FLORES VERGARA, MIGUEL ANGEL. Diversity of Scaffold/Matrix Attachment Regions (S/MARs) in *Arabidopsis* is Revealed by Analysis of Sequence Characteristics, Nucleosome Occupancy, Epigenetic Marks, and Gene Expression. (Under the direction of Dr. George C. Allen and Dr. William F. Thompson.)

Eukaryotic chromatin is organized as independent loops of varying sizes. Following histone extraction with lithium diodosalicylate (LIS), these loops can be visualized as a DNA halo anchored to the nuclear matrix structure. As a basic unit, the loop is thought to be essential for DNA replication, transcription and chromosomal packaging. The formation of each loop is dependent on a specific chromatin segment that must function as an anchor to the nuclear matrix. Sequences that attach specifically to the nuclear matrix have been termed scaffold/matