies as well.

This dissertation of Ramakrishna K. Chodavarapu is approved.

Steve Jacobsen

Jason Ernst

Amander T. Clark

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2012

Dedication

To the many who made this possible: to my family, to my friends, and to my mentors.

Contents

ABSTRACT OF THE DISSERTATION ii
Dedication iv
List of Figures and Tables vi
Acknowledgements vii
Vita viii
Publications and Presentations ix
Chapter 1: Introduction 1
DNA Methylation 3
Nucleosomes 5
References 9
Chapter 2: Methods 15
BS-seq analysis 16
MNase-seq analysis 18
References 21
Chapter 3: Relationship between nucleosome positioning and DNA methylation. 22
Relationship between nucleosome positioning and DNA methylation. 24
Method Summary 32
Figure Legends 34
References 35
Supplementary Methods 41
Supplementary Figure Legends 44
Chapter 4: Transcriptome and methylome interactions in rice hybrids. 73
Abstract 76
Introduction 77
Results and Discussion 78
Conclusion 85
Materials and Methods 87
Acknowledgements 90
References 91
Figure Legends 93
Chapter 5: Concluding Remarks 101

List of Figures and Tables

Chapter 3:

Figure 1 37
Figure 2 38
Figure 3 39
Figure 4 40

Supplementary Figure 1 49
Supplementary Figure 2 50
Supplementary Figure 3 51
Supplementary Figure 4 52
Supplementary Figure 5 53
Supplementary Figure 6 54
Supplementary Figure 7 55
Supplementary Figure 8 56
Supplementary Figure 9 57
Supplementary Figure 10 58
Supplementary Figure 11 59
Supplementary Figure 12 60
Supplementary Figure 13 61
Supplementary Figure 14 62
Supplementary Figure 15 63
Supplementary Figure 16 64
Supplementary Figure 17 65
Supplementary Figure 18 66
Supplementary Figure 19 67
Supplementary Figure 20 68
Supplementary Figure 21 69
Supplementary Figure 22 70
Supplementary Figure 23 71
Supplementary Figure 24 72

Chapter 4:

Figure 1 96
Figure 2 97
Figure 3 98

Supplementary Figure 1 100

Acknowledgements

I want to thank the members of the Pellegrini and Jacobsen laboratories for their invaluable guidance and discussions, especially Suhua Feng, Pao-Yang Chen, and Shawn Cokus. I would also like to thank all of the members of my committee, especially Steve Jacobsen and Matteo Pellegrini. I would also like to thank Sabeeha Merchant for her thoughts and guidance. Finally I would like to acknowledge my primary funding source, the Genetics Training Grant (GM07104) for their support, as well as other funding sources used for my projects.

Vita

2006-2007	Undergraduate Research, SRP
2007	B.S. Molecular, Cell, Developmental Biology B.A. Business Economics University of California, Los Angeles
2008-2010	Teaching assistant, MCDB 167A: Cell Biology Teaching assistant, LS 4: Genetics Teaching assistant, MCDB 172: Bioinformatics University of California, Los Angeles
2009-2012	Genetics Training Grant University of California, Los Angeles

Publications and Presentations

Ramakrishna K. Chodavarapu, Suhua Feng, Bo Ding, Stacey A. Simon, David Lopez, Yulin Jia, Guo-Liang Wang, Blake C. Meyers, Steven E. Jacobsen, and Matteo Pellegrini. (2012) Transcriptome and methylome interactions in rice hybrids. *Proc. Nat. Acad. Sci. U. S. A.*, 109: 12040-12045.

Ramakrishna K. Chodavarapu, Suhua Feng, Yana V. Bernatavichute, Pao-Yang Chen, Hume Stroud, Yanchun Yu, Jonathan Hetzel, Frank Kuo, Jin Kim, Shawn J. Cokus, David Casero, Maria Bernal, Peter Huijser, Amander T. Clark, Ute Krämer, Sabeeha S. Merchant, Xiaoyu Zhang, Steven E. Jacobsen, and Matteo Pellegrini. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature,* 466:388-392.

Poster presentation at the Education Academy of Computational Life Science, organized by Tokyo Institute of Technology (Sept. 5 2012)

Research talk at the UCLA MCDB Departmental Retreat (November 13-15, 2009)

Research talk at the Genetics Training Grant Symposium (June 22, 2010)

Chapter 1: Introduction

Despite having identical genetic material, disparate cell types are able to achieve distinct gene expression profiles and maintain their cellular identity through repeated cell divisions. Partly, this is due to the array of genetic tools at the cell's disposal including combinatorial regulation through transcription factors and signaling pathways. However, one of the keys is 'epigenetics': the stable and heritable mechanisms that act primarily by either disrupting, restricting, or fostering access to the underlying genetic material [1, 2].

Chromatin is the protein-DNA superstructure whose fundamental building blocks are nucleosomes. Nucleosomes are in turn made up of a histone octamer wrapped by 147 bp of DNA [3]. Each tail of the histone proteins contains several residues which can be post-translationally modified by mono-, di-, or tri-methylation, amongst a large list of other modifications. Based on the dynamic interactions of these histone tail modifications, the nearby chromatin can then switch between a transcriptionally active 'open' chromatin form, and a transcriptionally silent 'closed' form [4, 5]. Insofar as the structure of the DNA can be modified, and its accessibility to the RNA polymerase complex or to any number of other DNA binding proteins and enzymes be restricted or allowed without changing the primary DNA sequence, these changes are considered epigenetic factors governing changes in gene expression [6].

Epigenetic marks are usually enzyme-mediated chemical modifications to DNA, as in the case of DNA methylation, or to proteins that help regulate the chromatin superstructure, as in the case of post-translational modifications to histone tails [7]. Nucleosome positioning itself can be considered an epigenetic mark, as nucleosome position can restrict the ability of other proteins and enzymes to bind or recognize DNA as well. Nucleosomes can be packed tightly together such that little linker DNA exists as is the case in heterochromatin and transcriptionally silent regions, or they can be loosely packed, allowing for longer linker regions and the transcription machinery to gently pull them off one at a time [8].

Epigenetic processes have been known to act cooperatively. Histone deacetylases (HDACs) and Histone methyltransferases (HMTs) act cooperatively with DNA methyltransferases and nucleosome remodeling factors (NURFs) in order to establish a heritable repressive state at transcription start sites. This is important in development and differentiation, and may also be important in cancer development [9].

The complex machinery involved in initiating, maintaining, and modifying these marks ensures that epigenetic marks are tightly regulated: that transposons are heavily methylated and silenced, that genes are in open chromatin and can be read, that nucleosomes protect the exonic regions, etc. Moreover, unlike genetic mutations where all succeeding cells and generations must carry the mutation, epigenetic changes allow for a much more transient, gradual, and heterogeneous control of gene expression.

DNA Methylation

DNA methylation is a common mechanism by which a cytosine nucleotide is chemically modified by the addition of a methyl group by a DNA methyltransferase [10]. Often, heavy DNA methylation leads to the repression of gene expression, and is therefore commonly found amongst transposable elements and is significantly enriched over heterochromatin [11, 12]. Indeed, it has been shown that reducing DNA methylation leads to the mobilization of transposable elements [13]. DNA methylation at regulatory sites can also lead to gene silencing, although moderate methylation over coding regions does not seem to directly affect transcription.

In plants, DNA methylation occurs in three contexts, and is regulated in a context specific manner by different enzymes [10]. The CG context is symmetric across the DNA strands, which is thought to provide a hemi-methylated template for the maintenance of methylation at these sites after DNA replication, is maintained by the DNA methyltransferase

MET1 [11, 14]. The CHG context, where H is any nucleotide but G, is likewise symmetric across the two strands, however it is maintained by a different methyltransferase, CMT3 [15]. Studies have shown that CMT3 has a chromodomain, a domain that binds specifically to dimethylated H3K9, and is thought to be the link between H3K9me2 methylation and CHG DNA methylation [16, 17]. CHH methylation, which is not symmetric and thus cannot be copied from the opposite DNA strand, is mediated by an RNA directed DNA methylation pathway that targets *de novo* methylation by DNA methyltransferase DRM2 using siRNAs presented by AGO4 [18, 19]. That each context is controlled by distinct genetic pathways in Arabidopsis is underscored by the fact that each context tends to have a distinct distribution of methylation: CG sites tend to be either unmethylated or highly methylated, while methylated CHG sites show a uniform distribution between 20-100% methylation, and methylated CHH sites tend to be methylated at a rate below 10%. This also explains the vast difference in average methylation levels: ~25% for CG sites, ~7% for CHG sites, and ~2% for CHH sites [11]. Although the overall percentages are different, the trends are consistent in other plants [20].

Despite a lack of methylation in the model organisms, *S cerevisae* and *C elegans*, DNA methylation and even DNA methyltransferases are conserved across most eukaryotic organisms [7, 20, 21]. Unlike in plants, however, mammalian DNA methylation is primarily limited to the CG context. Moreover, most CGs tend to be fully methylated, resulting in very high levels of CG methylation [20, 22]. A significant exception to this are CpG islands, which tend to remain unmethylated. In mammals, CG methylation is maintained by the maintenance DNA methyltransferase, DNMT1. However, mammals do have *de novo* methyltransferases, DNMT 3a and 3b, although they are primarily active in initiating non-CG methylation in embryonic stem cells. Importantly, in cancer cells, DNMT1 levels are decreased, but DNMT3a levels are increased leading to hyper-methylation at specific sites [9, 23].

Methylation at multiple genomic features appears to be conserved across plants and animals. Transposons and other repetitive DNA are heavily methylated in many organisms, including Arabidopsis and rice. Methylation profiles over genes also appear to be conserved, with a general depletion immediately prior to, and immediately following genes, and an increase over the gene body [20]. Several factors help in explaining this. First, the depletion immediately upstream and downstream of genes may be due to regulatory regions being bound by transcription factors, preventing access to DNA methyltransferases. In mammals, CpG islands sometimes precede the gene, and as noted, are generally unmethylated. Second, DNA methylation is increased over the exon, and this likely accounts for the increase over the gene body.

Studies of hybrid vigor, the phenomenon where the offspring of two related but different species demonstrate superior performance, have revealed altered transcription levels amongst many genes in hybrids. It has been thought that the magnitude of hybrid vigor is correlated with variation between the parental genomes. However, it has been shown that crosses between genetically similar parents also produce significant hybrid vigor, suggesting that genetics is only part of the story [24-26]. DNA methylation has been shown to be faithfully inherited in offspring, although less so than genetic material. It has also been shown that in the case of hybrids, methylation patterns are largely inherited faithfully, except in regions where the parents themselves have divergent patterns and may be one of the causes of the hybrid vigor phenomenon [27, 28].

Nucleosomes

Nucleosomes lead to the compaction of DNA into chromatin. Approximately 147 bases of DNA are wrapped around a histone octamer, forming the basic DNA packaging unit - a nucleosome [3, 29]. The histone core is composed of two hetero-tetramers of H2A, H2B, H3, and H4 histones. Nucleosomes are typically separated by linker DNA that can range in size

from 10 to 50 bases. DNA methylation and histone tail modifications have been established as important epigenetic mechanisms for transcriptional regulation [30]. However, nucleosome positioning itself likely also plays a role in epigenetic regulation, as the position of nucleosomes can render chromatin more or less accessible to transcription [8, 31].

The position of an individual nucleosome is likely to be determined by a number of factors including DNA sequence, DNA conformation, as well as the position of neighboring nucleosomes. By using traditional sequencing approaches, multiple studies have found that nucleosome-bound DNA is enriched for sequences containing 10 base periodicities in AA, TT, and TA dinucleotides [32-34]. Similar results have also been found using high-throughput sequencing approaches [35-38]. These dinucleotide preferences are likely due to the fact that nucleosome-DNA interactions require the DNA to be strongly bent and that certain base steps are more amenable to bending [39]. This dinucleotide periodicity has been hypothesized to be part of a "nucleosome code" that determines nucleosome positioning [34]. However, while such a sequence pattern may be enriched in nucleosome-bound DNA, it appears to be neither necessary nor sufficient to completely determine nucleosome positions [8].

Previous studies of nucleosome localization demonstrated that nucleosome positions near the transcription start sites are often well defined. In Saccharomyces cerevisiae, there is a well-positioned nucleosome 300 to 150 bases upstream of the TSS (the -1 nucleosome) and another well-positioned nucleosome immediately following the TSS (+1 nucleosome) [35, 40]. In between these two nucleosomes, there is a relatively nucleosome-free region (NFR). It has been hypothesized that the two well-positioned nucleosomes flanking the TSS may play a role in the recruitment of the transcriptional machinery, thus defining the position of the TSS [8, 35]. Evidence suggests that TSS flanking nucleosomes become displaced from the TSS prior to transcription because they provide steric hindrance to RNA polymerase and other transcriptional machinery [41, 42]. Studies in S. cerevisiae and Drosophila melanogaster have also reported

an NFR preceding the transcription terminus [36, 38]. Recent work has shown that these NFR regions may instead be regions that contain highly unstable nucleosomes that contain H3.3 and H2AZ histone variants [43].

Runs of A/T nucleotides have been shown to be favored at sites where the minor groove faces toward the histone core, while runs of G/C nucleotides have been shown to be favored at sites where the minor groove faces away from the histone core [33]. The out of phase peaks in the frequencies of these dinucleotides leads to optimal bending of DNA, as A/T nucleotides cause a negative base roll and G/C nucleotides cause a positive base roll [39].

Since methyltransferases access the major groove, there may be opportunities for methyltransferases to methylate DNA that is still bound to nucleosomes. It has been proposed that chromatin remodeling enzymes are required for DNA methyltransferases to gain access to the DNA, and indeed DRD1, DDM1 and LSH1 are remodelers known to be important factors controlling DNA methylation [44]. A 10-nucleotide periodicity in CHH methylation data was reported when an autocorrelation analysis on the methylation pattern of the whole genome was performed [11]. One interpretation of this was that the structure of the DRM2 enzyme might be responsible for this pattern, since the orthologous enzymes in mammals are known to act as heteromeric complexes in which two methyltransferase active sites show a spacing equivalent to roughly 10 nucleotides of DNA [45]. However, a more general explanation is that nucleosomes are to some extent dictating access to the DNA and therefore setting the register of methylation for all DNA methyltransferases. In addition, some evidence suggests that DNA methylation may influence nucleosome positioning, because DNA methylation in the major groove is antagonistic to the compression needed for sharp DNA bending, disfavoring methylated DNA in the major groove facing the nucleosome surface [31]. This could suggest a co-evolution of DNA methylation patterning and nucleosomal positioning in the genome.

Nucleosomal preferences could also partially explain the strong sequence preferences that was previously observed for CHG and CHH methylation [11].

In the case of CMT3, it is predicted that this enzyme is recruited or activated by histone H3 lysine 9 dimethylation, because a knockout of histone H3 lysine 9 methyltransferase mimics a knockout of CMT3, and because the CMT3 chromodomain can bind to methylated histones [17]. Furthermore, previous structural modeling of CMT3 suggested that the chromodomain is on the opposite face of the molecule as the DNA methyltransferase catalytic domain [46]. Thus CMT3 may be positioned such that the beginning and ends of nucleosome bound DNA are more susceptible to methylation.

Previous theoretical studies of intron-exon boundaries have generated models suggesting that nucleosomes should be enriched at splice sites [47, 48]. These studies hypothesized that the position of nucleosomes would be centered on the splice site, with half the nucleosome-bound DNA in the intron and the other half in the exon. It is becoming increasingly clear that splicing is coupled with transcription [49]. Indeed, the CTD of RNAPII and chromatin elongation factors help recruit splicing factors to the nascent transcript [50, 51]. A slow RNAPII processivity could help define splice sites, as it reduces the likelihood of a stronger downstream splice site being transcribed [52]. Recently, it has been shown in C. elegans that there is increased H3K36me3 at exons relative to introns, and at exons that are constitutively expressed relative to alternatively spliced exons [53]. Other recent analyses suggested similar trends, and Pol II was found to be enriched on human exons relative to introns [54-56]. These results suggest that the phenomenon of nucleosome enrichment on exons is widely conserved throughout eukaryotic organisms.

References

1. Margueron, R. and D. Reinberg, *Chromatin structure and the inheritance of epigenetic information.* Nat Rev Genet, 2010. **11**(4): p. 285-96.

2. Bird, A., *Perceptions of epigenetics.* Nature, 2007. **447**(7143): p. 396-8.

3. Kornberg, R.D. and Y. Lorch, *Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome.* Cell, 1999. **98**(3): p. 285-94.

4. Campos, E.I. and D. Reinberg, *Histones: annotating chromatin.* Annu Rev Genet, 2009. **43**: p. 559-99.

5. Bartova, E., et al., *Histone modifications and nuclear architecture: a review.* J Histochem Cytochem, 2008. **56**(8): p. 711-21.

6. Croft, J.A., et al., *Differences in the localization and morphology of chromosomes in the human nucleus.* J Cell Biol, 1999. **145**(6): p. 1119-31.

7. Feng, S., S.E. Jacobsen, and W. Reik, *Epigenetic reprogramming in plant and animal development.* Science, 2010. **330**(6004): p. 622-7.

8. Jiang, C. and B.F. Pugh, *Nucleosome positioning and gene regulation: advances through genomics.* Nat Rev Genet, 2009. **10**(3): p. 161-72.

9. Jones, P.A. and S.B. Baylin, *The epigenomics of cancer.* Cell, 2007. **128**(4): p. 683-92.

10. Chan, S.W., I.R. Henderson, and S.E. Jacobsen, *Gardening the genome: DNA methylation in Arabidopsis thaliana.* Nat Rev Genet, 2005. **6**(5): p. 351-60.

11. Cokus, S.J., et al., *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008. **452**(7184): p. 215-9.

12. Zilberman, D., et al., *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.* Nat Genet, 2007. **39**(1): p. 61-9.

13. Miura, A., et al., *Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis.* Nature, 2001. **411**(6834): p. 212-4.

14. Vongs, A., et al., *Arabidopsis thaliana DNA methylation mutants.* Science, 1993. **260**(5116): p. 1926-8.

15. Lindroth, A.M., et al., *Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation.* Science, 2001. **292**(5524): p. 2077-80.

16. Cedar, H. and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms.* Nat Rev Genet, 2009. **10**(5): p. 295-304.

17. Lindroth, A.M., et al., *Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3.* EMBO J, 2004. **23**(21): p. 4286-96.

18. Wassenegger, M., et al., *RNA-directed de novo methylation of genomic sequences in plants.* Cell, 1994. **76**(3): p. 567-76.

19. Henderson, I.R. and S.E. Jacobsen, *Epigenetic inheritance in plants.* Nature, 2007. **447**(7143): p. 418-24.

20. Feng, S., et al., *Conservation and divergence of methylation patterning in plants and animals.* Proc Natl Acad Sci U S A.

21. Law, J.A. and S.E. Jacobsen, *Establishing, maintaining and modifying DNA methylation patterns in plants and animals.* Nat Rev Genet, 2010. **11**(3): p. 204-20.

22. Ehrlich, M., et al., *Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells.* Nucleic Acids Res, 1982. **10**(8): p. 2709-21.

23. Jones, P.A., *DNA methylation and cancer.* Oncogene, 2002. **21**(35): p. 5358-60.

24. Birchler, J.A., D.L. Auger, and N.C. Riddle, *In search of the molecular basis of heterosis.* Plant Cell, 2003. **15**(10): p. 2236-9.

25. Birchler, J.A., et al., *Heterosis.* Plant Cell, 2010. **22**(7): p. 2105-12.

26. Springer, N.M. and R.M. Stupar, *Allelic variation and heterosis in maize: how do two halves make more than a whole?* Genome Res, 2007. **17**(3): p. 264-75.

27. Becker, C., et al., *Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.* Nature, 2011. **480**(7376): p. 245-9.

28. Schmitz, R.J., et al., *Transgenerational epigenetic instability is a source of novel methylation variants.* Science, 2011. **334**(6054): p. 369-73.

29. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.

30. Qiu, J., *Epigenetics: unfinished symphony.* Nature, 2006. **441**(7090): p. 143-5.

31. Segal, E. and J. Widom, *What controls nucleosome positions?* Trends Genet, 2009. **25**(8): p. 335-43.

32. Ioshikhes, I., et al., *Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences.* J Mol Biol, 1996. **262**(2): p. 129-39.

33. Satchwell, S.C., H.R. Drew, and A.A. Travers, *Sequence periodicities in chicken nucleosome core DNA.* J Mol Biol, 1986. **191**(4): p. 659-75.

34. Segal, E., et al., *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.

35. Albert, I., et al., *Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome.* Nature, 2007. **446**(7135): p. 572-6.

36. Mavrich, T.N., et al., *A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome.* Genome Res, 2008. **18**(7): p. 1073-83.

37. Valouev, A., et al., *A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning.* Genome Res, 2008. **18**(7): p. 1051-63.

38. Mavrich, T.N., et al., *Nucleosome organization in the Drosophila genome.* Nature, 2008. **453**(7193): p. 358-62.

39. Widom, J., *Role of DNA sequence in nucleosome stability and dynamics.* Q Rev Biophys, 2001. **34**(3): p. 269-324.

40. Yuan, G.C., et al., *Genome-scale identification of nucleosome positions in S. cerevisiae.* Science, 2005. **309**(5734): p. 626-30.

41. Li, B., M. Carey, and J.L. Workman, *The role of chromatin during transcription.* Cell, 2007. **128**(4): p. 707-19.

42. Workman, J.L., *Nucleosome displacement in transcription.* Genes Dev, 2006. **20**(15): p. 2009-17.

43. Jin, C., et al., *H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions.* Nat Genet, 2009. **41**(8): p. 941-5.

44. Kobor, M.S. and M.C. Lorincz, *H2A.Z and DNA methylation: irreconcilable differences.* Trends Biochem Sci, 2009. **34**(4): p. 158-61.

45. Jia, D., et al., *Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation.* Nature, 2007. **449**(7159): p. 248-51.

46. Henikoff, S. and L. Comai, *A DNA methyltransferase homolog with a chromodomain exists in multiple polymorphic forms in Arabidopsis.* Genetics, 1998. **149**(1): p. 307-18.

47. Kogan, S. and E.N. Trifonov, *Gene splice sites correlate with nucleosome positions.* Gene, 2005. **352**: p. 57-62.

48. Denisov, D.A., E.S. Shpigelman, and E.N. Trifonov, *Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes.* Gene, 1997. **205**(1-2): p. 145-9.

49. Allemand, E., E. Batsche, and C. Muchardt, *Splicing, transcription, and chromatin: a menage a trois.* Curr Opin Genet Dev, 2008. **18**(2): p. 145-51.

50. Goldstrohm, A.C., et al., *The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1.* Mol Cell Biol, 2001. **21**(22): p. 7617-28.

51. Howe, K.J., *RNA polymerase II conducts a symphony of pre-mRNA processing activities.* Biochim Biophys Acta, 2002. **1577**(2): p. 308-24.

52. de la Mata, M., et al., *A slow RNA polymerase II affects alternative splicing in vivo.* Mol Cell, 2003. **12**(2): p. 525-32.

53. Kolasinska-Zwierz, P., et al., *Differential chromatin marking of introns and expressed exons by H3K36me3.* Nat Genet, 2009. **41**(3): p. 376-81.

54. Andersson, R., et al., *Nucleosomes are well positioned in exons and carry characteristic histone modifications.* Genome Res, 2009.

55. Tilgner, H., et al., *Nucleosome positioning as a determinant of exon recognition.* Nat Struct Mol Biol, 2009.

56. Schwartz, S., E. Meshorer, and G. Ast, *Chromatin organization marks exon-intron structure.* Nat Struct Mol Biol, 2009.

Chapter 2: Methods

BS-seq analysis

Treatment of genomic DNA with sodium bisulfite results in the conversion of unmethylated cytosines into uracils, which are then converted into thymines during PCR, while methylated cytosines remain unconverted. This DNA can then be sequenced using a next gen sequencer and the reads aligned back to the genome using a specialized aligner such as BS-seeker that accounts for cytosine conversion [1, 2]. Along with the regular quality checks performed on NGS data, an additional check should be done to ensure that unconverted reads are not included in the read pool. This is done by eliminating reads that contain at least three CHH sites that are methylated, as CHH methylation is generally infrequent and it is unlikely that all of the CHH sites are methylated and more likely that the read simply did not properly undergo conversion [3]. The output from an aligner like BS-seeker can be processed such that it ultimately results in discrete counts of unconverted cytosines and converted thymines, allowing the measurement of DNA methylation frequency at a single base resolution across multiple cells. The methylation rate can be calculated by dividing the number of cytosines found at a particular site by the number of reads (both cytosines and thymines) that were found at that site. Unless taken as large aggregates, sites with low coverage, usually below 5, should be eliminated, as they are prone to counting biases.

Although it is tempting to consider these individual methylation frequency as the 'true' methylation frequency in cells, what is actually being measured is not methylation frequency, but discrete counts that imply a certain methylation frequency. That is, the discrete counts provide an estimate of actual methylation, and not a direct measurement of it; as when one picks red or green marbles several times from a bag containing some distribution of red and green marbles and estimates the fraction of marbles that are red based on his results. It is not necessary to consider this fact while taking averages over large numbers of sites, as underestimates are as likely as overestimates, and on average one will negate the other.

However, when comparing methylation levels at one site across two or more samples, this uncertainty should be considered. A site can be called differentially methylated if indeed, the most likely probability (that is, the mean methylation frequency) is outside of the confidence interval of the methylation frequency likelihood distribution of the other sample. The methylation frequency likelihood distribution is effectively modeled by the binomial distribution, which takes into account the implied frequency and the number of reads that provided information. However, ensuring different methylation frequency distributions is not sufficient. While this accounts for statistical significance, it overlooks biological significance. Ultimately, a threshold should be chosen for the difference in the implied methylation frequencies. The threshold chosen should be context-specific, as methylation frequencies vary wildly based on context and organism.

Along a similar line, when considering rates of differential methylation, it is important to consider only the pool of sites for which coverage is in a similar range in all of the samples. Because the likely methylation distributions vary with coverage, where sites with high coverage have sharp distributions, and those with low coverage having wide distributions, it is important to distinguish sites for which a differential methylation call cannot be made from those that are not differentially methylated. Ultimately, these two groups (those that are not differentially methylated and those that cannot be called as differentially methylated) do not change the numerator in the calculation; but the denominator changes vastly. Therefore, the results are influenced by coverage.

Once individual differentially methylated sites have been identified, regions of differential methylation can be identified, as differential methylation tends to be clustered. In order to identify clusters, a simple binomial based model can be utilized, such that the frequency of differential methylation and the number of differentially methylated sites and the number of

cytosines in the cluster are considered. Once a region is identified as a DMR, the region can be expanded or contracted until the optimal cluster size is found.

MNase-seq analysis

Treating chromatin with Micrococcal nuclease (MNase), a DNase, results in variable sized chromatin fragments that are protected from digestion due to the presence of proteins bound to DNA [4]. Because histones account for the vast majority of protein bound to DNA, this becomes a way of detecting primarily nucleosomes. Furthermore, size selecting the resulting chromatin fragments, such that only ~150 bp fragments remain, ensures that only nucleosomes remain. Once the protein is washed away from the DNA, the resulting DNA fragments can be sequenced and further analyzed [5, 6].

Each read represents the first few bases, depending on the sequencing length, of chromatin fragments, and may be thought of as representing one end of a nucleosome. In general, multiple reads indicate a nucleosome that is better positioned -- that is, where the nucleosome begins at the same loci in multiple cells in the sample, and is therefore frequently found at that locus. However, because of PCR biases and potential mapping artifacts, if reads disproportionately pile up at one locus, their number must be artificially reduced, such that they do not disproportionately influence generalized conclusions. In deciding the cutoff, it is generally a good idea to look at a histogram of the distribution of the number of reads at each site and identify a limit where the distribution becomes sparse.

The above generates a profile of nucleosome start site likelihood. However, a profile of nucleosome occupancy can be generated by ‘smearing’ the nucleosome start site profile over the length of nucleosome-occupied DNA (147 bp). In fact, because nucleosome occupancy is non-stranded (the double helix wraps the nucleosome), the nucleosome start sites on the positive strand should be smeared downstream (in the 5’-3’ direction) and added to the start

sites on the negative strand, which will be smeared upstream (also in the 5'-3' direction but on the negative strand). This profile can then be used in answering questions related to nucleosome occupancy.

Furthermore, additional, more sophisticated methods may be employed in analyzing MNase-seq data. Because we know the length of nucleosome-occupied DNA (147 bp), we should be able to determine where the nucleosome 'end' site is, given a start site. Moreover, we can recognize the presence of this end site in our data, as it will be present on the opposite strand 147 bp away. Furthermore, we would expect well positioned nucleosomes to have reads on one strand followed by reads 147 bp away on the opposite strand, with very few reads in between. Some have proposed using this idea as a method of assigning a score representing the likelihood of nucleosomal presence [7].

Ultimately, as in ChIP-seq, it is useful to generate a genomic DNA (i.e. remove all proteins) MNase control. With the control, it would be possible to account for biases resulting from the MNase digestion itself, from PCR, from sequencing, etc. A Poisson-based model can be employed in determining regions of about 200 bp that show significant enrichment in nucleosomes compared to genomic DNA. However, because regions are being identified, the sharp boundaries generated by nucleosome start sites would be lost. Also, because of the Poisson distribution handles large counts, even small relative differences are shown as significant. Therefore, simply ranking windows in terms of their 'signal to noise' fraction, where the reads from the genomic DNA digestion are considered noise and the reads from chromatin digestion are considered signal, proved to be a better route.

These methods are reminiscent of methods employed in ChIP-seq analysis, where the problem is similar: given an IP (MNase-seq data) and an input control (MNase on genomic DNA) find the regions that are enriched. However, the one advantage in this analysis over ChIP-seq is that the size of the region of interest is known and thus, the algorithms can take this

into account. In ChIP-seq, if the region size is unknown, as in the case of histone modifications, areas of enrichment have to be dynamically expanded until the areas are optimal for the IP in question. But because of the inherent similarity, and because in many ways MNase-seq can be treated as a specialized case of ChIP-seq, the methods employed in existing ChIP-seq analysis software can be co-opted.

Others have found that nucleosomes have a DNA code, that is, a preference to bind to DNA where certain dinucelotide frequencies are present [5]. They have taken this observation and created a nucleosome occupancy prediction algorithm given a DNA sequence [8]. This artificial data, although trained on yeast nucleosome data, has faithfully recapitulated many of the interesting patterns observed in MNase data. Moreover, this data also allows for the relatively quick analysis of sequence perturbations, such as identifying if GC content, or splice site junctions are responsible for certain nucleosome occupancy patterns. Of course, there are limits to this algorithm and the data it produces. It is well known that nucleosomes can be, and are frequently, rearranged to suite the cell's needs – during transcription, replication, etc. However, because the algorithm only takes the genome sequence as an input, it is not possible to study nucleosome dynamics in this manner.

References

1. Feng, S., et al., *Determining DNA methylation profiles using sequencing.* Methods Mol Biol, 2011. **733**: p. 223-38.
2. Chen, P.Y., S.J. Cokus, and M. Pellegrini, *BS Seeker: precise mapping for bisulfite sequencing.* BMC Bioinformatics, 2010. **11**: p. 203.
3. Cokus, S.J., et al., *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008. **452**(7184): p. 215-9.
4. Heins, J.N., et al., *Characterization of a nuclease produced by Staphylococcus aureus.* J Biol Chem, 1967. **242**(5): p. 1016-20.
5. Segal, E., et al., *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.
6. Yuan, G.C., et al., *Genome-scale identification of nucleosome positions in S. cerevisiae.* Science, 2005. **309**(5734): p. 626-30.
7. Valouev, A., et al., *Determinants of nucleosome organization in primary human cells.* Nature, 2011. **474**(7352): p. 516-20.
8. Kaplan, N., et al., *The DNA-encoded nucleosome organization of a eukaryotic genome.* Nature, 2009. **458**(7236): p. 362-6.

Chapter 3: Relationship between nucleosome positioning and DNA methylation.

The following was first published in the journal Nature on May 30th, 2010.

Ramakrishna K. Chodavarapu, Suhua Feng, Yana V. Bernatavichute, Pao-Yang Chen, Hume Stroud, Yanchun Yu, Jonathan Hetzel, Frank Kuo, Jin Kim, Shawn J. Cokus, David Casero, Maria Bernal, Peter Huijser, Amander T. Clark, Ute Krämer, Sabeeha S. Merchant, Xiaoyu Zhang, Steven E. Jacobsen, and Matteo Pellegrini. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature,* 466:388-392.

Relationship between nucleosome positioning and DNA methylation.

Ramakrishna K. Chodavarapu[1#], Suhua Feng[1,2#], Yana V. Bernatavichute[1,3], Pao-Yang Chen[1], Hume Stroud[1], Yanchun Yu[4], Jonathan Hetzel[1], Frank Kuo[1], Jin Kim[1], Shawn J. Cokus[1], David Casero[1], Maria Bernal[5], Peter Huijser[6], Amander T. Clark[1,7], Ute Krämer[5], Sabeeha S. Merchant[3,8], Xiaoyu Zhang[4], Steven E. Jacobsen[1,2,3*], and Matteo Pellegrini[1,3*]

1. Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, California, United States of America
2. Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, California, United States of America
3. Molecular Biology Institute, University of California Los Angeles, Los Angeles, California, United States of America
4. Department of Plant Biology, University of Georgia, Athens, GA 30602, United States of America
5. University of Heidelberg, BIOQUANT 23, R. 645, Im Neuenheimer Feld 267, D-69120 Heidelberg, Germany
6. Department of Molecular Plant Genetics, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany
7. Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, California, United States of America
8. Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, United States of America

#these authors contributed equally

*corresponding authors

Nucleosomes compact and regulate access to DNA in the nucleus, and are composed of approximately 147 bases of DNA wrapped around a histone octamer[1, 2]. Here we report a genome-wide nucleosome positioning analysis of *Arabidopsis thaliana* utilizing massively parallel sequencing of mononucleosomes. By combining this data with profiles of DNA methylation at single base resolution, we identified ten base periodicities in the DNA methylation status of nucleosome-bound DNA and found that nucleosomal DNA was more highly methylated than flanking DNA. These results suggest that nucleosome positioning strongly influences DNA methylation patterning throughout the genome and that DNA methyltransferases preferentially target nucleosome-bound DNA. We also observed similar trends in human nucleosomal DNA suggesting that the relationships between nucleosomes and DNA methyltransferases are conserved. Finally, as has been observed in animals, nucleosomes were highly enriched on exons, and preferentially positioned at intron-exon and exon-intron boundaries. RNA Pol II was also enriched on exons relative to introns, consistent with the hypothesis that nucleosome positioning regulates Pol II processivity. DNA methylation is enriched on exons, consistent with the targeting of DNA methylation to nucleosomes, and suggesting a role for DNA methylation in exon definition.

To investigate the position of nucleosomes in *Arabidopsis thaliana*, we sequenced Micrococcal Nuclease (MNase) digested nucleosomal DNA using an Illumina GAII sequencer to achieve a roughly 68-fold coverage of nucleosomes (see Supplementary Methods). The data are displayed in a modified UCSC genome browser (http://apis.pellegrini.mcdb.ucla.edu:8081/Ath-Nucleosomes/, (name = reviewer, password = DkZmv9wJ) along with nucleosome density tracks that allow easy visualization of nucleosome positions throughout the genome (Fig. 1a).

To obtain a chromosomal view of nucleosome content, we plotted reads in bins of 100 kilobases tiling the chromosomes. As a control for biases in mapping and sequencing we also sequenced a library of randomly sheared Arabidopsis genomic DNA. We then normalized the nucleosome counts by the number of uniquely mapping genomic DNA counts within each bin along the chromosomes. Nucleosome content was relatively uniform throughout the euchromatic regions of chromosomes, but showed significant enrichment in pericentromeric heterochromatin regions, (Fig. 1b, Supplementary Fig. 1). To confirm these results, we performed ChIP-Seq utilizing an antibody against unmodified histone H3 (Fig. 1b, Supplementary Fig. 1). These observation suggests that nucleosomes are more densely packed in pericentromeric regions that are transcriptionally silent, are rich in transposons and contain heavily methylated DNA, than in the euchromatic arms[3].

By examining the relationship between reads that map to opposite strands of DNA, we observed a peak of reads on the reverse strand that occurred approximately 145 -170 bases downstream of the reads on the forward strand (Supplementary Fig. 2a). This broad peak in the general vicinity of the accepted size of a nucleosome suggests that many nucleosomes are well positioned, leading to the repeated sequencing of both the forward and reverse complement of the corresponding nucleosome regions. Similarly, when we plotted the correlation between positive strand reads with other positive strand reads we saw a progressively decreasing

correlation from the start of the nucleosome (Fig. 1c, Supplementary Fig. 2b), with smaller peaks spaced at 174 and 355 base pairs from the starting position of the reference reads suggesting some preference for regular spacing of nucleosomes genome wide. Assuming that nucleosomes are wrapped around 147 base pairs of DNA, the average length of the linkers of regularly spaced nucleosomes is about 30 base pairs. Consistent with the higher content of nucleosomes in pericentromeric heterochromatin regions, we found that the peaks at 174 and 355 were more pronounced in these regions than in the euchromatic arms (Fig. 1c), suggesting that pericentromeric nucleosomes have a higher tendency to be in regularly spaced arrays. These results are consistent with earlier evidence for more regular spacing of nucleosomes in Drosophila heterochromatin[4].

Similar to findings in animal and fungal high-throughput nucleosome sequencing studies[5-8], we found 10 base periodicities in WW (W = A or T) dinucleotides, and SS (S = G or C) dinucleotides that were 5 bases out of phase with the WW dinucleotides (Fig. 1d, Supplementary Figs 3-5). WW dinucleotides are favored at sites where the minor groove faces the histone core, while SS dinucleotides are favored at sites where the minor groove faces away from the histone core[9, 10]. The out of phase peaks in the frequencies of these dinucleotides leads to optimal bending of DNA, as A/T nucleotides cause a negative base roll and G/C nucleotides cause a positive base roll[9, 10].

To study the relationship of DNA methylation with nucleosome positions, we utilized our single nucleotide resolution whole genome bisulfite-sequencing data[3]. Arabidopsis cytosines are methylated by one of three different DNA methyltransferases depending on their sequence context. CG sites are methylated by MET1, and CHG sites (where H is A, C or T) are methylated by CMT3. Finally, CHH sites are methylated by DRM2, a *de novo* methyltransferase that is targeted by small RNAs[11]. Remarkably, all three types of methylation showed a 10 base periodicity on nucleosomal DNA, which was in phase with the WW dinucleotides and out of

phase with the SS dinucleotides (Figs 1d, 2a-f, Supplementary Fig. 6). These methylation preferences were not correlated with preferences for CG, CHG or CHH sequences at these locations (Supplementary Fig. 7). Since DNA methyltransferases access the major groove, this methylation would be on DNA which is on the outside of the nucleosome (minor groove facing the histones) and thus more accessible to the DNA methyltransferases. This in turn suggests that DNA might be in part methylated as the DNA is still bound to nucleosomes, leading to the observed 10 base pair periodicity. It has been proposed that chromatin remodeling enzymes are required for DNA methyltransferases to gain access to the DNA, and indeed DRD1, DDM1 and LSH1 are remodelers known to be important factors controlling DNA methylation[12]. However, our data suggests that nucleosomal DNA may also be a substrate for DNA methyltransferases *in vivo*, demonstrating a prominent role of the nucleosome in determining methylation patterning throughout the genome.

We previously reported a 10-nucleotide periodicity in CHH methylation data when performing autocorrelation analysis on the methylation pattern of the whole genome[3]. Our previous interpretation was that the structure of the DRM2 enzyme might be responsible for this pattern, since the orthologous Dnmt3 enzymes in mammals are known to act as heteromeric complexes in which two methyltransferase active sites have a spacing equivalent to roughly 10 nucleotides of DNA[13]. However, the current data in which we see 10 base pair periodicities for all types of methylation suggest a more general explanation: that nucleosomes are to some extent dictating access to the DNA and therefore setting the register of methylation for all DNA methyltransferases. Nucleosomal preferences could also partially explain the strong sequence preferences that we previously observed for CHG and CHH methylation[3]. Highly methylated cytosines tended to be immediately followed by A/T but not C, consistent our finding that DNA methylation is out of phase with CC dinucleotides (Supplementary Fig. 6).

On a larger scale, we also observed that levels of all three types of DNA methylation were higher in nucleosome spanning DNA than flanking DNA, indicating that nucleosome bound DNA is enriched for DNA methylation (Fig. 2, Supplementary Fig. 8). This finding supports the view that nucleosomes are preferentially targeted by DNA methyltransferases. In the case of CMT3, it is predicted that this enzyme is recruited or activated by histone H3 lysine 9 dimethylation, because a knockout of histone H3 lysine 9 methyltransferase mimics a knock out of CMT3, and because the CMT3 chromodomain can bind to methylated histones[14]. However our data suggest that all types of methylation are enriched on nucleosome bound DNA suggesting that nucleosomes or histone modifications may assist in the recruitment of all of the Arabidopsis DNA methyltransferases.

To test whether the patterns of DNA methylation on nucleosomal DNA are also found in human, we utilized previously published MNase nucleosome sequencing data[15] together with single nucleotide resolution bisulfite sequencing data on the human embryonic stem cell line HSF1 (see Supplementary Methods). We found that human nucleosome-bound DNA also showed a ten base periodicity in its CG, CHG, and CHH methylation status (Fig. 3). Furthermore, as in *Arabidopsis*, the overall level of methylation was higher on nucleosome bound DNA than flanking regions. This suggests that nucleosome architecture plays a role in shaping DNA methylation patterning in the human genome, and is consistent with the recent finding of stable anchoring of Dnmt3 DNA methyltransferases to mammalian chromatin[16] and more generally with the conservation of DNA methyltransferase function in plants and animals[17]. We also analyzed DNA methylation patterning on nucleosomes in different regions of the Arabidopsis and human genomes, including genes, promoters, pericentromeric regions, and euchromatic arm regions (Supplementary Fig. 9-17) and found that the 10-nucleotide periodicity was found in all cases, suggesting that the relationship between nucleosome positioning and DNA methylation is general.

We observed that nucleosomes were much more abundant in exons than in introns (see Fig. 1a for an example), consistent with findings from several recent nucleosome positioning studies in other organisms[18-23]. The number of nucleosomes per base pair in introns was only 63% of the level found in exons. Furthermore, we found a strong peak of nucleosome start sites at intron-exon junctions and at exon-intron junctions (Fig. 4, Supplementary Fig. 18). The enrichment of nucleosomes in exons was not solely due to higher GC content (Supplementary Fig. 19a-c), or because of consensus splice site sequences (Supplemental Fig. 19d-f), and was confirmed using independent chromatin immunoprecipitation methods (Supplementary Fig. 20). Longer exons contained a higher number of nucleosomes. Examination of exons in the size ranges of 170-240 bp, 315-330 bp, 480-550 bp, and 645-715 bp revealed peaks of one, two, three or four well positioned nucleosomes respectively (Fig. 4a), suggesting that nucleosome are phased within exons.

Strong nucleosome enrichment on exons was detected both for genes that are highly expressed, and those not expressed in the shoot tissue used for analysis of nucleosomes (Supplementary Fig. 21), suggesting that DNA sequences position nucleosomes in the absence of active transcription and splicing. Consistent with this hypothesis, utilizing a nucleosome positioning prediction algorithm[24] we found similar patterns of nucleosome positioning in introns and exons between the experimental and theoretical datasets (Supplementary Fig. 22). Similarly, using theoretically predicted nucleosome positions, we observed similar patterns of DNA methylation on nucleosome bound DNA (Supplementary Fig. 23), as well as enrichment of predicted nucleosomes in pericentromeric regions (Supplementary Fig. 24).

Because nucleosomes present a barrier to RNA polymerase II (Pol II) transcription, we tested for Pol II occupancy in exons using a chromatin immunoprecipitation microarray approach. We observed significant enrichment of Pol II in exons relative to introns, consistent with the hypothesis that Pol II is paused on exonic DNA (Fig. 4). One possibility is that Pol II

stalling on exons could enhance efficient splicing of upstream exons, thus aiding in the fidelity of exon definition. This is consistent with the finding of Pol II enrichment on human exons and suggests that Pol II enrichment on exons might be a common eukaryotic feature[20]. Furthermore, particular histone modifications have been recently shown to recruit splicing regulators, providing additional possible mechanisms for the regulation of splicing by nucleosome positioning[25].

Because of the enhancement of DNA methylation over nucleosomal DNA, and the enrichment of nucleosomes on exons, a prediction is that DNA methylation should be enhanced on exons relative to introns. Indeed, we found depletion of DNA methylation in introns and enrichment in exons in a pattern that was similar to nucleosome occupancy (Fig. 4b). This is consistent with recent findings of enhanced exonic methylation in other plant species, as well as in the Chlamydomonas, honeybee, Ciona and human genomes[26, 27]. This suggests the possibility that DNA methylation, which frequently exists in the transcribed regions of active genes, could play a conserved role in exon definition or splicing regulation. These findings also reinforce the view that nucleosomal positions play an important role in shaping the methylation landscape of eukaryotic genomes.

Method Summary

Approximately 300ng of MNase digested mononucleosome DNA were used for Illumina library generation following manufacturer instructions. Libraries were sequenced using Illumina Genome Analyzer II following manufacturer instructions. Resulting reads were mapped to the TAIR7 annotation of the Arabidopsis genome and reads that mapped to multiple sites were eliminated. A total of 24.2 million reads were mapped to the forward strand and 23.9 million reads to the reverse strand, generating a 68-fold coverage of nucleosome space.

Pol II ChIP-chip was performed as previously described[28]. Briefly, crosslinked chromatin was extracted, sonicated, and used in ChIP with an antibody against Pol II CTD (ab817; Abcam, Cambridge, MA). Pol II-bound DNA and input genomic DNA were extracted, amplified, labeled and hybridized to the Afffymetrix tiling microarrays as previously described[28]. Three biological replicates were performed and the $\log_2$ ratios of Pol II-bound RNA over input DNA were calculated using the Tiling Analysis Software (Affymetrix; Santa Clara, CA).

Acknowledgments

We thank members of our laboratories for input and guidance, and the Department of Energy, Institute for Genomics and Proteomics and Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA for support. R. K. C was supported by NIH training grant GM07104. S.F. was a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. P.C. is funded by Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA. M.B was supported by an Alexander-von-Humboldt Fellowship and U.K. by a Heisenberg Fellowship. X.Z. was supported by a Faculty Research Grant (JR-040) from the University of Georgia. Research was supported by NIH grants GM60398 and GM42143, and NSF Plant Genome Research Program #0701745. S. E. J. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions

S.E.J. and M.P. designed the research. S.F., Y.V.B, H.S., Y.Y., J.H., M.B., P.H., U.K., S.S.M., X.Z. and A.T.C. performed experiments. R.K.C, P.Y.C, F.K., J.K., and S.J.C analyzed data. R.K.C., S. F., S.E.J, and M.P wrote the paper.

Author Information

All sequencing files have been deposited in GEO under accession code GSE21673. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.E.J. (jacobsen@ucla.edu) or M.P. (matteop@mcdb.ucla.edu).

Figure Legends

Figure 1 | Characterization of *A. thaliana* nucleosome data.

a, A representative UCSC Browser screenshot of a gene, nucleosome reads found in the region, and the calculated nucleosome densities. **b,** A chromosomal view of nucleosome read counts and ChIP-seq read counts in 100kb tiles along chromosome 1 shows nucleosome enrichment in pericentromeric regions. **c,** An autocorrelation of positive stranded reads shows local peaks at positions 174 and 355 bases. **d**, AA and GC dinucleotide and CG methylation profiles show a 10 base periodicity over nucleosome-bound DNA, with DNA methylation profiles in phase with WW dinucleotide profiles, and out of phase with SS dinucleotide profiles.

Figure 2 | DNA methylation profiles of nucleosome-bound DNA in *Arabidopsis*.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). Fast Fourier Transforms (FFTs) can be used to deconstruct inherent frequencies in a complex signal. The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(a)**, CHG **(b)**, and CHH **(b)** contexts.

Figure 3 | DNA methylation profiles of nucleosome-bound DNA in Human.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts.

Figure 4 | Nucleosome and PolII levels in Exons.

We performed chromatin immunoprecipitation with an antibody against RNA polymerase II (Pol II), and hybridized the resulting DNA to a whole genome Affymetrix microarray. We normalized these data to randomly sheared genomic DNA to control for probe efficiencies. **a,** Nucleosomes are phased in exons. Exon size limits were selected such that the exon could house no more than 1, 2, 3, or 4 nucleosomes. Nucleosome midpoints are plotted over these intron-exon boundaries. **b,** Each exon was divided into 25 equal sized bins and the nucleosome midpoints, Pol II and CG methylation levels are plotted over the exon and flanking introns.

References

1. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-60 (1997).
2. Kornberg, R.D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-94 (1999).
3. Cokus, S.J. et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215-9 (2008).
4. Sun, F.L., Cuaycong, M.H. & Elgin, S.C. Long-range nucleosome ordering is associated with gene silencing in Drosophila melanogaster pericentric heterochromatin. *Mol Cell Biol* **21**, 2867-79 (2001).
5. Albert, I. et al. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572-6 (2007).
6. Mavrich, T.N. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**, 1073-83 (2008).
7. Valouev, A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051-63 (2008).
8. Mavrich, T.N. et al. Nucleosome organization in the Drosophila genome. *Nature* **453**, 358-62 (2008).
9. Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* **34**, 269-324 (2001).
10. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-8 (2006).
11. Henderson, I.R. & Jacobsen, S.E. Epigenetic inheritance in plants. *Nature* **447**, 418-24 (2007).
12. Kobor, M.S. & Lorincz, M.C. H2A.Z and DNA methylation: irreconcilable differences. *Trends Biochem Sci* **34**, 158-61 (2009).
13. Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**, 248-51 (2007).
14. Lindroth, A.M. et al. Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. *EMBO J* **23**, 4286-96 (2004).
15. Schones, D.E. et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-98 (2008).
16. Jeong, S. et al. Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA. *Mol Cell Biol* **29**, 5366-76 (2009).
17. Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**, 204-220 (2010).
18. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* (2009).
19. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* (2009).
20. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* (2009).
21. Kolasinska-Zwierz, P. et al. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41**, 376-81 (2009).
22. Spies, N., Nielsen, C.B., Padgett, R.A. & Burge, C.B. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36**, 245-54 (2009).
23. Ponts, N. et al. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Res* **20**, 228-38.
24. Kaplan, N. et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-6 (2009).

25. Luco, R.F. et al. Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000 (2010).
26. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*.
27. Laurent, L. et al. Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320-31 (2010).
28. Zhang, X. et al. Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol* **5**, e129 (2007).

Figure 1

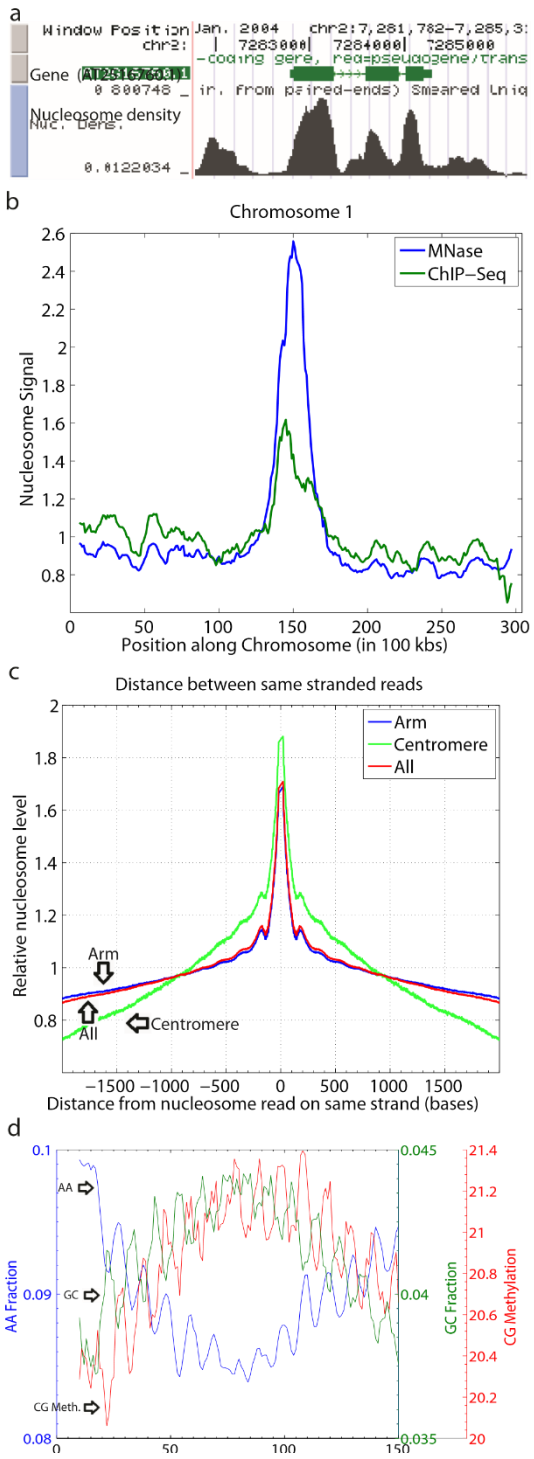

a
Window Position
Jan. 2004 chr2:7,281,782-7,285,3
chr2:
7283000
7284000
7285000
Gene
Nucleosome density
Nuc. Dens.
b
Chromosome 1
MNase
ChIP–Seq
Nucleosome Signal
Position along Chromosome (in 100 kbs)
c
Distance between same stranded reads
Arm
Centromere
All
Relative nucleosome level
Distance from nucleosome read on same strand (bases)
d
AA
GC
CG Meth.
AA Fraction
GC Fraction
CG Methylation

Figure 2

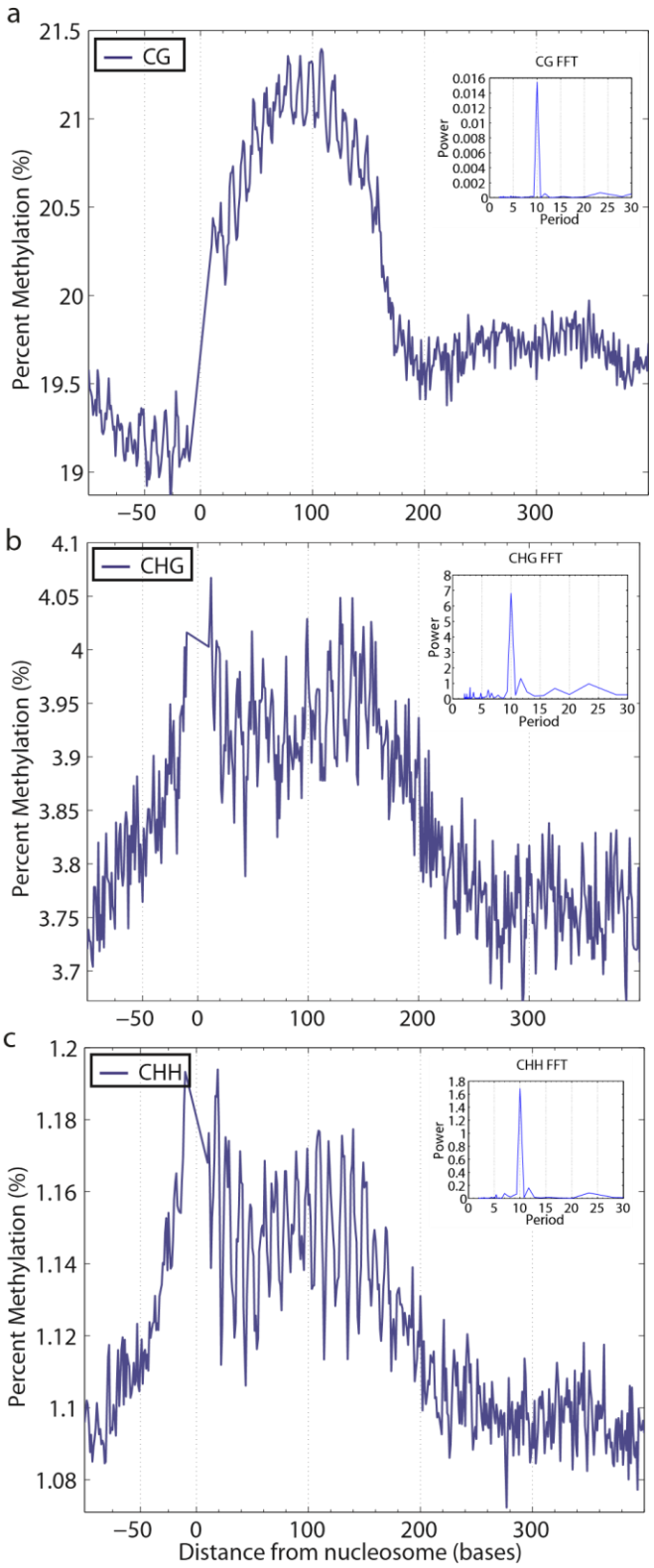

a
CG
CG FFT
Power
Period
Percent Methylation (%)
b
CHG
CHG FFT
Power
Period
Percent Methylation (%)
c
CHH
CHH FFT
Power
Period
Percent Methylation (%)
Distance from nucleosome (bases)

Figure 3

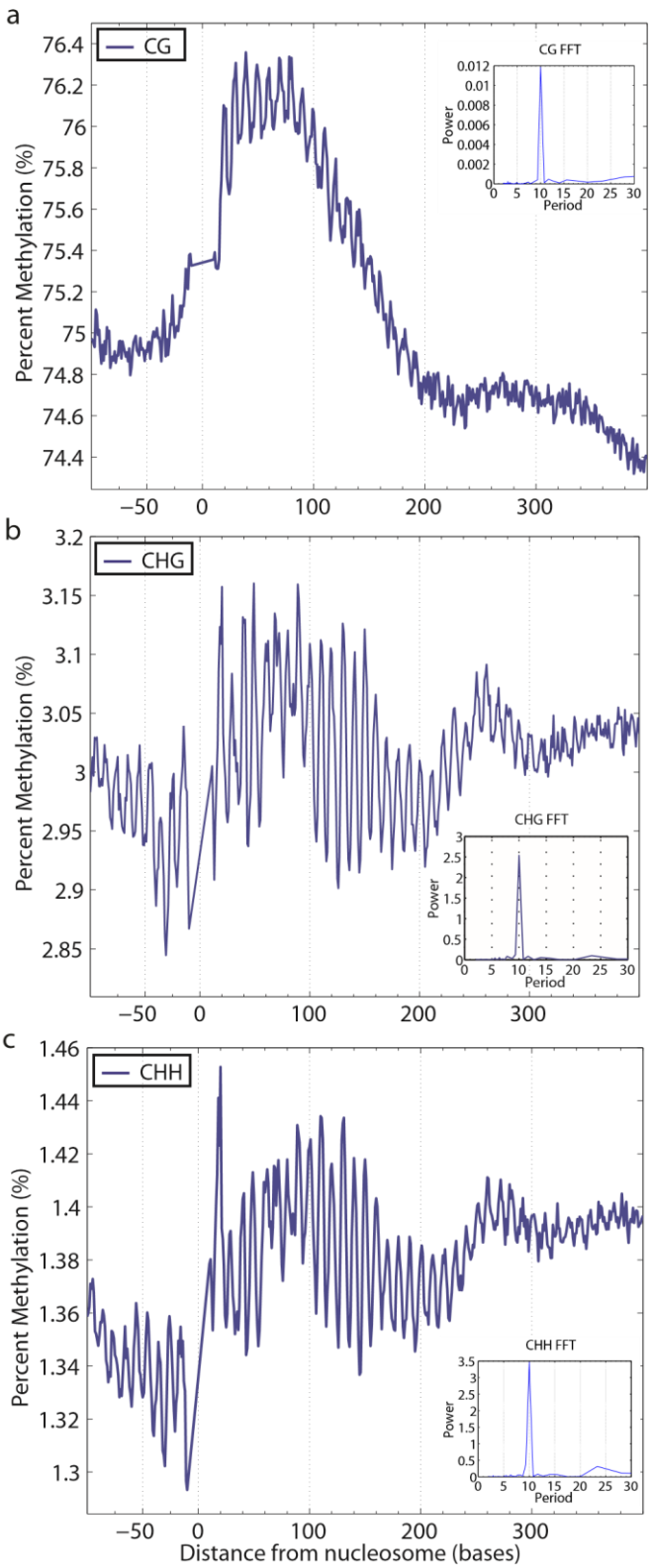
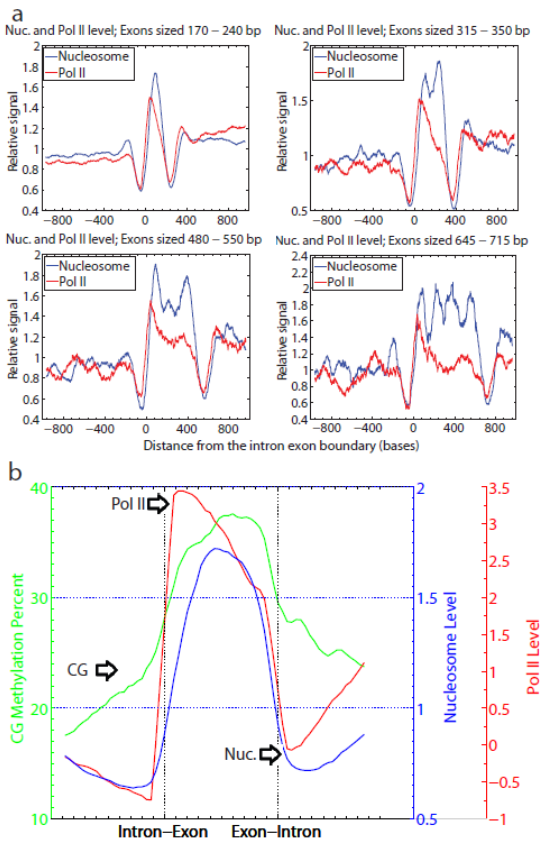

a
CG
CG FFT
Power
Period
Percent Methylation (%)
b
CHG
CHG FFT
Power
Period
Percent Methylation (%)
c
CHH
CHH FFT
Power
Period
Percent Methylation (%)
Distance from nucleosome (bases)

Figure 4


a
Nuc. and Pol II level; Exons sized 170 – 240 bp
Nuc. and Pol II level; Exons sized 315 – 350 bp
Nuc. and Pol II level; Exons sized 480 – 550 bp
Nuc. and Pol II level; Exons sized 645 – 715 bp
Nucleosome
Pol II
Relative signal
Distance from the intron exon boundary (bases)
b
Pol II
CG
Nuc.
CG Methylation Percent
Nucleosome Level
Pol II Level
Intron–Exon
Exon–Intron

Supplementary Methods

Arabidopsis nucleosome preparation

Arabidopsis nuclei were prepared as described previously (Bernatavichute et al. 2008 PLoS ONE 3: e3156) with the following modifications. After spinning the nuclei pellet at 3000 rpm for 5 min (SS-34, Sorvall), the pellet was resuspended in 5 ml of HBB buffer (25 mM Tris-Cl pH 7.6, 0.44 M sucrose, 10 mM MgCl2, 0.1% Triton-X, 10 mM beta-mercaptoethanol), applied to a 40/60% Percoll (GE Healthcare) gradient in HBB and spun 2000 rpm for 20 min (SS-34, Sorvall). Isolated nuclei were washed two times in HBB buffer and flash frozen in liquid nitrogen in HBC buffer (25 mM Tris-Cl pH 7.6, 25 mM Tris-Cl pH 7.6, 440 mM sucrose, 10 mM MgCl2 and 0.1% Triton-X, 10 mM beta-mercaptoethanol, 20% glycerol). Nuclei from 1/5 of each preparation were treated with four ul of RNAse A, 10 ug/ul, (Qiagen) and digested with 0.02 U/ul (final concentration) of Micrococcal Nuclease (Takara), 3 min. Digestion was done in digestion buffer (16 mM Tris-Cl, pH 7.6, 50 mM NaCl, 2.5 mM CaCl2, 0.01 mM PMSF and EDTA-free protease inhibitor cocktail (Roche)) and stopped with 10 mM EDTA and treated with Proteinase K (Roche). DNA was purified either with phenol:chlorophorm extraction and precipitated with salts and ethanol or with ChIP DNA Clean and Concentrator kit (Zymo Research). Purified DNA was run on 2% agarose gel and bands corresponding to ~150 bp were cut and purified with a Gel Purification kit (Qiagen). We cloned and performed traditional DNA sequencing of digested fragments to check the quality of the library. The average length of 55 fragments was 146 nucleotides.

RNA-Seq expression

Plants were grown in a climate-controlled growth chamber with a photoperiod of 11 h light, 13 h dark, at a photon efflux density of 145 µmol m^{-2} s^{-1} during the day, and day and night

temperatures of 20 ºC and 18 ºC, respectively. 6-week-old plants were harvested for RNA extraction.

To generate mRNA expression profiles, total RNA was isolated from 100-mg subsamples of tissues. RNase-free DNase set for on-column DNase digest and RNeasy spin columns (Qiagen, Hilden, Germany) were used according to the manufacturer's instructions for subsequent digestion of genomic DNA and RNA purification.

Ten µg of total RNA per sample were used to generate the cDNA Colony Template Libraries (CTL) for high-throughput DNA sequencing using Illumina technology according to the Whole Transcriptome Shotgun protocol (Fasteris SA, Genome Analyzer Service, Switzerland). CTLs were sequenced to obtain a library of 36 base-pair long reads.

H3-Chip-seq

Chromatin was immunoprecipitated using an antibody against histone H3 as previously described (Bernatavichute et al. 2008 PLoS ONE 3: e3156). Libraries were constructed and DNA was sequenced according to manufacturer instructions (Illumina).

Nucleosome predictions

Nucleosome predictions were calculated using a previously described algorithm (Kaplan et al. 2009 Nature 458: 362-366). The predictions were made on the TAIR7 version of the *Arabidopsis* genome. The probability scores for nucleosome start sites were used for further analysis.

Human DNA methylation

The HSF1 hESC line was cultured as described previously (Xiao et al. 1999 Embo J 18: 5943-5952). BS-Seq DNA library construction, high-throughput Illumina sequencing by and analysis were performed similar to that previously described (Cokus et al. 2008 Nature 452: 215-219). 684 million 47 nucleotide reads were used in the analysis of DNA methylation of nucleosomal DNA.

Data analysis

Nucleosome-seq and Chip-seq and were mapped using Bowtie (Langmead et al. 2009 Genome Biol 10: R25) with a tolerance of 3 mismatches. mRNA-Seq sequence reads were aligned against the set of transcript sequences from the TAIR v7.0 database using SOAP (Li et al. 2008 Bioinformatics 24: 713-714). A tolerance of up to two-mismatches was employed and reads with more than one match to the reference sequences were eliminated. Genes with more than one isoform were collapsed into a single gene. The total number of alignment hits was used as an estimate for gene expression levels. Because cDNA synthesis was performed using oligo dT priming before random fragmentation, we chose not to normalize transcript count estimates by the gene length. Instead, genes under 1000 bases were discarded due to disproportionate representation in the expression data.

Supplementary Figure Legends

Supplementary Figure 1 | Chromosomal views of nucleosome read distributions.

a-d, A chromosomal view of nucleosome read counts and ChIP-seq read counts in 100kb bins along chromosome 2-5 shows nucleosome enrichment in pericentromeric regions.

Supplementary Figure 2 | Relationship between forward and reverse stranded reads, and between forward stranded reads and other forward stranded reads.

a, The starts of nucleosome reads on the reverse strand peaked 148 bases away from nucleosome reads on the forward strand. **b,** A zoomed in version of **Figure 1c** is provided to demonstrate a ten base periodicity. The 10 base periodicities in both panels is likely due to the rotational setting tendencies of the histone/DNA interface (Albert et al. 2007 Nature 446: 572-576). Because of a 10 base periodicity in some dinucleotides (**Supplementary Figure 3**), anything but a 10 base translational shift would result in a change in the bases that are accessible to histones and would therefore be disfavored.

Supplementary Figure 3 | Dinucleotide profiles across the nucleosome.

a-f, The fraction of each dinucleotide is plotted as a function of the distance from the start of nucleosome reads. AA **(a)**, CT **(b)**, CC **(c)**, AC **(d)**, TT **(e)**, and CG **(f)** dinucleotide plots are shown. For clarity, 10 bases around the beginning of reads have been omitted since these may reflect sequence biases of the micrococcal nuclease enzyme.

Supplementary Figure 4 | FFTs of dinucleotide profiles

a-f, FFTs are calculated for each of the dinucleotide profiles in Supplementary Figure 3. AA **(a)**, CT **(b)**, CC **(c)**, AC **(d)**, TT **(e)**, CG **(f)**, all show a period of 10 bases over the nucleosome-bound region.

Supplementary Figure 5 | AA/TA/TT dinucleotide frequency in Arabidopsis vs. in Saccharomyces

Arabidopsis AA/TA/TT dinucleotide frequency is compared to that in Saccharomyces (Segal et al. 2006 Nature 442: 772-778).

Supplementary Figure 6 | CHG and CHH methylation occurs in phase with AA dinucleotides.

a,b, As with CG methylation (Figure 1D), CHG **(a)** and CHH **(b)** methylation profiles are also in phase with AA dinucleotides and out of phase with GC dinucleotides. Lines were drawn to indicate peaks in AA dinucleotides. **c-e,** For additional reference, CC dinucleotides are plotted along with CG methylation **(c)**, CHG methylation **(d)**, and CHH methylation **(e)**.

Supplementary Figure 7 | CG, CHG and CHH methylation vs CG, CHG, and CHH sequences.

CG methylation **(a)**, CHG **(b)** and CHH **(c)** methylation profiles are plotted along with CG dinucleotides, CHG trinucleotides, and CHH trinucleotides. Lines were drawn to indicate peaks in AA dinucleotides.

Supplementary Figure 8 | Methylation profiles at H3 ChIP-Seq reads.

As an independent confirmation of the enrichment of methylation on nucleosomes, chromatin immunoprecipitation was performed using an antibody against unmodified histone H3 and the resulting fragments were sequenced and mapped. **a-c,** The weighted average percent CG **(a)**, CHG **(b)**, and CHH **(c)** DNA methylation was calculated and plotted at each distance from ChIP-seq read start sites (0).

Supplementary Figure 9 | DNA methylation profiles of nucleosome-bound DNA in Arabidopsis promoters.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in promoter regions - defined as 5 kb upstream of TSSs - were considered for this analysis.

Supplementary Figure 10 | DNA methylation profiles of nucleosome-bound DNA in Arabidopsis genes.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in genes were considered for this analysis.

Supplementary Figure 11 | DNA methylation profiles of nucleosome-bound DNA in Arabidopsis repeats.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in repetitive regions were considered for this analysis. Tandem repeats, inverted repeats, dispersed repeats, and telomeric regions were included.

Supplementary Figure 12 | DNA methylation profiles of nucleosome-bound DNA in Arabidopsis euchromatic regions.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in euchromatic regions were considered for this analysis.

Supplementary Figure 13 | DNA methylation profiles of nucleosome-bound DNA in Arabidopsis centromeric regions.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in centromeric regions were considered for this analysis.

Supplementary Figure 14 | DNA methylation profiles of nucleosome-bound DNA in Human promoters.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in promoter regions - defined as 5 kb upstream of TSSs - were considered for this analysis.

Supplementary Figure 15 | DNA methylation profiles of nucleosome-bound DNA in Human genes.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in genes were considered for this analysis.

Supplementary Figure 16 | DNA methylation profiles of nucleosome-bound DNA in Human repeats.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes in repeats were considered for this analysis.

Supplementary Figure 17 | DNA methylation profiles of nucleosome-bound DNA in Human CpG islands.

The weighted average percent DNA methylation was calculated and plotted at each distance from nucleosome start sites (0). **a, b, c,** CG methylation **(a)**, CHG methylation **(b)**, and CHH methylation **(c)** each show a 10 base periodicity over nucleosome-bound DNA (1-147 bases). **d, e, f,** The FFTs calculated over the region of the nucleosome demonstrate this periodicity in CG **(d)**, CHG **(e)**, and CHH **(f)** contexts. Only nucleosomes within 1 kb of CpG islands were considered for this analysis.

Supplementary Figure 18 | Nucleosome and Pol II levels in Exons.

We performed chromatin immunoprecipitation with an antibody against RNA polymerase II (Pol II), and hybridized the resulting DNA to a whole genome Affymetrix microarray. We normalized

these data to randomly sheared genomic DNA to control for probe efficiencies. Nucleosome midpoints and Pol II levels are plotted over the intron-exon boundary.

Supplementary Figure 19 | Nucleosome enrichment over exons is GC independent and not due to consensus site sequences.

a-c, Exons were categorized based upon different GC content constraints to address the possibility that GC content of the DNA alone could explain nucleosomal positioning at the intron-exon boundary, as exons are known to have a higher GC content than introns. **a,** Exons were scored based on their GC content and separated into groups - low GC (bottom 33%) and high GC (top 33%). **b,** Exons were scored based on the GC content of the immediately preceding intron and similarly segregated into low and high GC groups. **c,** Finally, exons were scored by the change in GC content over the intron-exon boundary, and similarly segregated. **a-c,** In each panel, the mean GC content in introns of the low and high groups is indicated on the left, and the mean GC content in exons of the low and high groups is indicated on the right. In each case, the nucleosome profile at the boundary retains its general shape, suggesting that GC content alone cannot explain the nucleosome patterning. **d,** A nucleosome prediction algorithm was used to profile the likelihood of nucleosomes starting across the intron-exon boundary. The profiles for unmodified exon boundaries and boundaries where the consensus site is randomized are very similar, suggesting that the consensus splice site is not necessary to position nucleosomes. **e,** The consensus site is shown, with the arrow representing the start of the exon. **f,** The same loci are shown, but where the nucleotides were randomized.

Supplementary Figure 20 | Microarrray confirmation of nucleosome profile at intron-exon boundaries.

a-c, We also sought to determine whether the enrichment of nucleosomes in exons relative to introns could be an artifact of our technique of isolating mononucleosomes after MNase digestion. We analyzed our previously published dataset in which we immunoprecipitated formaldehyde crosslinked and sonicated chromatin with antibodies against the unmodified C terminus of histone H3 **(a)**, H3K4me1 **(b)**, or H3K27me3 **(**c**)**, and hybridized the resulting DNA to a whole genome Affymetrix tiling array (Zhang et al. 2007 PLoS Biol 5: e129; Zhang et al. 2009 Genome Biol 10: R62). We normalized this data with control DNA from crosslinked and sonicated input chromatin that had not undergone immunoprecipitation, to control for probe hybridization efficiencies. H3 **(a)**, H3K4me1 **(b)**, and H3K27me3 **(c)** ChIP-chip signals over the intron-exon boundary confirm results presented in Figure 4a.

Supplementary Figure 21 | Expression independence of nucleosome profile at intron-exon boundary.

a, To test if active transcription or splicing is causing the nucleosomes to become phased on exons, we categorized exons into two groups: highly and poorly transcribed groups. The highly transcribed group refers to those top 1/3rd exons that were derived from genes with the highest RNA-seq transcript counts. The poorly transcribed group refers to those exons that were derived from genes with no RNA-seq transcript counts. The two profiles are very similar, suggesting that nucleosome positioning signals at the exon are not strongly dependent on transcription or splicing and are instead likely to be hard coded in the exonic DNA sequences.

Supplementary Figure 22 | Predicted nucleosomes at the intron-exon boundary

a-c, Exons were categorized in the manner described in Supplementary Figure 9 and predicted nucleosome start sites were plotted around exons in each of these categories. **d-h** Exons were size selected as described in Figure 4; predicted nucleosome start sites and MNase nucleosome start sites were plotted over these exons.

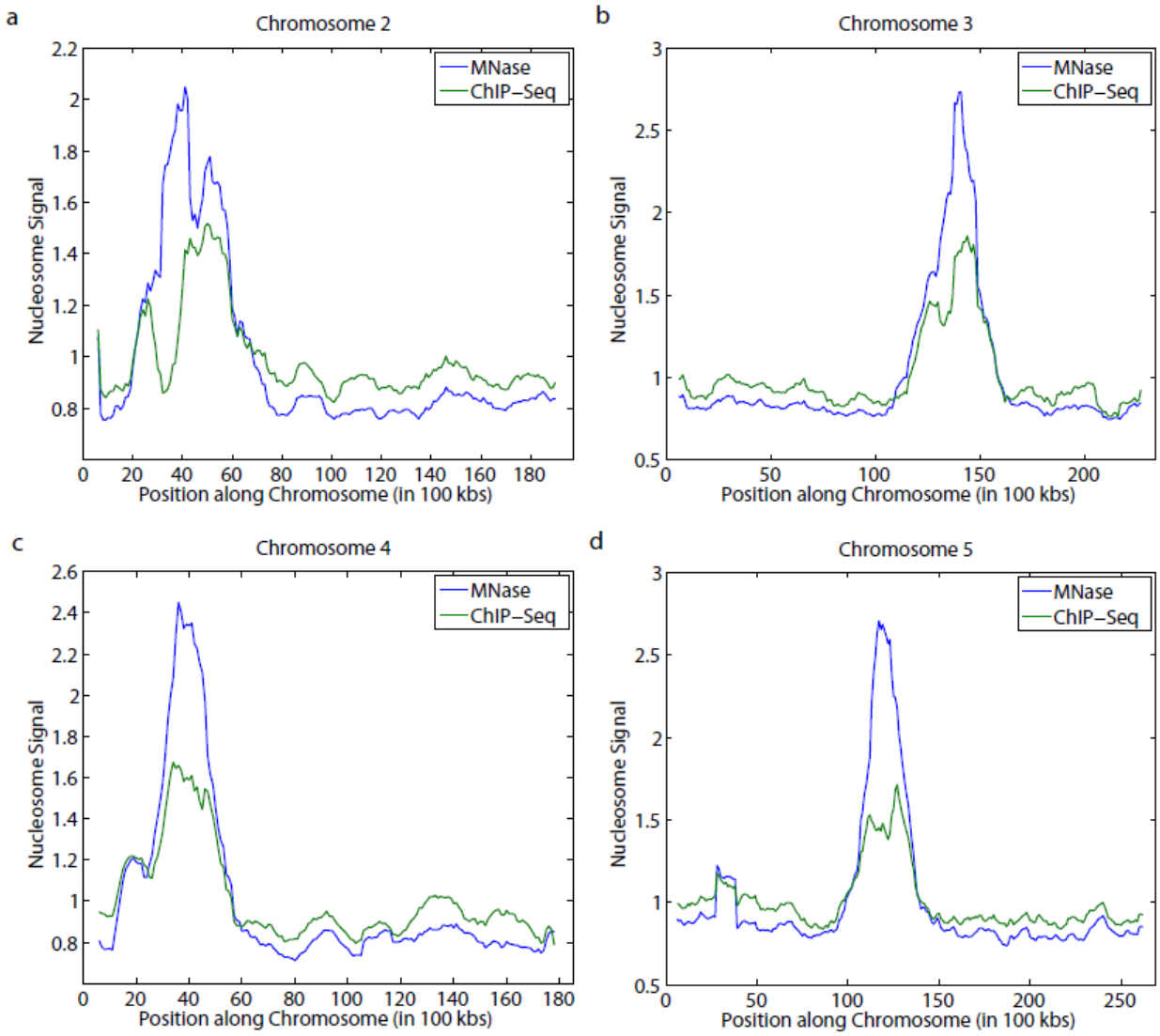
Supplementary Figure 23 | Methylation profiles of the highest scoring and lowest scoring predicted nucleosome start sites.

a-f, Predicted nucleosomes were separated into high and low scoring subsets. CG **(a, d)**, CHG **(b, e)**, and CHH **(c, f)** methylation was profiled across these sites. These results confirm our findings from MNase data. Moreover, high scoring nucleosomes have higher CG methylation over nucleosomes and low scoring nucleosomes have lower CG methylation over nucleosomes.

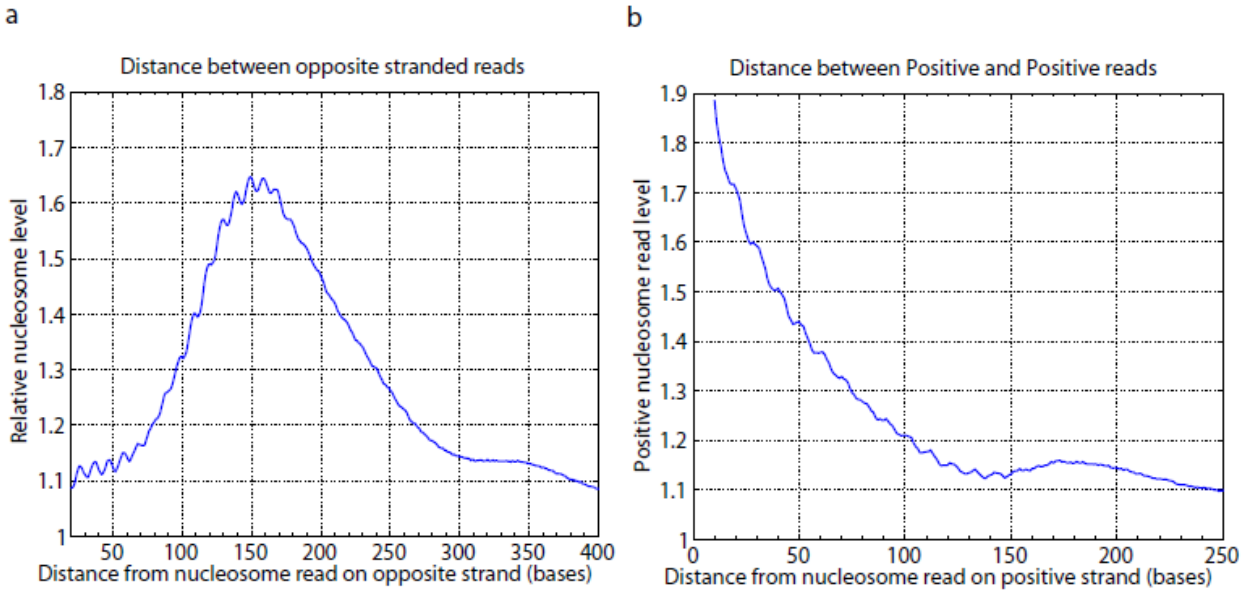
Supplementary Figure 24 | Chromosomal view of predicted nucleosome start sites

a-e Predicted nucleosome start sites in 100 kb bins along chromosomes 1 **(a)**, 2 **(b)**, 3 **(c)**, 4 **(d)**, and 5 **(e)** are shown.

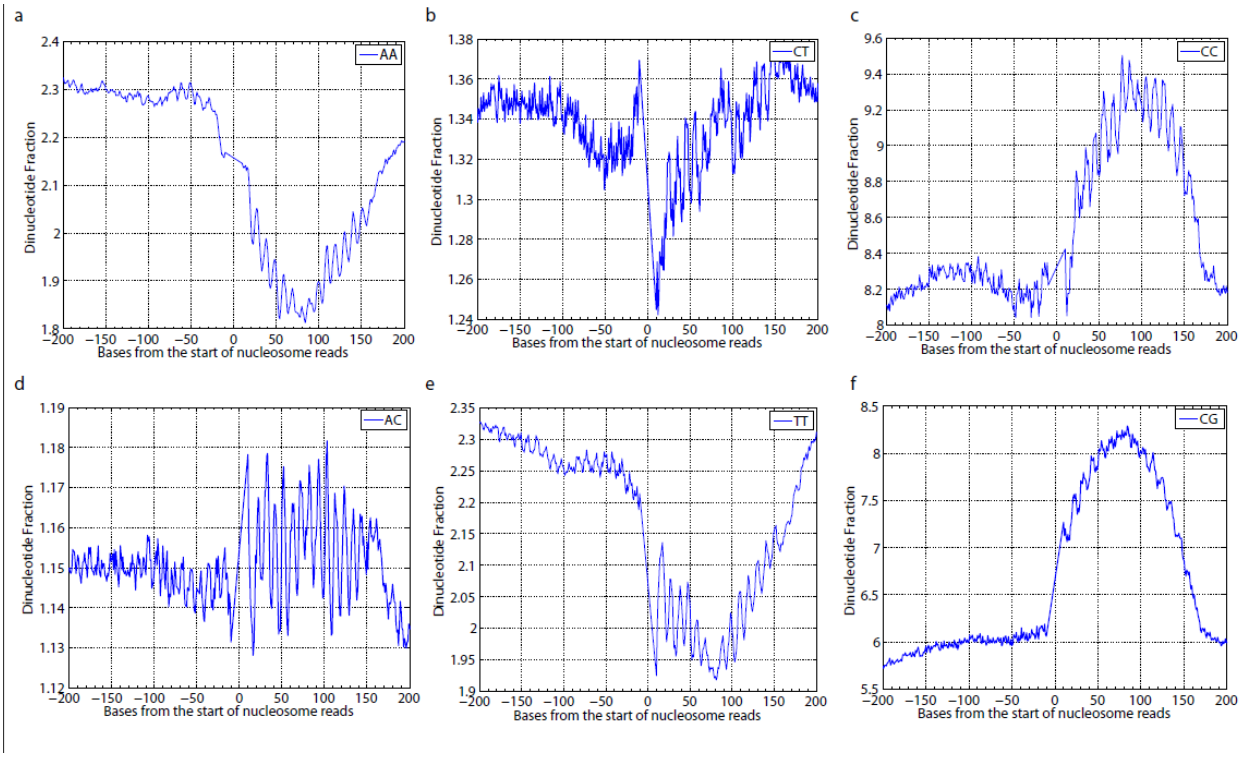
Supplementary Figure 1



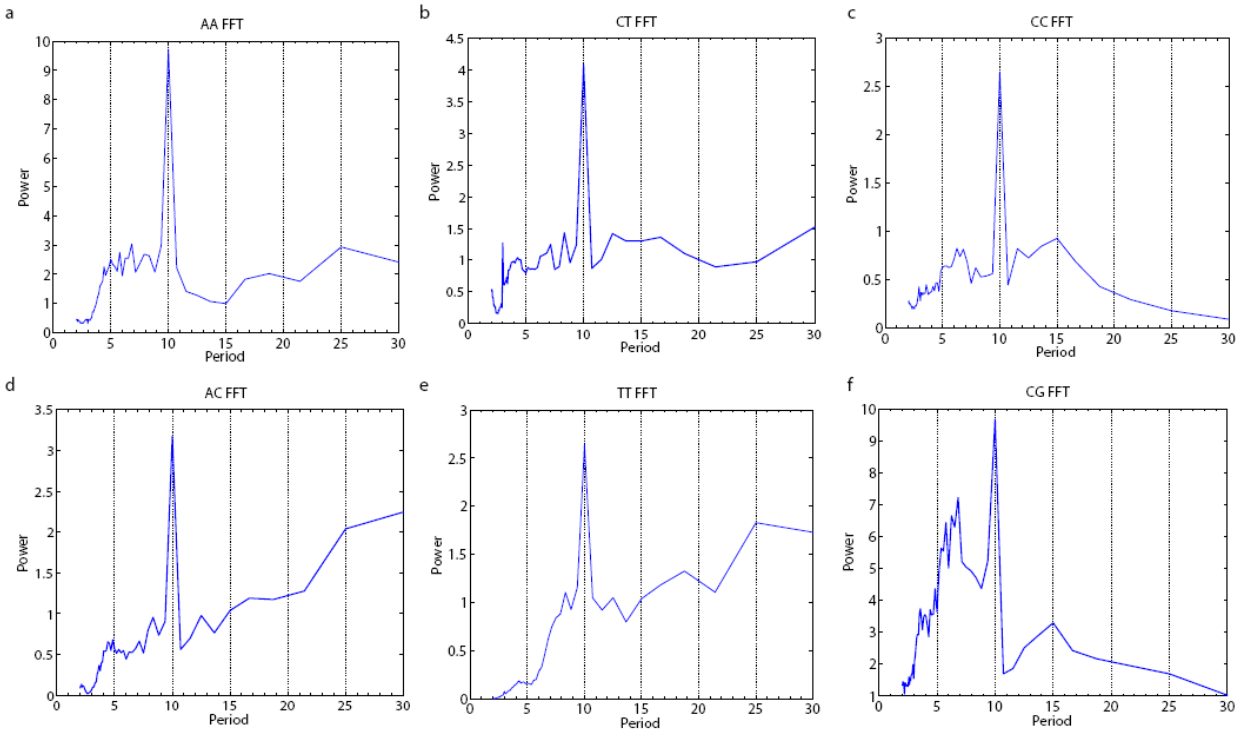
Supplementary Figure 2



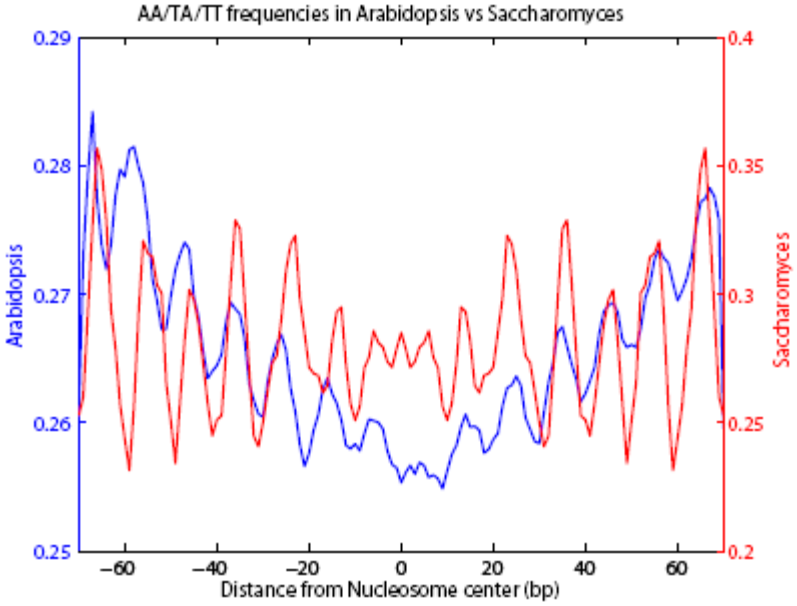
Supplementary Figure 3



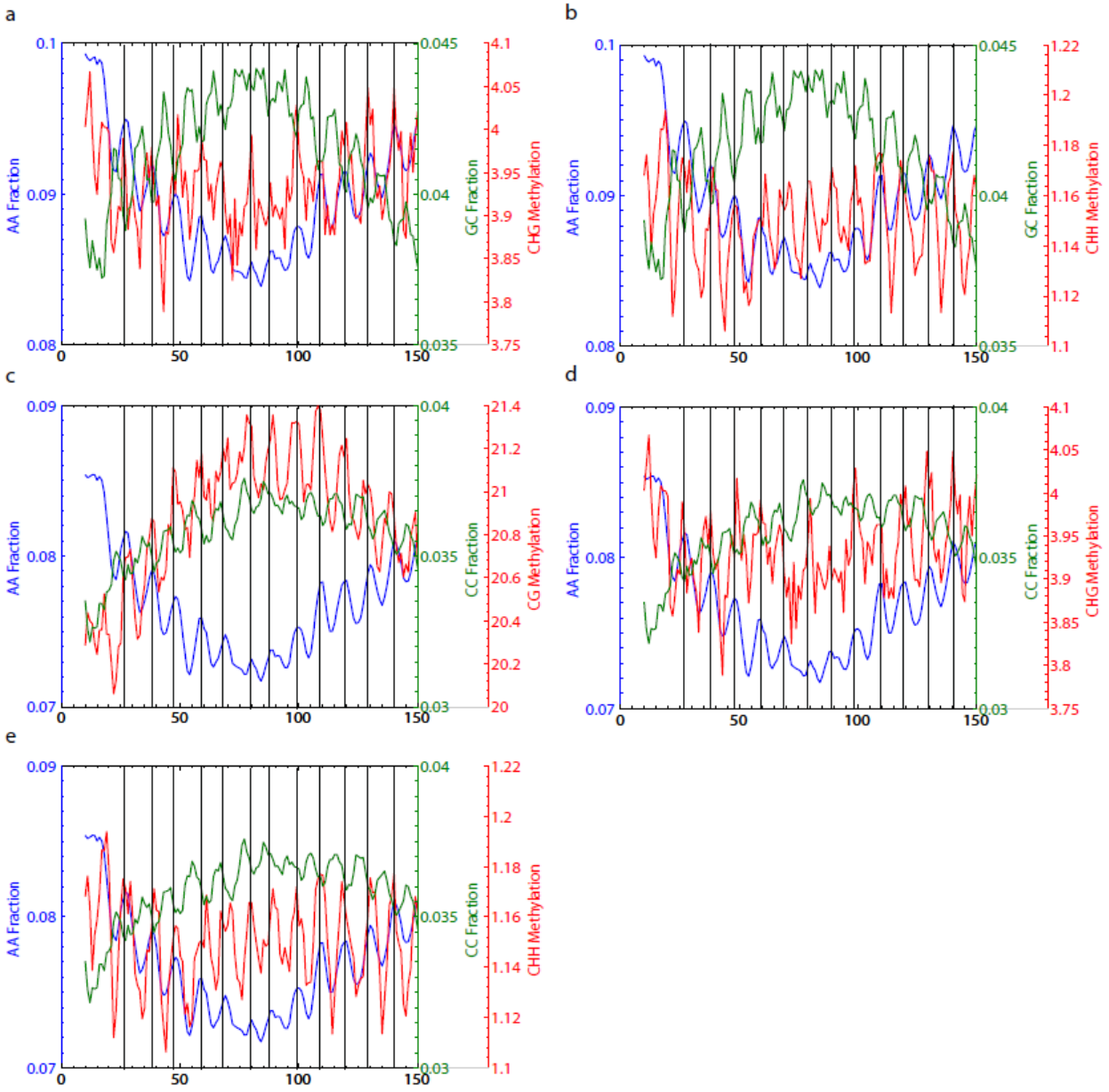
Supplementary Figure 4



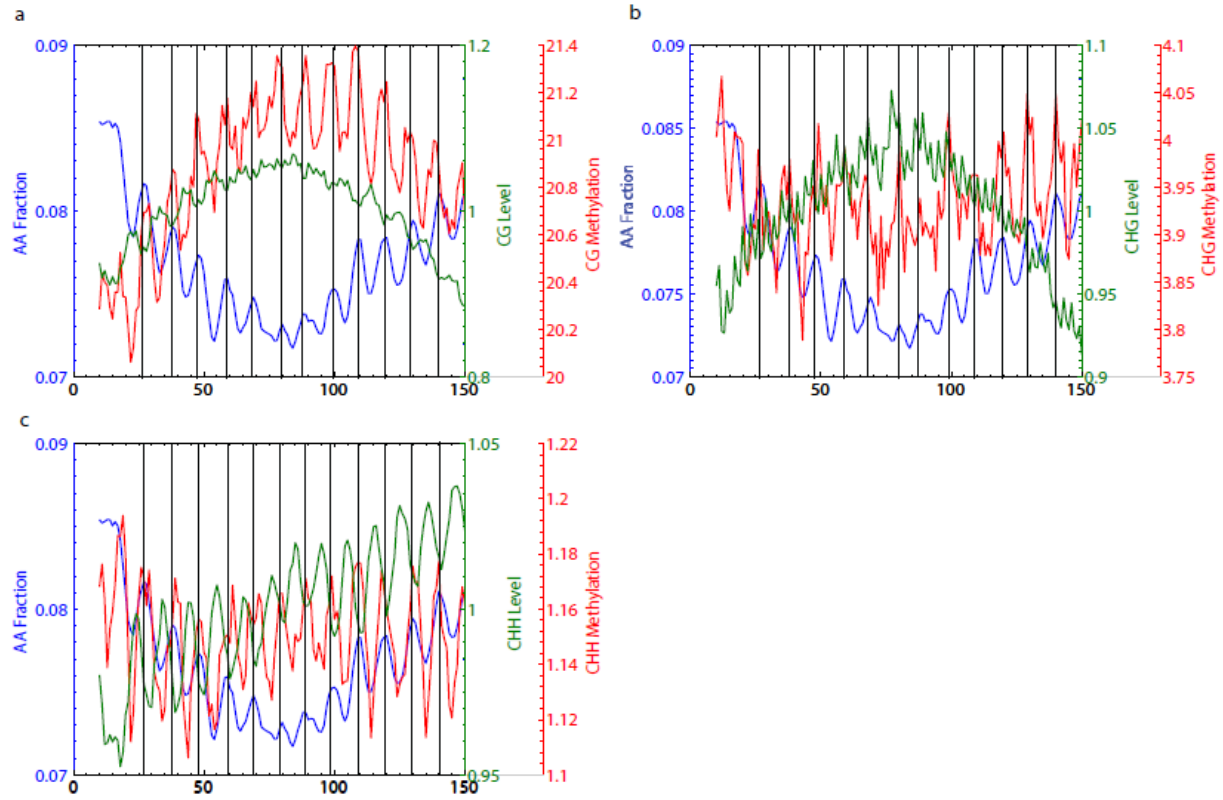
Supplementary Figure 5



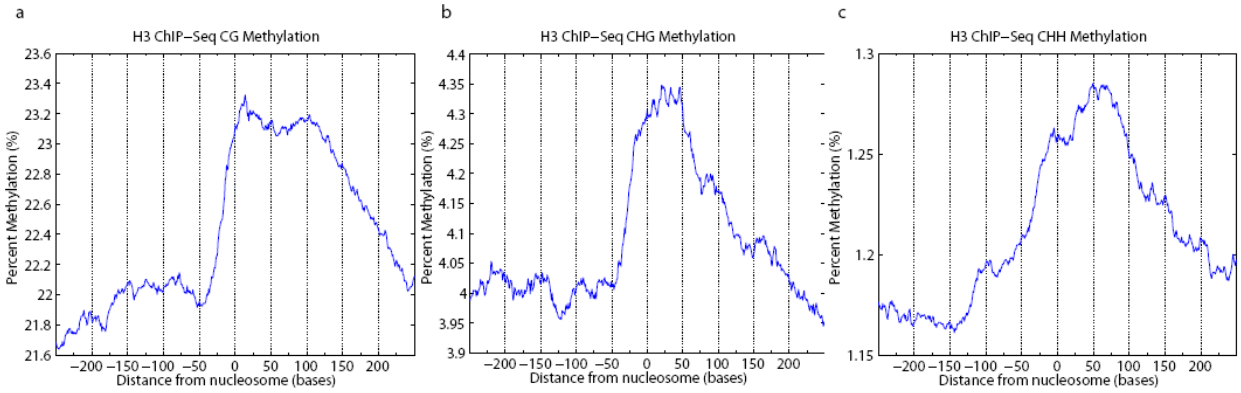
Supplementary Figure 6



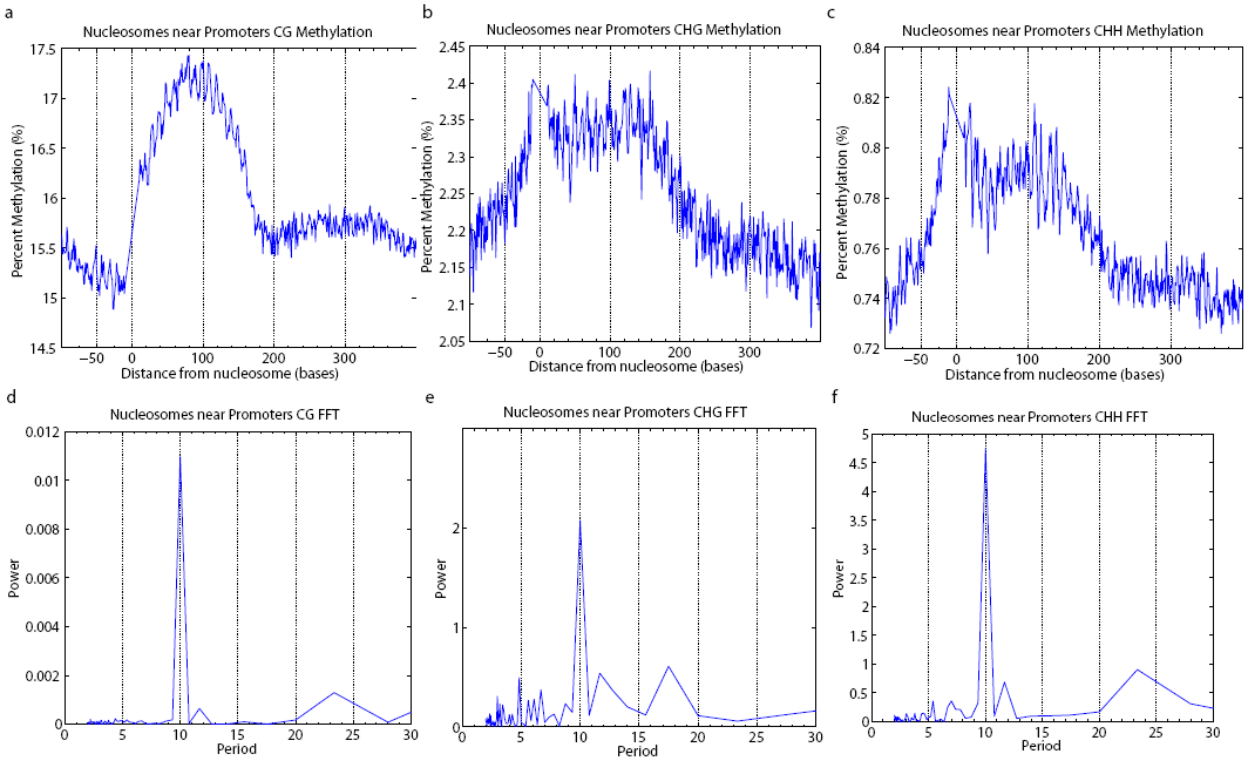
Supplementary Figure 7



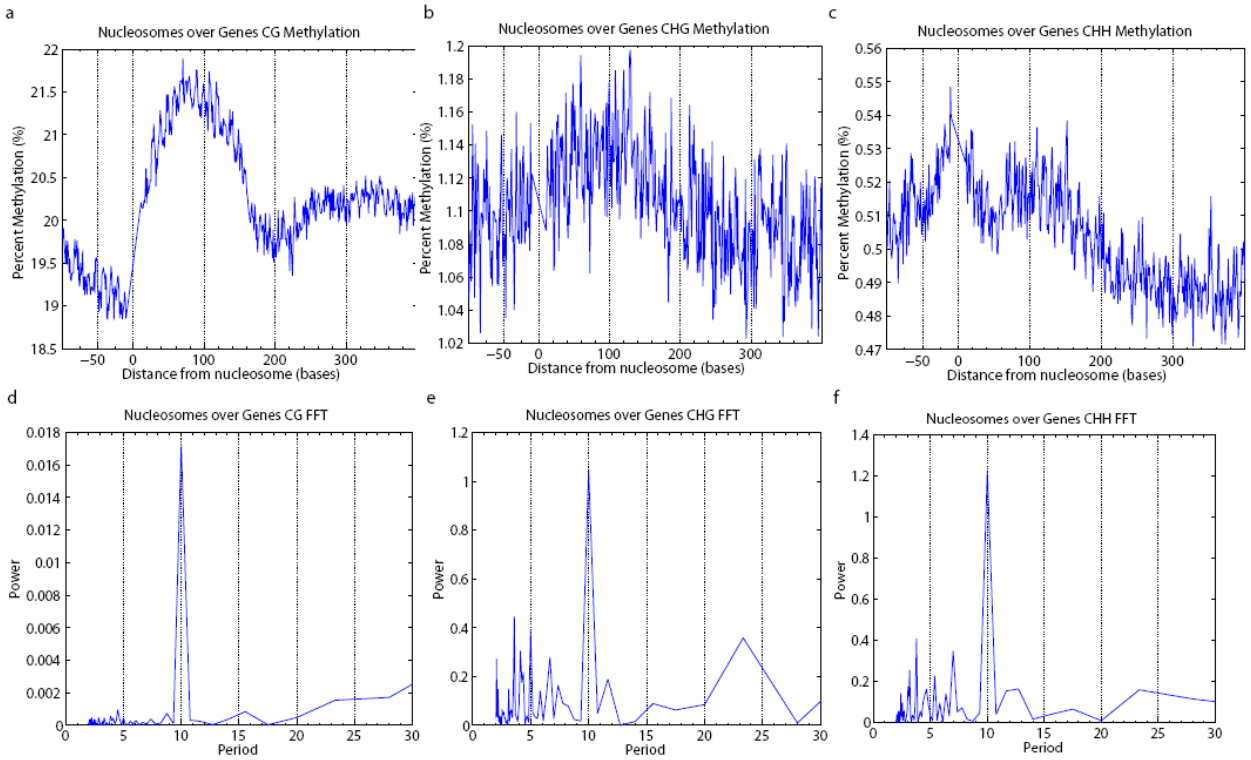
Supplementary Figure 8



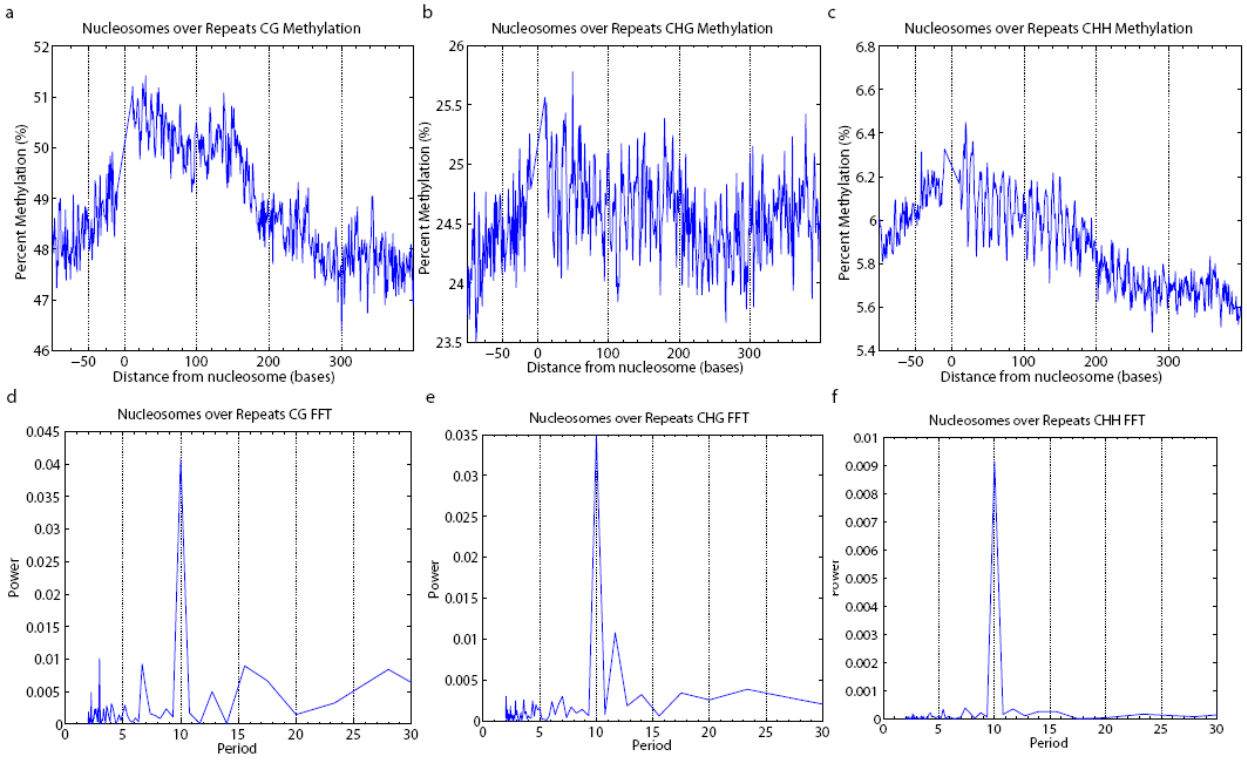
Supplementary Figure 9



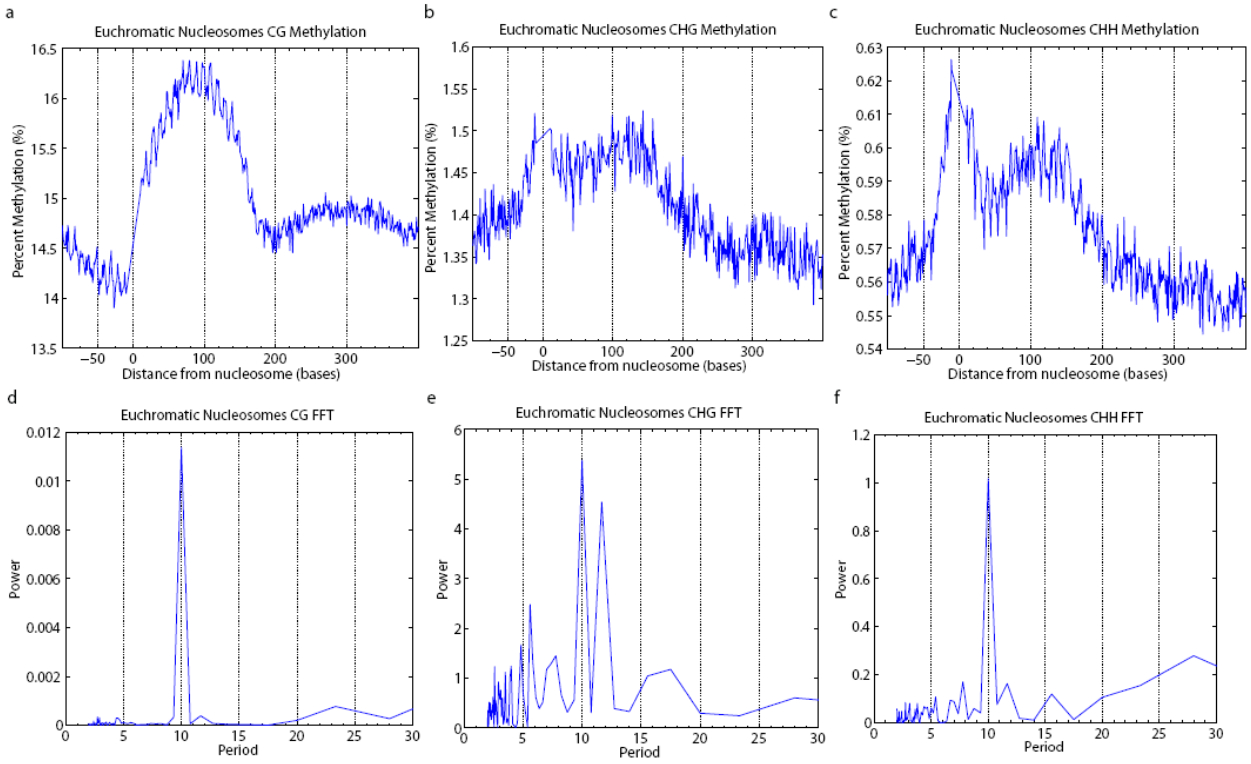
Supplementary Figure 10



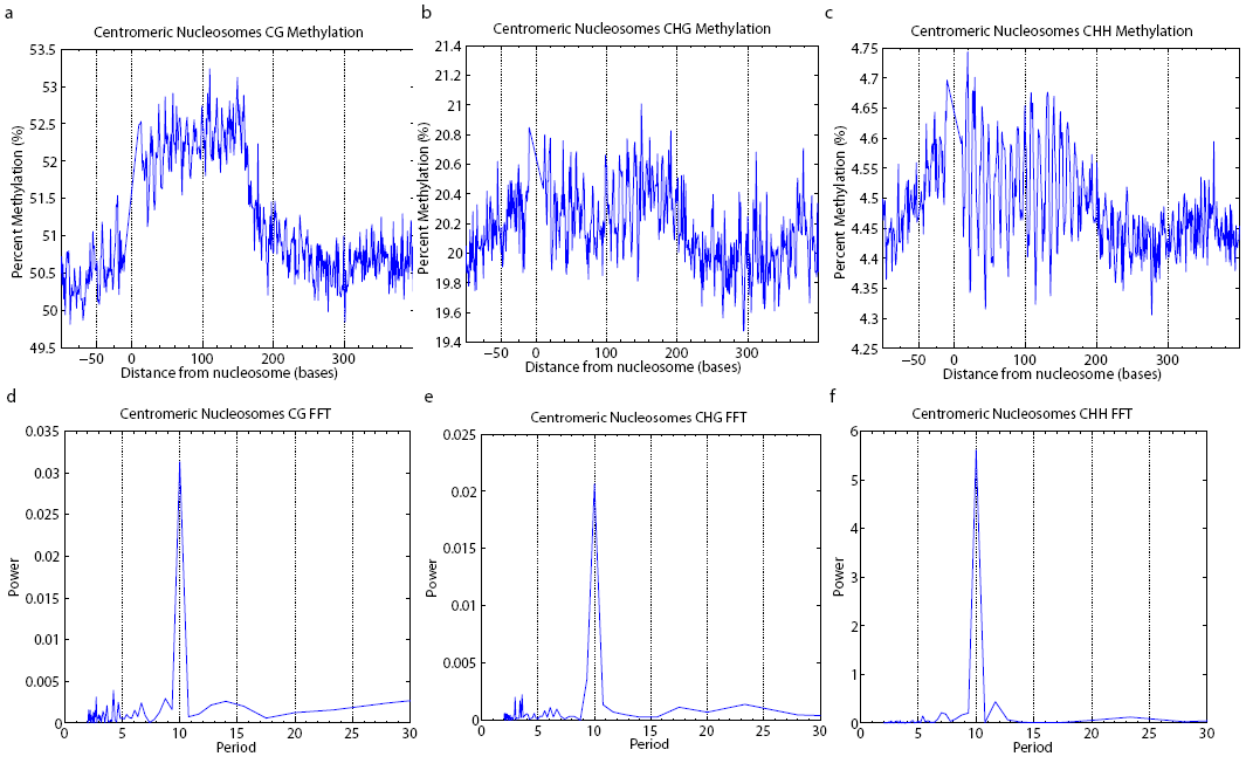
Supplementary Figure 11



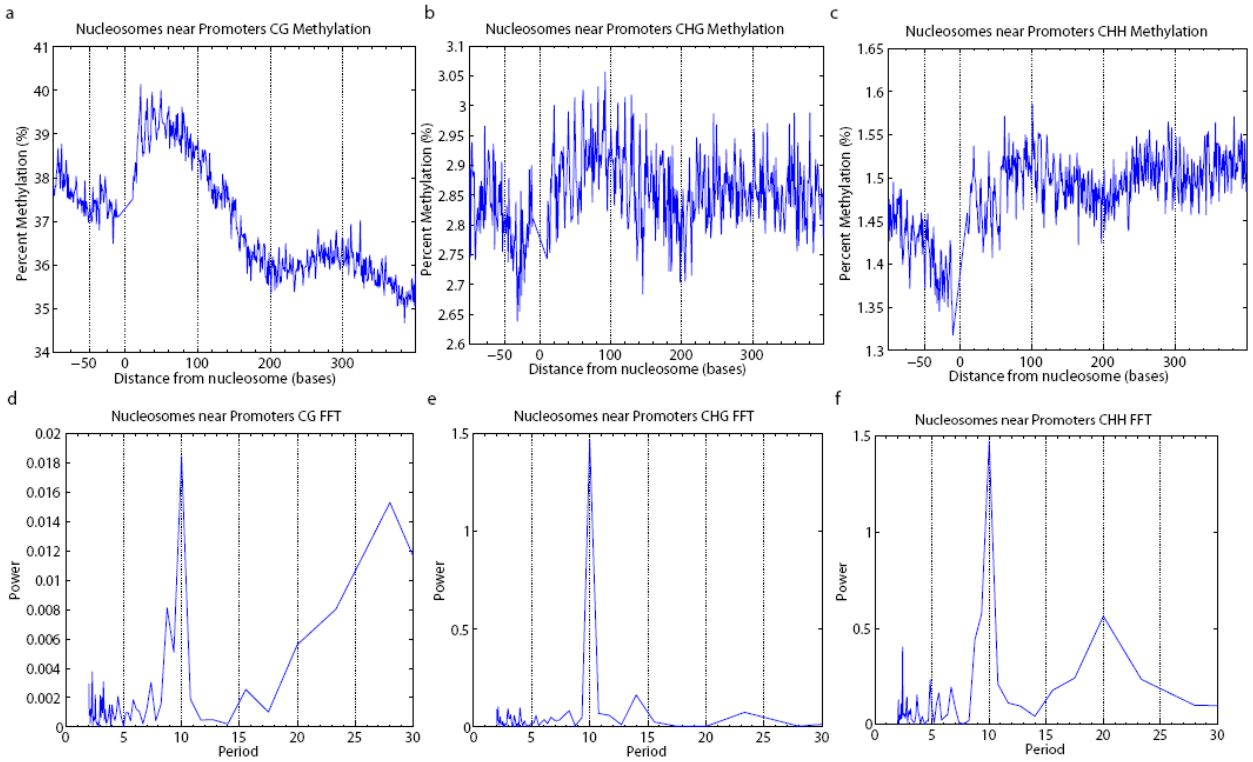
Supplementary Figure 12



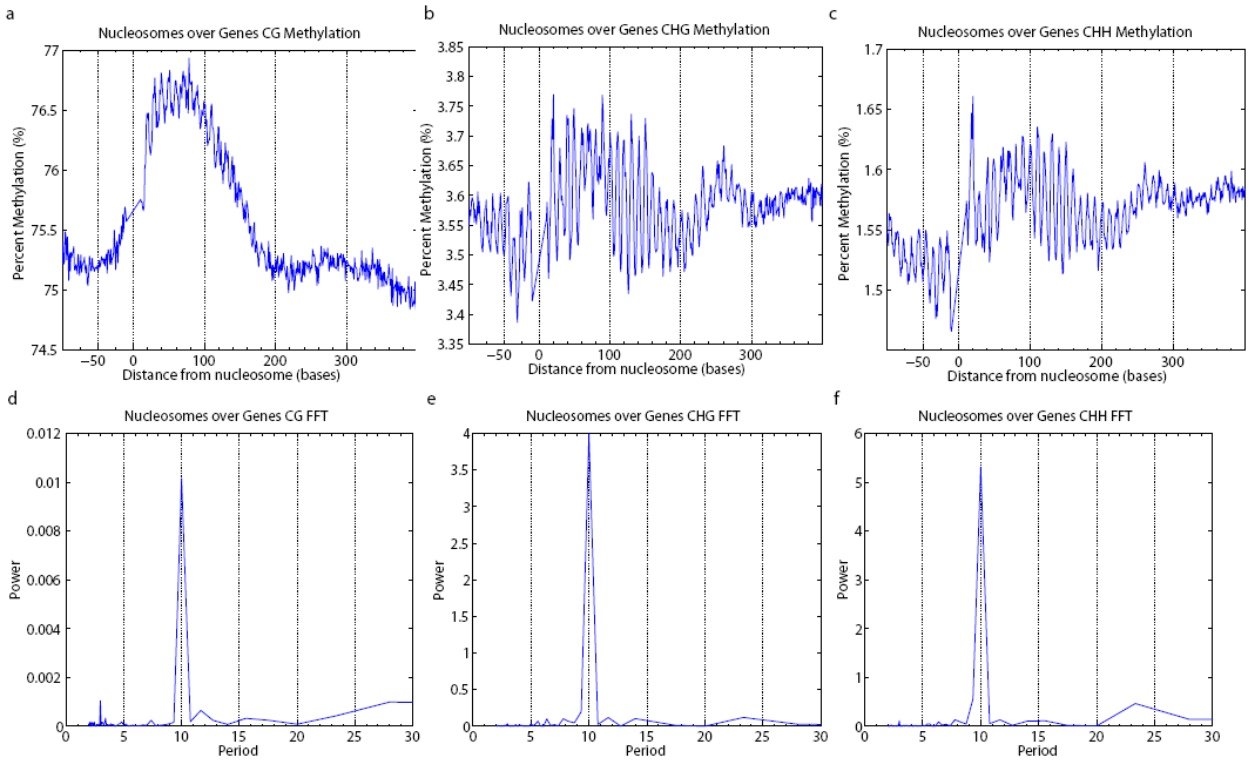
Supplementary Figure 13



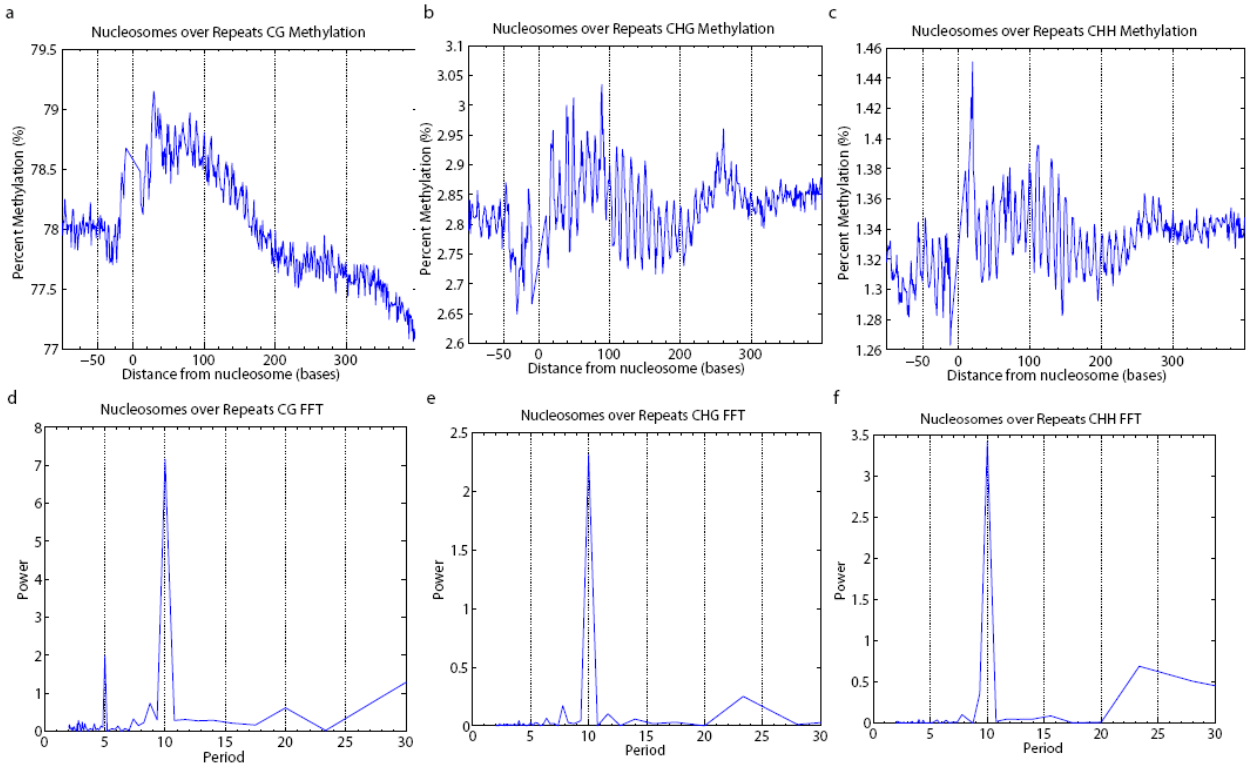
Supplementary Figure 14



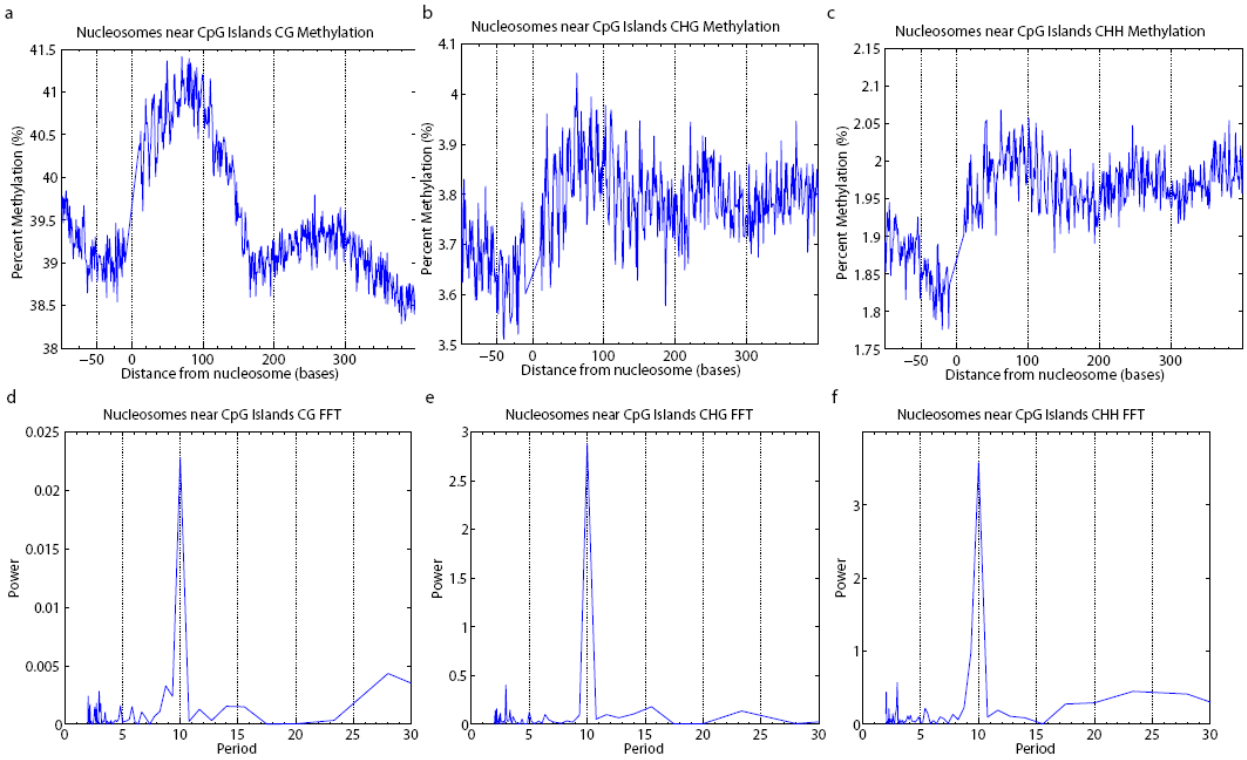
Supplementary Figure 15



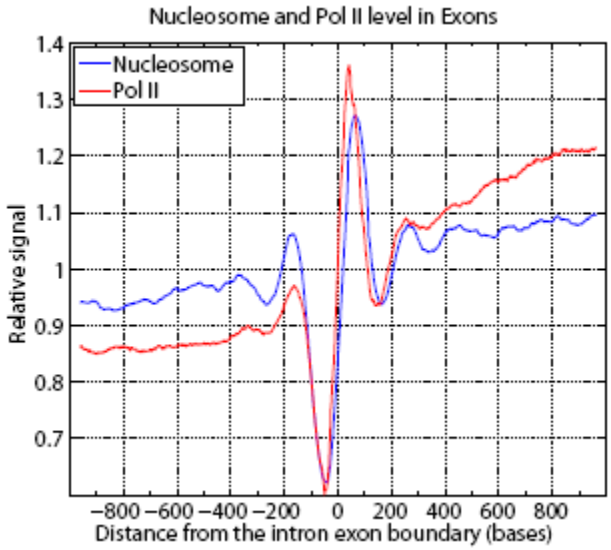
Supplementary Figure 16



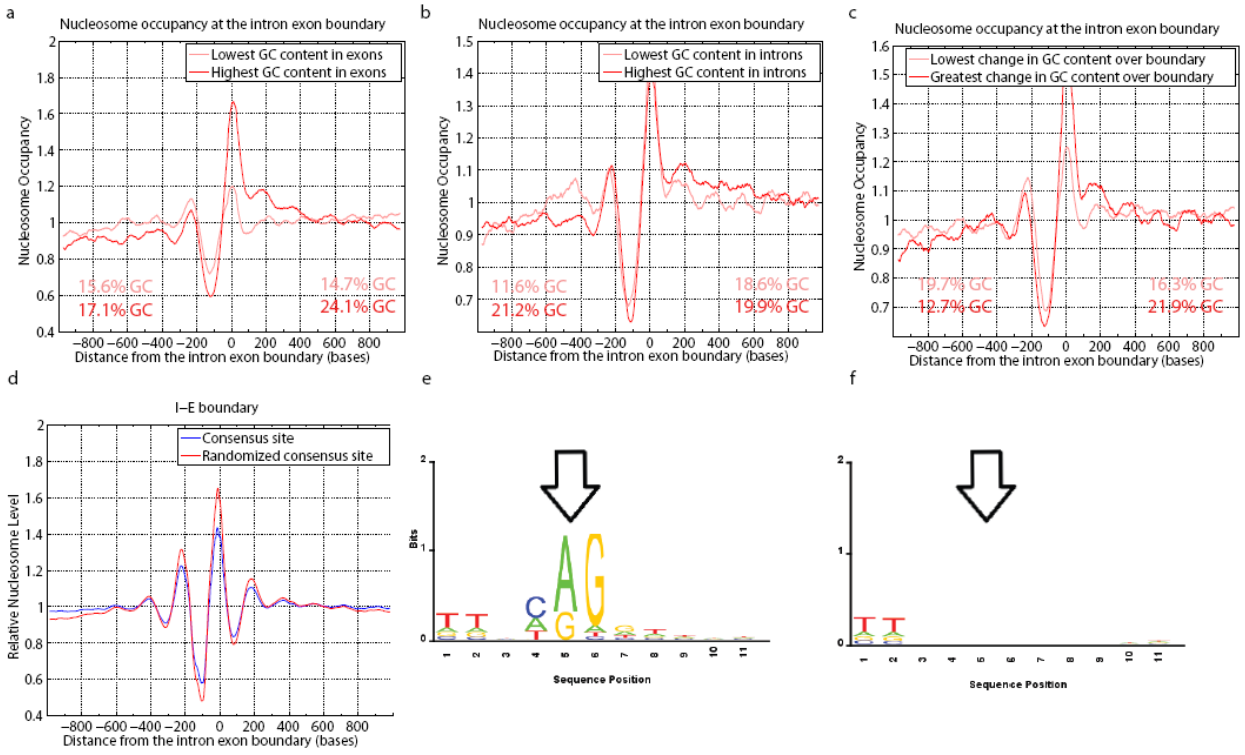
Supplementary Figure 17



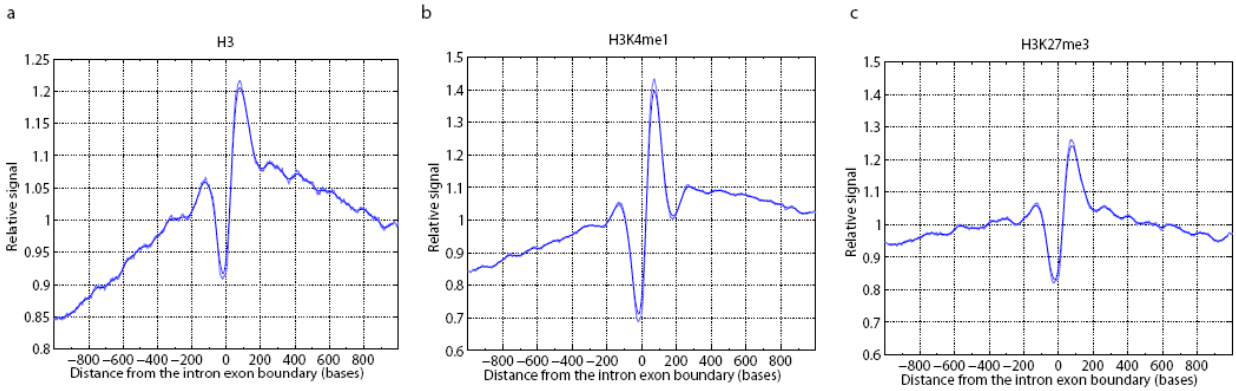
Supplementary Figure 18



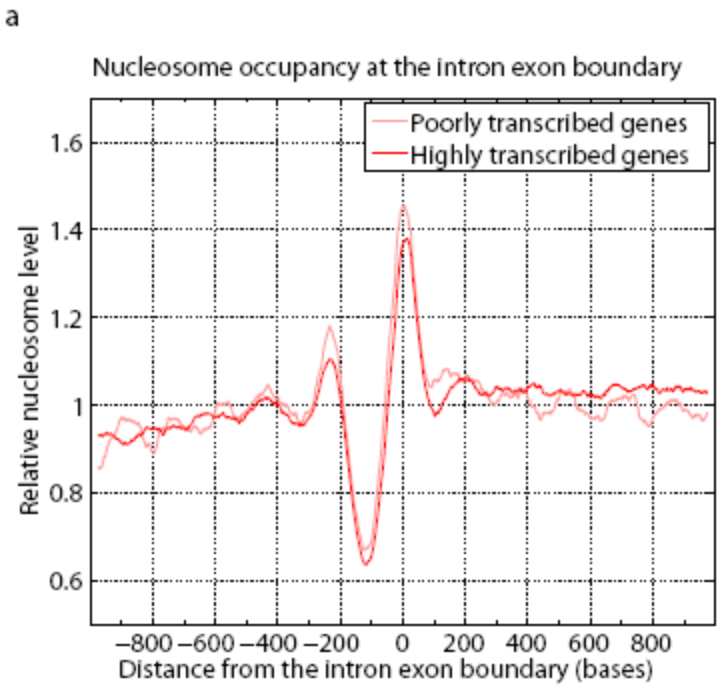
Supplementary Figure 19



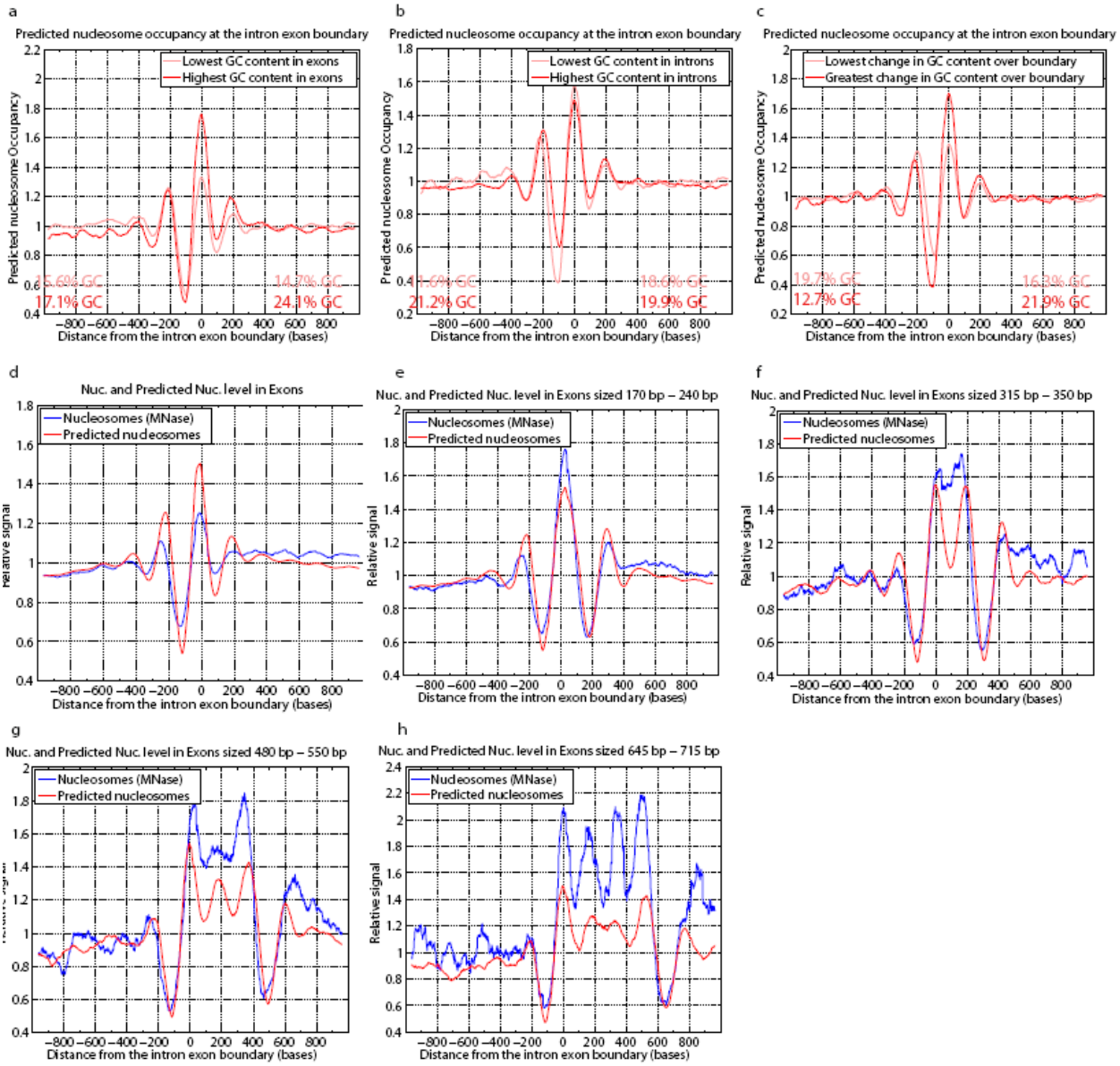
Supplementary Figure 20



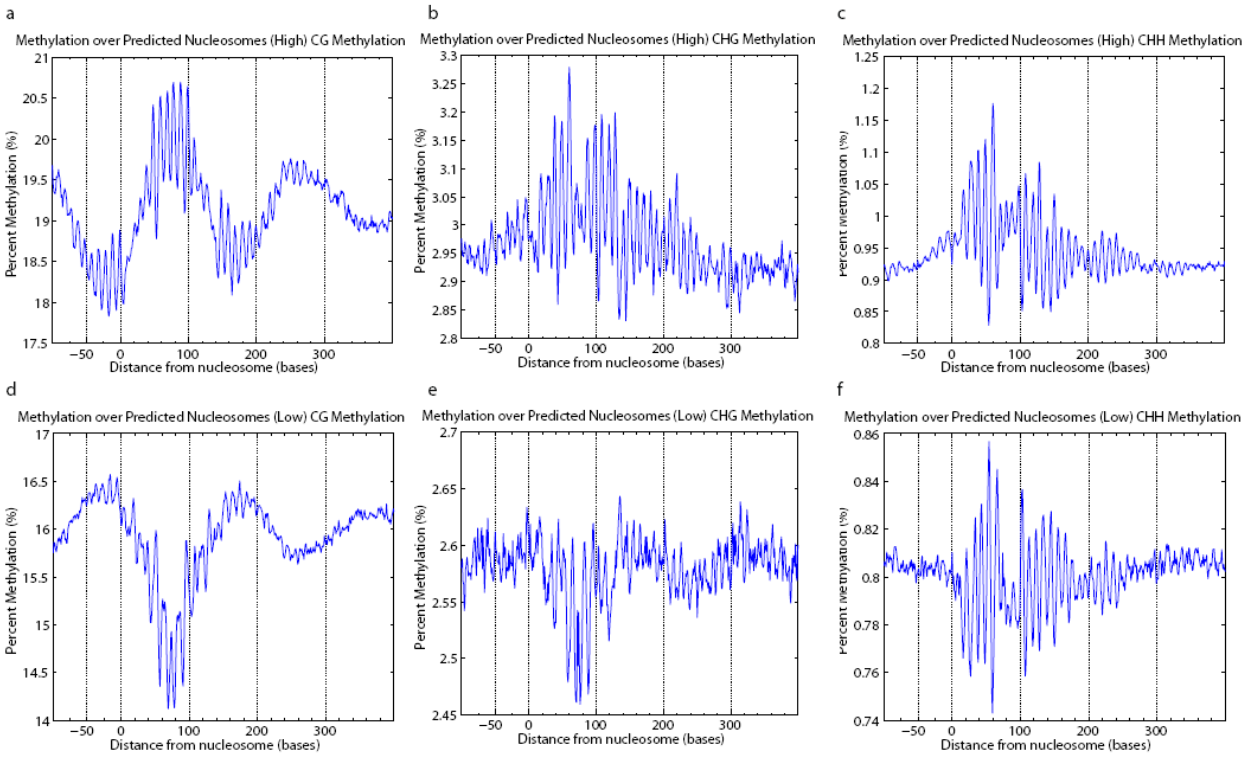
Supplementary Figure 21



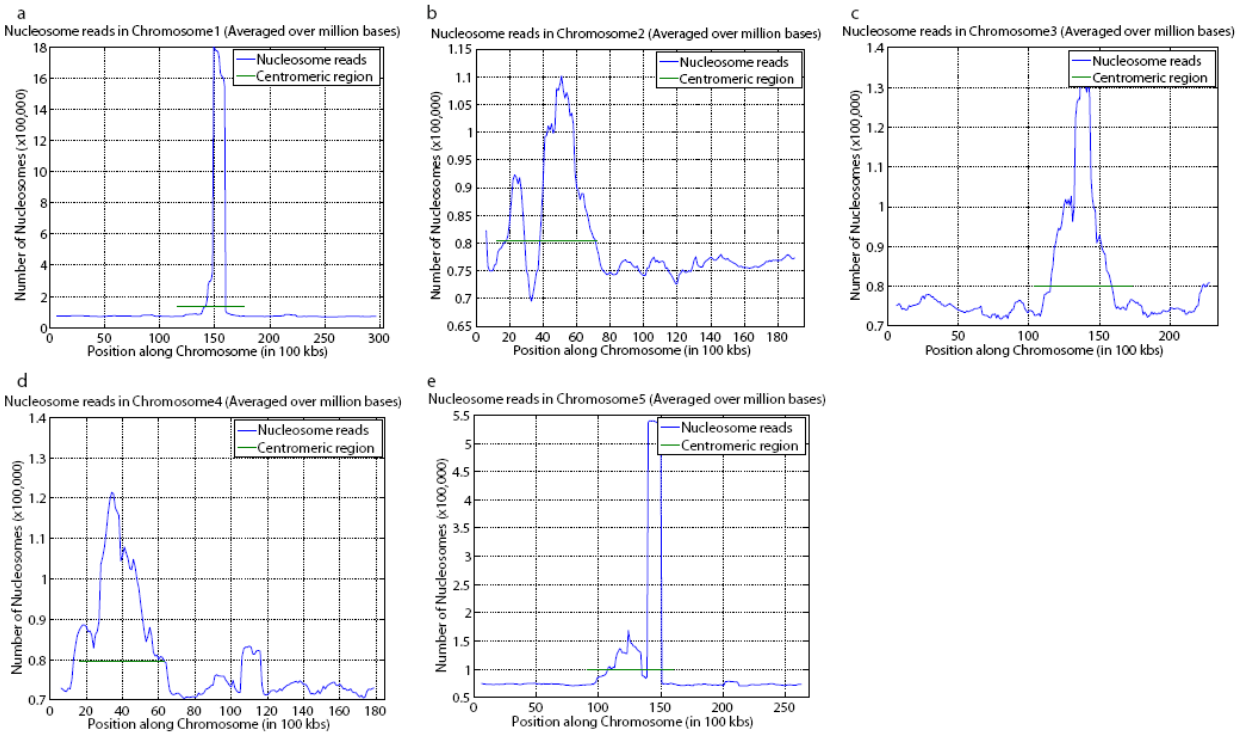
Supplementary Figure 22



Supplementary Figure 23



Supplementary Figure 24



Chapter 4: Transcriptome and methylome interactions in rice hybrids.

The following was first published in the journal PNAS on July 9th, 2012.

Ramakrishna K. Chodavarapu, Suhua Feng, Bo Ding, Stacey A. Simon, David Lopez, Yulin Jia, Guo-Liang Wang, Blake C. Meyers, Steven E. Jacobsen, and Matteo Pellegrini. (2012) Transcriptome and methylome interactions in rice hybrids. *Proc. Nat. Acad. Sci. U. S. A.*, 109: 12040-12045.

Transcriptome and methylome interactions in rice hybrids

Ramakrishna K. Chodavarapu[1], Suhua Feng[1,2], Bo Ding[4], Stacey A. Simon[5], David Lopez[1], Yulin Jia[6], Guo-Liang Wang[4], Blake C. Meyers[5], Steven E. Jacobsen[1,2,3,7*], and Matteo Pellegrini[1,3,7*]

9. Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, California, United States of America
10. Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, California, United States of America
11. Molecular Biology Institute, University of California Los Angeles, Los Angeles, California, United States of America
12. Department of Plant Pathology, Ohio State University, Columbus, Ohio, United States of America
13. Department of Plant and Soil Sciences, Delaware Biotechnology Institute, University of Delaware, Newark, Delaware, United States of America
14. USDA-ARS Dale Bumpers National Rice Research Center, Stuttgart, Arkansas, United States of America
15. Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research, University of California, Los Angeles, California, United States of America

*corresponding authors

Classification: Biological Sciences, Plant Biology, Biophysics and Computational Biology

Abstract

DNA methylation is a heritable epigenetic mark that controls gene expression, is responsive to environmental stresses and in plants is also may play a role in heterosis. To determine the degree to which DNA methylation is inherited in rice, and how it both influences and is affected by transcription, we performed genome-wide measurements of these patterns through an integrative analysis of BS-seq, RNA-seq, and siRNA-seq data in two inbred parents of the Nipponbare (NPB) and Indica (93-11) variety of rice and their hybrid offspring. We show that SNPs occur at a rate of about 1/253 bp between the two parents and that these are faithfully transmitted into the hybrids. We use the presence of these SNPs to reconstruct the two chromosomes in the hybrids according to their parental origin. We found that unlike genetic inheritance, epigenetic heritability is quite variable. Cytosines were found to be differentially methylated (epimutated) at a rate of 7.48% (1/15 cytosines) between the NPB and 93-11 parental strains. We also observed that 0.79% of cytosines were epimutated between the parent and corresponding hybrid chromosome. We found that these epimutations are often clustered on the chromosomes, with clusters representing 20% of all epimutations between parental ecotypes, and 2 to 5% in F1 plants. Epimutation clusters are also strongly associated with regions where the production of siRNA differs between parents. Finally, we identified genes with both allele-specific expression patterns that were strongly inherited as well as those differentially expressed between hybrids and the corresponding parental chromosome. We conclude that much of the misinheritance of expression levels is likely due to epimutations and *trans* effects.

Introduction

DNA methylation is an epigenetic mark that can often lead to the repression of gene expression (1, 2). It is enriched in heterochromatin and when present at regulatory sites, usually acts as a repressor of expression, most notably in transposons (3). However, it is also found over coding regions, where it likely does not directly affect transcription and is associated with moderately expressed genes(2, 4-6). In plants, DNA methylation occurs in three different contexts – CG, CHG, and CHH (where H is any nucleotide but G). In Arabidopsis, each context is maintained by different enzymes: MET1 for CG sites, CMT3 for CHG sites and DRM2 for CHH sites. CG and CHG sites are symmetric across the two DNA strands, which is thought to be important for the maintenance of methylation at these sites following DNA replication. In contrast, CHH sites are not symmetric, and their methylation is mediated by RNA directed DNA Methylation pathways (RdDM) which use siRNAs to initiate de novo methylation(3). Cellular methylation states tend to persist during cell division, and recent studies in Arabidopsis have also shown that DNA methylation is faithfully inherited across generations (7, 8). Nonetheless, we are only beginning to understand how different methylation patterns from inbred parents may 'interact' during the generation of their hybrid progeny (9, 10).

In this study, we generated integrative maps of whole genome cytosine methylation profiles (BS-Seq) and transcriptional profiles (RNA-seq), to characterize two rice subspecies, *Oryza sativa* spp japonica (Nipponbare) and *Oryza sativa* spp indica (93-11) and their two reciprocal hybrid offspring. Using a combination of BS-Seq, RNA-seq and sRNA-seq we were able to generate allele-specific patterns of methylation and transcription in the hybrids, and thus directly measure the degree to which these are altered between the corresponding parental and F1 chromosomes.

Results and Discussion

We sequenced bisulfite treated DNA from two cultivated subspecies of *Oryza sativa*, Nipponbare (NPB) and 93-11, as well as the F1 hybrids resulting from reciprocal crosses. We achieved coverage levels of 17x for NPB, 10x for 93-11, and about 25x for each of the two crosses (supporting information (SI) Table 1).

SNPs are much more common between varieties, and very infrequent across generations

Using the MSU Rice Genome v6.1 annotation (11) as a reference, we mapped the bisulfite converted reads to the genome using BS Seeker (12). From these alignments we determined single nucleotide polymorphisms (SNPs) between the NPB and 93-11 parental plants. Although there is ambiguity due to cytosine conversion in bisulfite treated DNA, we can distinguish conversions from mutations by analyzing the sequence of the complimentary base of each cytosine, and as a result, were able to call all possible types of SNPs. For our SNP calls we required that we had at least three reads on each strand and over 90% agreement between them before making a call (Fig. 1 *a*). Using these SNP calls we reconstructed two genomes, one for each parent. We then mapped F1 reads against these inferred genomes to identify the parental chromosome from which they were derived.

We found one SNP every 253 base pairs between the two parental strains (Fig. 1 *b*), representing a level of divergence that is similar to previous *japonica-indica* comparisons (13). Not surprisingly, regions enriched with genes had lower densities of SNPs. Of these SNPs, the most frequent were C to T changes, or their complements corresponding to a G to A mutations on the opposite strand. Combined, these types of mutations accounted for 73% of the SNPs that were found between varieties (Fig. 1 *c* and SI Table 2), which represents a three-fold enrichment compared to the expected value if all types of mutations were equally likely. We also observe that methylated cytosines mutate more than 3 times more frequently than non-methylated cytosines (Fig. 1 *d*), and they most often mutate to thymines. This

is consistent with previous observations that the deamination of methylcytosine, which results in thymine, occurs at higher rates than other spontaneous mutations (14-16).

Next, we compared mutation rates between the parental strains and the hybrids. Not surprisingly, we found that SNPs were nearly absent across generations. If, as is likely, the few SNPs we call are false positives, we estimate the false discovery rate of our SNP calls to be less than 0.0003%.

We also found a small number of SNPs between the reference MSU *O. sativa japonica* genome and our assembled version (1/90651 bp, 2,487 total) (SI Table 2), and considerably more SNPs by comparing our assembled version of the 93-11 genome with the reference Bejing Genome Institute's *O. sativa* ssp. *indica* genome (1/51 bp), consistent with the fact that the 93-11 assembly is not as complete as that of the NPB cultivar (17). The fact that we only align reads that map to a unique position in the genome, have at most three mismatches with the reference genome, do not permit insertions or deletions, and require a minimum of 3 reads on both strands, allowed us to only call bases over about half of the genome in both NPB and 93-11, and thus we are reporting only a subset of all true SNPs.

Epimutations occur frequently and in clusters between NPB and 93-11

Using the BS-Seq reads, we were able to quantitatively identify the methylation status of cytosines in the parental genomes. By partitioning the F1 reads between their NPB or 93-11 chromosomes, we were also able to identify the allele specific methylation status of cytosines in F1 plants. However, because the partitioning of reads between NPB and 93-11 parents can only occur at polymorphic sites, we were only able to do this for 6.2% of all cytosines with sufficient coverage.

When comparing sites where each library had similar levels of coverage, we found that NPB had a higher level of methylation than 93-11 at CG and CHG sites, but that 93-11 had more methylation at CHH sites (Fig. 2 *a*). As one might expect, on a global level, the hybrids had methylation levels that were half way

in between the parental methylation levels. However, the methylation of individual chromosomes in the hybrids closely matched that of the corresponding parent.

We used a binomial based approach to determine whether sites were differentially methylated between samples, and refer to these events as 'epimutations'. As discussed in detail in the methods section, this was accomplished by requiring that the methylation level in both samples be outside of the 95% confidence interval of methylation of each other (SI Fig. 1). Further, we required that the methylation difference was greater than a threshold, defined as the average methylation level for each particular sequence context. In order to prevent coverage biases, we also restricted our analysis only to those sites that had comparable coverage in both samples.

Overall, we found that the 'epimutation' rate was 7% (1/15 cytosines) between the NPB and 93-11 parental strains. Since cytosines represent about 20% of all bases, this leads to an overall epimutation rate of one in every 75 bases, which is considerably higher than the SNP rate of one in every 253 bases. The epimutation rates were higher at CG (17%) and CHG (11%) sites, compared to CHH (3%) sites.

The epimutations between the parental strains are likely driven by both the cis genetic differences between the two cultivars as well as stochastic effects. To determine the rates of epimutation within the same local genetic background, we compared the methylation of the parental strains to the corresponding chromosomes in the hybrids. Here we found that only about 0.8% (1/126) of cytosines have altered methylation (Fig. 2 *b*, SI Table 3). Among these sites we found that CHG epimutations were the most common. Other studies have found CHG sites are associated with silent chromatin and H3K9 dimethylation, indicating that the repression of these sites may be variable between the parent and F1 plants (18).

We next asked whether epimutation events are isolated and independent, or whether they tend to occur in clusters. We searched for 100bp sized windows in which epimutations were enriched. Enriched

windows were defined as those having at least five, and a minimum of 20% of all cytosines that are differentially methylated between two samples. Using this approach we identified 16,865 epimutation clusters between NPB and 93-11. An example of such a cluster is shown in Fig 2 *c*. Of these clusters, 5,152 had epimutations with significantly greater methylation in 93-11 and 8,584 with greater methylation in NPB.

In contrast, there were fewer clusters of differential methylation between parents and F1 plants. We identified an average of 340 clusters meeting our criteria between NPB and the NPB chromosomes in the two F1 hybrids, and 75 between 93-11 and the 93-11 chromosomes in the two F1 hybrids. A reduced occurrence of clustered epimutations is expected since there are far fewer epimutations between parents and F1s than between the two parents. However, we found that as a fraction of the total, transgenerational epimuations were less clustered than those between the parents, with clusters representing 20% of all epimutations between parents, but only 5% in NPB F1 plants and 2% in 93-11 F1 plants.

Intriguingly, we found that a significant number (~45%) of transgenerational epimutations (between parent and F1 plants) were also identified as epimutations between the two parental varieties (Fig. 2 *b*). Conversely, about 15% of the epimutations that were found between parental cultivars were also found to be epimutated across generations. This suggests that regions of discordant methylation between the two inbred parents are often aberrantly inherited in the F1 plants, a result that is consistent with recent findings in Arabidopsis hybrids (10). We found similar results for epimutation clusters, as 20% of those found in NPB F1 plants and nearly 35% of the clusters found in 93-11 F1 plants had some overlap with NPB-93-11 epimutation clusters.

We next asked whether epimutation clusters are correlated with siRNAs, and whether this could explain the overlap we observe between inter-parental and trangenerational epimutations. The RNA-directed

DNA methylation (RdDM) pathway is associated with feedback loops between the production of siRNAs and the recruitment of DRM methyltransferases (3, 19). However, the activation of this feedback is likely dependent on the dosage of siRNAs. As a result, we expect that some of the epimutation clusters are associated with differential siRNA production between the parental cultivars. At these loci the hybrid plants would produce siRNAs at a level that is intermediate between the two parents. In some cases this level may be sufficient to methylate both chromosomes in the hybrid, leading to the hypermethylation of the locus with respect to the unmethylated parent. Conversely, in other cases, the siRNA dosage may be insufficient to methylate the locus in either chromosome, leading to the hypomethylation of the locus with respect to the hypermethylated parent.

To understand the relationship between epimutation clusters and siRNAs we sequenced small RNAs from the parental and F1 samples. From the sequenced reads we selected 24 nt small RNAs that mapped uniquely to the genome, as we expect these to be enriched for siRNAs instead of miRNAs (20, 21). We found that 29% of the epimutation clusters had siRNAs associated with them. We also observed that epimutation clusters that were hypermethylated in NPB versus 93-11, also had more siRNAs in NPB versus 93-11. That is, within a cluster there was a strong correlation (rho = 0.7) between the fraction of epimutations where methylation was increased in NPB and the fraction of siRNAs belonging to NPB (Fig. 2 *d*). Due to the paucity of SNPs in 24 nucleotide siRNAs we were unable to assign a sufficient number of siRNAs uniquely to the parental chromosomes, and thus were unable to analyze parent specific siRNA levels in the hybrids.

Genes expressed in an allele-specific manner are enriched for epimutations

To study the inheritance of gene expression levels in the hybrid strains we measured transcript levels using RNA-seq. We obtained RNA-seq data from the same samples that were used for methylation profiling. Reads were mapped using TopHat (22) and in the case of F1 plants, we assigned them to individual chromosomes using the SNP data discussed above. Based on the counts of SNP containing

reads, we computed the allele-specific expression levels in the F1 plants. We also used this data to estimate expression levels in both parents.

We first compared transcription between the NPB and 93-11 varieties. We defined genes as differentially expressed if the difference in their counts was at least 100 RPKM and had a P value below 10^{-5} (based on a Poisson model) and they had at least a two-fold change. We found 1525 transcripts that met these criteria (Fig. 3 *a,* SI Table 4 and SI Table 5).

To assess the association between methylation and transcription we measured both methylation and epimutations within transcribed loci and their immediate 1 kb upstream region. In order to find genes that were enriched for differential methylation, we used a binomial based test to determine whether a significant number of sites were differentially methylated. We found that 491 out of 25,640 (SI Table 6 and SI Table 7) genes had significant enrichment for interparental epimutations (p value < 0.01). Of these 45 were also differentially expressed, indicating that there is a significant enrichment for differential methylation among the differentially expressed genes (p=0.0024).

We next set out to determine how transcriptional differences between the parents were inherited in the hybrid plants, by analyzing RNA-seq reads corresponding to SNPs. We calculated F1 allele specific expression estimates and used a t-test to identify genes that were expressed in an allele specific manner between the parents as well as between the corresponding chromosomes in the F1 hybrids. As in our previous comparison, we also required that the expression fold change between the two alleles be at least two fold and have a difference of at least 100 RPKM. We found 85 genes that fit these criteria (SI Table 8), and thus had significantly different expression levels between NPB and 93-11 chromosomes.

Allele-specific expression is potentially mediated by both genetic and epigenetic factors. We found that most (79%) genes had at least 1 SNP between NPB and 93-11, and 37% had more than 5 SNPs. In contrast, genes that were expressed in an allele-specific manner had a significantly higher SNP rate (Fig.

3 *b*). We also found that of the 85 genes that were expressed in an allele-specific manner, a significant number, 15, were also differentially methylated ($p = 10^{-26}$). A plot of epimutation densities over allele-specific genes also shows that the level of epimutations tends to rise significantly over the body of these genes (Fig. 3 *c*). Thus we observe that the genetic differences that underlie allele-specific genes also correlates with dramatic differences in their methylation state, and that the combination of these two affects is likely mediating the allele-specific expression differences we observed.

To disentangle genetic and epigenetic factors that mediate expression changes, we also looked for genes that were differentially expressed between the parents and their corresponding chromosomes in the hybrid. We found 1619 genes that had transgenerational differential expression in at least one parent-F1 comparison using the same criteria described above. Of these, we found only a single case that was both differentially expressed and differentially methylated. Since most of these genes have similar methylation profiles and identical genetic backgrounds, their transcriptional changes may be mediated by *trans* effects in the hybrids. The fact that this group of genes is much larger than the allele-specific genes, suggests that these *trans* effects dominate the genetically and epigenetically mediated allele-specific expression phenotypes.

Conclusion

While it is thought that DNA methylation patterns are to a large degree heritable in plants, the frequency of epimutations and their underlying causes are still poorly understood. By sequencing both parental and hybrid strains we were able to measure the frequency of transgenerational epimutations and found that they occur in about one in every 125 cytosines across genetically identical chromosomes. This shows that DNA methylation is in fact quite variable across generations, and suggests that this variability across inbred strains may have important phenotypic consequences.

There are two primary mechanisms that could account for changes in DNA methylation across generations. The first are stochastic effects that may accrue as somatic cells in the parents lead to germ line cells and eventually somatic cells of their progeny. The second cause of epigenetic variability may be due to *trans* effects, as the heterozygous hybrids are genetically different from the inbred parents. Specifically, we suggest that these *trans* effects may be in part mediated by RdDM pathways that mediate the interaction between the production of siRNAs and DNA methylation. Differences in siRNA production between parents at specific loci, would lead to the inheritance of average levels of siRNAs in the progeny, which could in turn lead to either the hyper- or hypo-methylation of the two alleles, depending on whether the dosage is sufficient or insufficient to activate RdDM.

It is likely that both these mechanisms are at play in the F1 hybrids we have analyzed. The fact that epimutations are clustered between parents suggests that siRNAs are likely mediating these differences, and we find a strong association between siRNA dosage differences and epimutation levels in our data. However, since the fraction of transgenerational epimutation clusters is quite low, stochastic effects must also be playing a role in the transmission of DNA methylation patterns across generations.

To understand whether epimutations have observable molecular phenotypes, we also examined transcription using RNA sequencing. We found that genes that are expressed in a strongly allele-specific manner across samples are enriched both for genetic and epigenetic differences, suggesting that these

two effects may be acting cooperatively to maintain these patterns. However, loci showing transgenerational gene expression changes across identical chromosomes are not significantly enriched for epimutations, suggesting that *trans* effects may be mediating the bulk of these changes. However, a single epimutation within a key cis-regulatory site may be sufficient to disrupt the expression of a gene, and therefore it is not possible to rule out stochastic epimutations as a significant driver of transgenerational transcriptional variability.

The two varieties of rice that we studied demonstrate significant heterosis, a phenomenon whereby matings by more distant plants, intra- or interspecies, result in offspring with superior agronomic traits such as tiller number and dry weight (9, 23-25). Previous studies of heterosis have revealed possible mechanisms to account for the increased vigor, but the mechanisms that underlie this complex phenotype are still poorly understood, and the role of DNA methylation in heterosis is unclear (26-29). Our results suggest that the majority of gene expression changes in hybrids are not associated with cis acting DNA methylation changes, and instead indicate that trans effects may play mediate the majority of the transcriptional differences in hybrid offspring. However, it is possible that a subset of the gene expression changes may also be due to intergenerational epimutations in the hybrids.

In conclusion, we have mapped the degree to which DNA methylation varies between parental and hybrid strains. These observations reinforce the notion of epigenetic plasticity and suggest that it may play important roles in both environmental adaptation as well as hybrid vigor. In the future it will be important to elucidate the magnitude and molecular details of these mechanisms.

Materials and Methods

Plant materials and growth conditions

Oryza sativa ssp. *japonica* (Nipponbare, NPB), *Oryza sativa* ssp. *indica* (93-11) and their reciprocal cross F1s (NPB × 93-11 and 93-11× NPB) were used in this study. Rice seeds were surface sterilized with 40% (V/V) sodium hypochlorite solution and transferred on 1/2 Murashige and Skoog medium. After germination, rice seedlings were transplanted into soil and grown at 26□/20□ under a 10h light/14h dark cycle in growth chamber. Fully expanded leaves from 6-week-old plants were collected for library construction.

Library Construction

BS-seq libraries were made from genomic DNA isolated from leaf tissues of NPB, 93-11 and their two reciprocal hybrid offspring, 93-11 X NPB and NPB X 93-11 by a previously published method using pre-methylated Illumina adapters (30).

RNA-seq libraries were made from total RNA isolated from leaf tissues using TRIzol reagent (Invitrogen, Life Technologies, USA). Total RNA (10 ug) for each sample were used for purifying the poly-A containing mRNA molecules, RNA amplification and synthesis of double stranded cDNAs that were ligated to adapters, thus preparing the libraries for sequencing on the Illumina GAII. The RNA-Seq library preparation protocol was followed as described in Illumina's standard protocol for RNA-seq libraries (Illumina, San Diego, CA). Libraries were sequenced on an Illumina GAII at the Delaware Biotechnology Institute (Newark, DE). This generated 74,171,650 total reads from NPB, 71, 535, 918 total reads from Oryza sativa indica (93-11), 74,999,214 total reads from the reciprocal hybrid offspring 93-11 X NPB and 78,592,119 total reads from the reciprocal hybrid offspring NPB X 93-11. The mRNA sequence data are available from NCBI's Gene Expression Omnibus (GEO) and are accessible via GEO Series accession number XXXXXX.

Small RNA libraries were constructed from total RNA isolated from leaf tissues using TRIzol reagent

(Invitrogen, Life Technologies, USA). The sRNA library preparation protocol is based on Illumina's sRNA Sequencing Sample Preparation Guide (Illumina, San Diego, CA). The libraries were sequenced on an Illumina GAII at the Delaware Biotechnology Institute (Newark, DE). The sRNA sequence data are available from NCBI's Gene Expression Omnibus (GEO) and are accessible via GEO Series accession number XXXXXX.

Data processing and Analysis

Bisulfite sequenced reads from all libraries were mapped to the *Oryza sativa* ssp. *japonica* (cv Nipponbare) v6.1 reference genome (11) using BS-seeker (12) allowing for up to three mismatches. Reads that mapped with equal efficiency to multiple sites were discarded. Coverage was about 20x for NPB and 12x for 93-11 and about 25x for the hybrids. Using these reads as a scaffold, custom NPB and 93-11 genomes specific to our samples were built. Genotypes were called only when coverage consisted of at least three reads per strand and over 90% of the reads were in agreement. These custom genomes were used to call SNPs.

DNA methylation was calculated as the percent of cytosines that failed to undergo bisulfite conversion. DNA methylation context was determined based on the sequence in NPB. For most analyses, the sites which were considered was limited to those that had comparable coverage levels in all libraries which was defined to be at least 5, and not more than 25 reads.

Epimutations were determined by finding sites where the implied methylation level of each sample was outside of the 5% confidence intervals of the other sample based on a binomial distribution. The difference in implied methylation levels also needed to be greater than a threshold based on the mean context specific methylation (SI Figure 1).

Transcripts enriched for differential methylation were identified by using a binomial test. The inputs for this test were the number of epimutations present, the number of cytosines in the gene with sufficient

coverage in the respective libraries, and the overall epimutation rate for the genome. Transcripts with P values below .01 were considered to be enriched for epimutations.

RNA-seq reads were mapped using the default parameters of TopHat (22). Coverage was about 20x in NPB, 15x in 93-11 and about 12x in the two hybrids. Differential expression was determined using a Poisson test. Starting with the number of reads per kilobase per million reads (RPKM) of one sample, we use the Poisson test to compute the probability of the observing the RPKM values in the other sample and vice versa. If either yielded a P value below 10^{-5} and there was at least a 2 fold change and an absolute difference of at least 100 RPKM, the gene was called differentially expressed.

siRNA reads were processed in order to trim the adapter sequence and then were mapped using Bowtie allowing for up to three mismatches and requiring matches to be unique (31). Matches were further size selected so that only 24 nt reads were kept.

Acknowledgements

We thank Jonathan Hetzel and William Wong for their assistance with sequencing. R.K.C. is supported by NIH training grant GM07104. S.F. is a Special Fellow of the Leukemia & Lymphoma Society. S.E.J. is an Investigator of the Howard Hughes Medical Institute. This project was supported by the United States National Science Foundation Plant Genome Research Program #0701745.

References

1. Bird A (2007) Perceptions of epigenetics. *Nature* 447(7143):396-398.
2. Suzuki MM & Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics* 9(6):465-476.
3. Law JA & Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* 11(3):204-220.
4. Zhang X*, et al.* (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* 126(6):1189-1201.
5. Zilberman D, Gehring M, Tran RK, Ballinger T, & Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics* 39(1):61-69.
6. Cokus SJ*, et al.* (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452(7184):215-219.
7. Schmitz RJ*, et al.* (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* 334(6054):369-373.
8. Becker C*, et al.* (2011) Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* 480(7376):245-249.
9. He G*, et al.* (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *The Plant cell* 22(1):17-33.
10. Greaves IK*, et al.* (2012) Trans Chromosomal Methylation in Arabidopsis hybrids. *Proceedings of the National Academy of Sciences of the United States of America* 109(9):3570-3575.
11. Ouyang S*, et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic acids research* 35(Database issue):D883-887.
12. Chen PY, Cokus SJ, & Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC bioinformatics* 11:203.
13. Nasu S*, et al.* (2002) Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (Oryza sativa, Oryza rufipogon) and establishment of SNP markers. *DNA research : an international journal for rapid publication of reports on genes and genomes* 9(5):163-171.
14. Yebra MJ & Bhagwat AS (1995) A cytosine methyltransferase converts 5-methylcytosine in DNA to thymine. *Biochemistry* 34(45):14752-14757.
15. Bransteitter R, Pham P, Scharff MD, & Goodman MF (2003) Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proceedings of the National Academy of Sciences of the United States of America* 100(7):4102-4107.
16. Morgan HD, Dean W, Coker HA, Reik W, & Petersen-Mahrt SK (2004) Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *The Journal of biological chemistry* 279(50):52353-52360.
17. Yu J*, et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296(5565):79-92.
18. Bernatavichute YV, Zhang X, Cokus S, Pellegrini M, & Jacobsen SE (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PloS one* 3(9):e3156.
19. Wassenegger M (2005) The role of the RNAi machinery in heterochromatin formation. *Cell* 122(1):13-16.
20. Zhang X, Henderson IR, Lu C, Green PJ, & Jacobsen SE (2007) Role of RNA polymerase IV in plant small RNA metabolism. *Proceedings of the National Academy of Sciences of the United States of America* 104(11):4536-4541.

21. Kasschau KD*, et al.* (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS biology* 5(3):e57.
22. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.
23. Birchler JA, Auger DL, & Riddle NC (2003) In search of the molecular basis of heterosis. *The Plant cell* 15(10):2236-2239.
24. Birchler JA, Yao H, Chudalayandi S, Vaiman D, & Veitia RA (2010) Heterosis. *The Plant cell* 22(7):2105-2112.
25. Zhang HY*, et al.* (2008) A genome-wide transcription analysis reveals a close correlation of promoter INDEL polymorphism and heterotic gene expression in rice hybrids. *Molecular plant* 1(5):720-731.
26. Song GS*, et al.* (2010) Comparative transcriptional profiling and preliminary study on heterosis mechanism of super-hybrid rice. *Molecular plant* 3(6):1012-1025.
27. Hochholdinger F & Hoecker N (2007) Towards the molecular basis of heterosis. *Trends in plant science* 12(9):427-432.
28. Ni Z*, et al.* (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature* 457(7227):327-331.
29. Springer NM & Stupar RM (2007) Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome Res* 17(3):264-275.
30. Feng S, Rubbi L, Jacobsen SE, & Pellegrini M (2011) Determining DNA methylation profiles using sequencing. *Methods in molecular biology* 733:223-238.
31. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10(3):R25.

Figure Legends

Figure 1: SNPs between NPB and 93-11

(a) Identification of SNPs required a minimum of three reads per strand in each parental strain. Although there is ambiguity due to cytosine conversion in bisulfite treated DNA, we can resolve this by considering the sequence of the reverse strand, and as a result, are able to call all possible types of SNPs. In the example, X indicates the location of the SNP and reads represent, in order, NPB forward strand, NPB reverse strand, 93-11 forward strand, 93-11 reverse strand. (*b*) SNP distribution over the genome is plotted on the bottom in windows of 10000 bases. Each color change represents a new chromosome. The top line represents the density of genes. (*c*) The fraction of the various types of SNPs found in NPB versus 93-11 is shown with the CT and GA SNPs grouped together and they represent nearly 74% of all SNPs. (*d*) The relative frequency of each SNP between NPB and 93-11 is shown per 1000 bases. Methyl-Cs convert to Ts at a rate that is nearly 5 times that of unmethylated Cs.

Figure 2: Epimutations

(*a*) NPB has slightly higher methylation levels at CG and CHG sites, while 93-11 has higher methylation levels at CHH sites. The methylation level of individual hybrid chromosomes is relatively unchanged compared to their parents. (*b*) The percentage of differentially methylated sites in each comparison in C (black), CG (red), CHG (green), and CHH (blue) contexts. The dark red fraction of each bar indicates the percentage of epimutations that were also divergent between the parents. For the 93-11 vs NPB comparison, the red portion indicates the percentage that was overlapping in any of the transgenerational comparisons. (c) A typical epimutation cluster showing increased methylation in 93-11 and nearly absent methylation in NPB. There is also a region of differential methylation in the 93-11 chromosome of a hybrid and the 93-11 parent that cannot be attributed to a lack of coverage. 9xN and Nx9 represent the reciprocal crosses between NPB and 9311, and |N and|9 represent the NPB and 9311

chromosomes within the hybrids. (*d*) Among epimutation clusters, as the fraction of methylation sites that are greater in NPB increases, so does the fraction of siRNAs that are found in NPB versus 93-11.

Figure 3: Transcript level changes

(*a*) Venn diagram showing the number of genes differentially methylated or differentially expressed between NPB and 93-11. (*b*) A metaplot of SNP density is displayed. Each gene was divided into 100 bins and the average SNP density is plotted in 100 bins upstream of the gene to 100 bins downstream of the gene. In allele specific genes, SNP density increases over the genic region, whereas for all genes, it decreases. (*c*) Genes are similarly broken into bins as in (*b*) and the epimutation density is plotted over all genes and allele specific genes. As in the case of SNP density, the epimutation density increases over allele specific genes compared to all genes.

SI Fig. 1 Epimutation Identification

The identification of epimutations (differentially methylated sites) requires that the implied methylation level of each sample be outside of the 5% confidence intervals based on a binomial distribution. It also requires that the methylation difference be greater than a threshold determined to be the overall mean of that context specific methylation

Figure 1

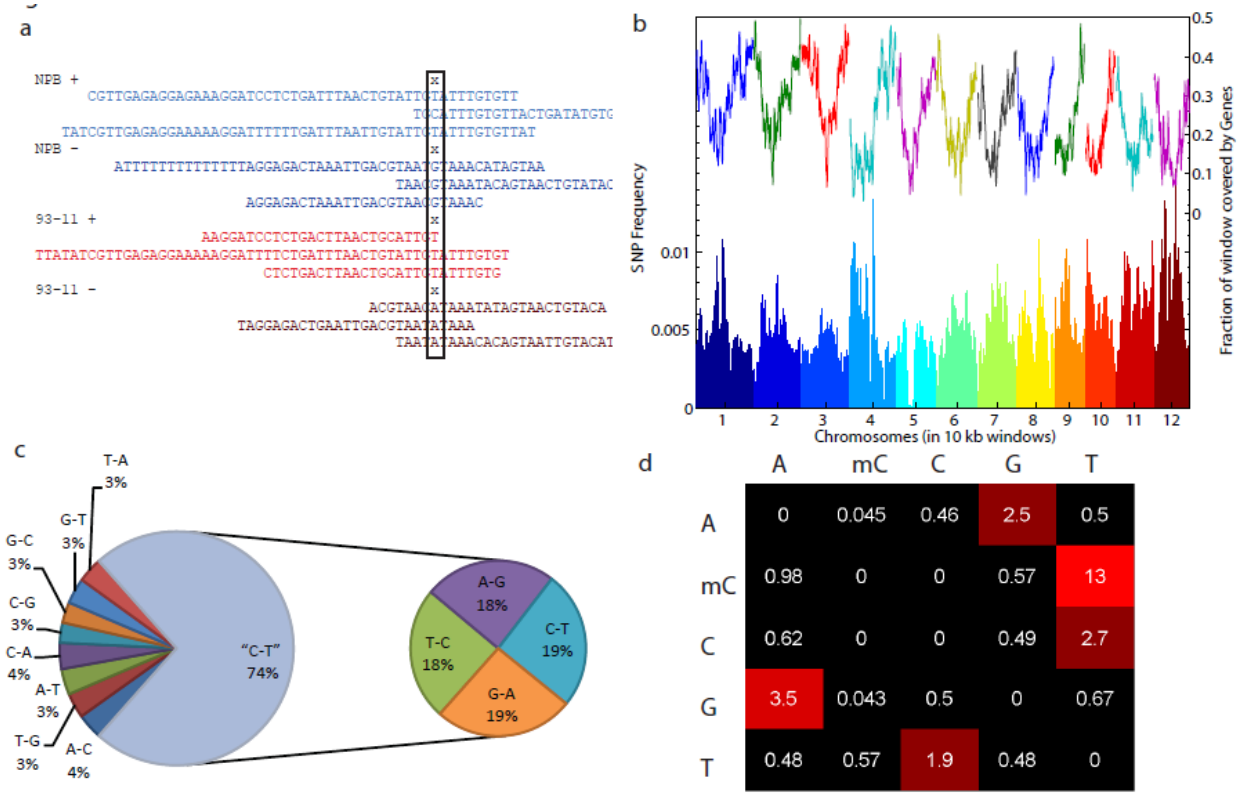

a
NPB +
NPB -
93-11 +
93-11 -
b
SNP Frequency
0.01
0.005
0
1 2 3 4 5 6 7 8 9 10 11 12
Chromosomes (in 10 kb windows)
0.5
0.4
0.3
0.2
0.1
0
Fraction of window covered by Genes
c
T-A 3%
G-T 3%
G-C 3%
C-G 3%
C-A 4%
A-T 3%
T-G 3%
A-C 4%
"C-T" 74%
A-G 18%
C-T 19%
T-C 18%
G-A 19%
d
A mC C G T
A 0 0.045 0.46 2.5 0.5
mC 0.98 0 0 0.57 13
C 0.62 0 0 0.49 2.7
G 3.5 0.043 0.5 0 0.67
T 0.48 0.57 1.9 0.48 0

Figure 2

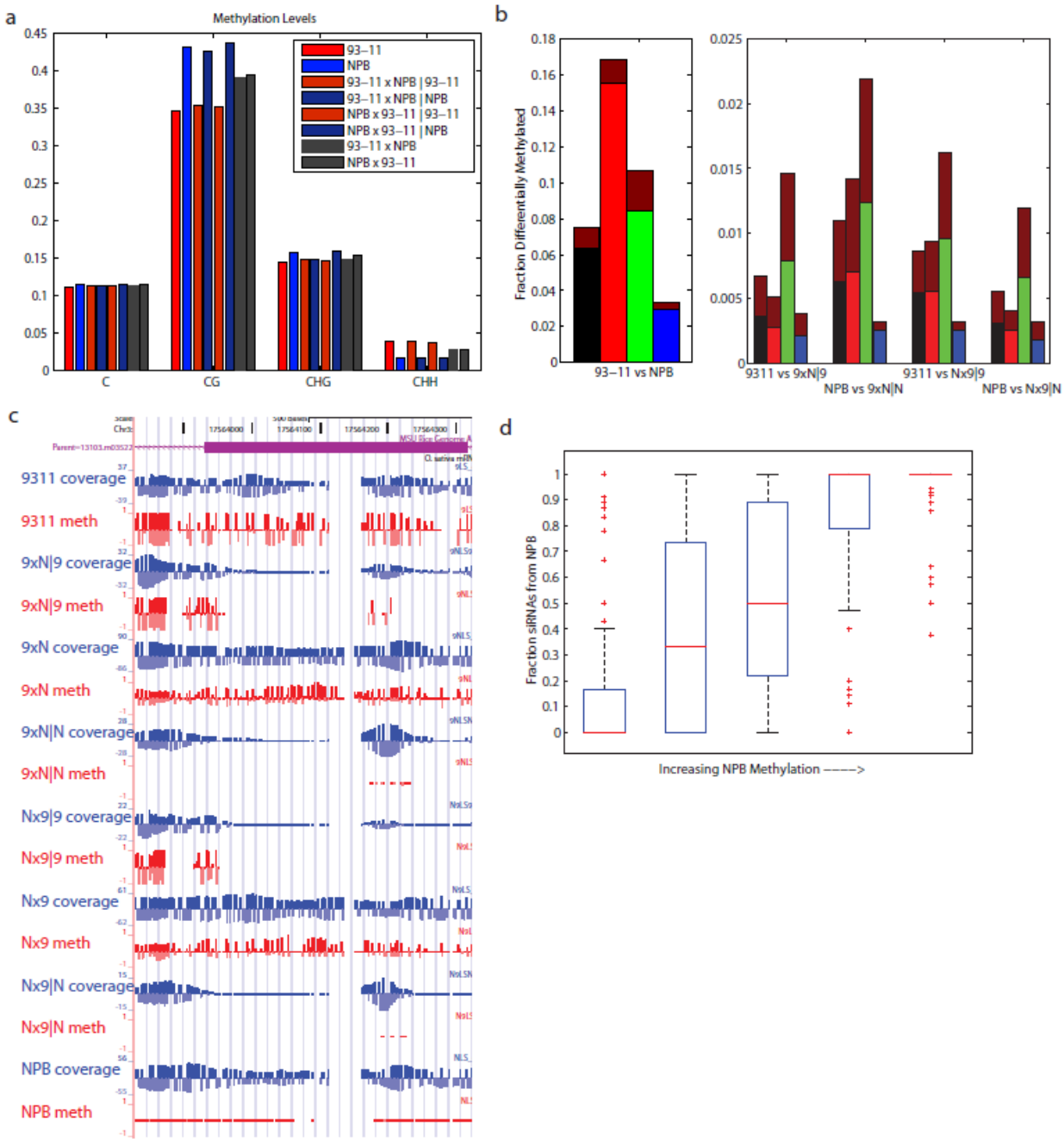
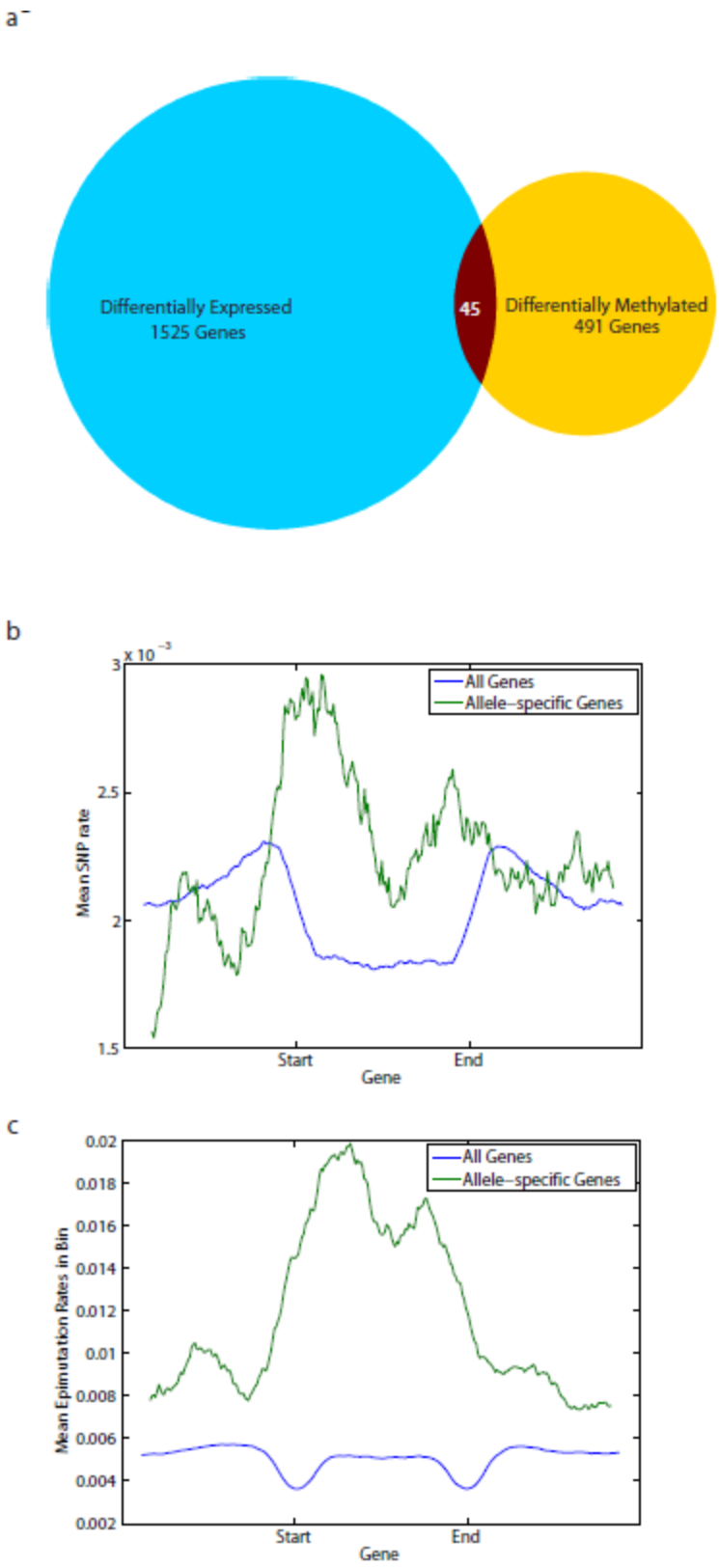
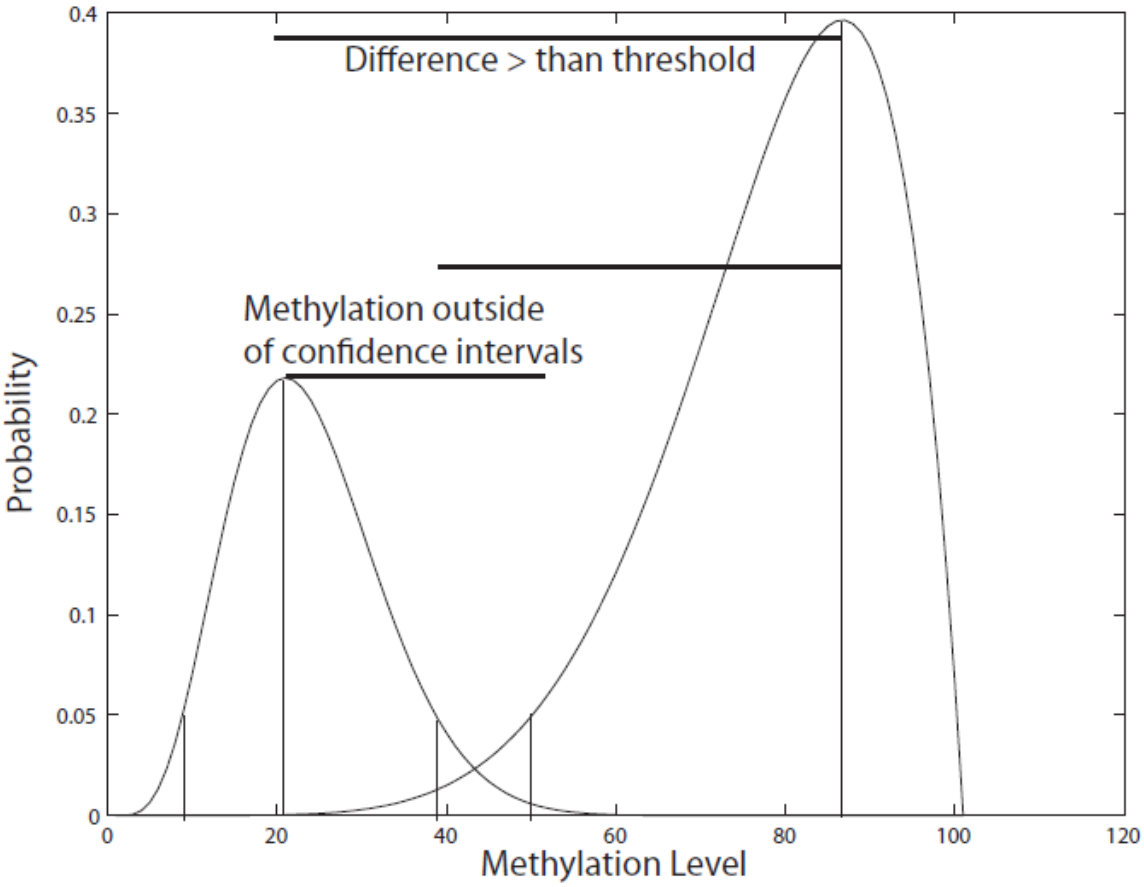

a
Methylation Levels
93–11
NPB
93–11 x NPB | 93–11
93–11 x NPB | NPB
NPB x 93–11 | 93–11
NPB x 93–11 | NPB
93–11 x NPB
NPB x 93–11
C
CG
CHG
CHH
b
Fraction Differentially Methylated
93–11 vs NPB
9311 vs 9xN|9
NPB vs 9xN|N
9311 vs Nx9|9
NPB vs Nx9|N
c
9311 coverage
9311 meth
9xN|9 coverage
9xN|9 meth
9xN coverage
9xN meth
9xN|N coverage
9xN|N meth
Nx9|9 coverage
Nx9|9 meth
Nx9 coverage
Nx9 meth
Nx9|N coverage
Nx9|N meth
NPB coverage
NPB meth
d
Fraction siRNAs from NPB
Increasing NPB Methylation ——–>

Figure 3


a
Differentially Expressed
1525 Genes
45
Differentially Methylated
491 Genes
b
x 10^-3
3
2.5
2
1.5
Mean SNP rate
All Genes
Allele–specific Genes
Start
End
Gene
c
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
Mean Epimutation Rates in Bin
All Genes
Allele–specific Genes
Start
End
Gene

Supplementary Figure 1



Chapter 5: Concluding Remarks

When Watson and Crick proposed the double-helix structure of DNA in their landmark paper in 1953, it must have seemed for a period soon after as if the secrets of the cell were on the verge of being exposed. And while many secrets were revealed, fast forward a few decades, and the complexities of the past were traded in for a new set of intricacies. The simple models of speciation and disease being caused by genetic mutations driving changes in proteins have been replaced by far more complex models that encompass the previous models, but also stress the importance in the timing of expression and the amount of expression. Models of gene expression and control have also evolved. What were "simple" models of gene expression under the combinatorial control of the limited numbers of transcription factors, have now given way to a far more complex story involving DNA methylation, histone codes, and siRNA pathways. Further complicating matters, all of these epigenetic and genetic mechanisms have significant amounts of cross-talk, meaning they must be considered together. This dissertation explored some facets of this epigenetic control mechanisms and their interplay.

The first part of this dissertation focused on nucleosome positioning in Arabidopsis. It described patterns in dinucleotide preferences for nucleosome bound DNA. It also showed the DNA methylation profiles of nucleosomes on a genome scale and identified ten base periodicities in the methylation status on nucleosome bound DNA in Arabidopsis and human stem cells. It also described an enrichment in DNA methylation over nucleosome bound DNA. Furthermore, the methylation pattern was in phase with the WW dinucleotide pattern, suggesting a possibility for DNA methylation to be placed on the nucleosomal DNA with a bound nucleosome.

The first section also described a striking phenomenon where nucleosomes were positioned at the intron-exon boundary in many genes, such that the nucleosomes were directly on exons. It was hypothesized that this may play a role in regulating Pol II processivity as ChIP-

chip results showed that Pol II stalled at the boundary. Moreover, it was also hypothesized that this stalling may play a role in splicing as other groups have suggested.

Further experiments and analyses performed since the time of publishing this article have shown that if only well-positioned nucleosomes are considered, determined in a manner prescribed in Chapter 2, then the methylation enrichment is even greater. In order to investigate the DNA methylation patterns further, it may be a good idea to bisulfite sequence MNase digested mononucleosomal DNA. Because sequencing this DNA would result in identifying the methylation status of the nucleosomal DNA directly, it may yield clearer results. Finally, because DNA methylation was also shown to be periodic over the nucleosomal fraction, it may be helpful to revisit the Kaplan et al. nucleosome prediction algorithm and update it to include DNA methylation data. That is, to generate a new nucleosome prediction algorithm that takes into account dinucleotide frequencies, but DNA methylation frequencies as well. Currently, experiments are underway in yeast to introduce DNA methylation into artificial chromosomes to see how that affects nucleosome positioning.

The second part of the dissertation showed an integrative study of gene mutations, DNA methylation, and gene expression in hybrid rice. The goal was to identify how faithfully genetic material and DNA methylation was inherited and to study how these may impact gene expression. First, this section showed the frequency of mutations from one generation to the next is immeasurably small. However, there was a sufficient abundance of SNPs (1/253 bp) to be able to distinguish some reads from the hybrids based on their parent of origin. Using the SNP data, the frequency of changes in DNA methylation was shown to be much greater than SNP frequency. The study found that of all of the DNA methylation differences found between parent and hybrid comparisons, nearly half were found to be different between the two parents themselves. Finally, the study identified genes that were differentially methylated and genes that were differentially expressed, and the genes that were both.

Recently, a new technique called Hi-C utilizes NGS technology to create a map of chromatin interactions on a genome scale. Using the likelihood of interactions as a proxy for distance between two linearly distant pieces of DNA can allow models of chromosome folding and chromatin structure to be generated. This, combined with other traditional methods like RNA-seq, ChIP-seq, BS-seq, and others that were used in this study, it may be possible to generate unprecedentedly integrative globes (vs. maps) of epigenetic interactions and can help elucidate gene control. The study did not identify the cause for most of the differential gene expression, although a significant number were found to be differentially methylated and differentially expressed. It was hypothesized that trans effects may have been responsible for the remainder of the differential gene expression. However, Hi-C may help in developing a more complete model the epigenetic factors that this dissertation has tried to address.

Because of the vast interplay between the epigenetic factors –the well established relationships between histone modifications and DNA methylation, between the well described relationship between siRNA pathway and DNA methylation, and between nucleosome positioning and methylation as shown here – it naturally lends itself to integrative analyses. Moreover, because the effects of epigenetics may not be as direct and as binary as genetics, it naturally lends itself to genome-wide analyses as well.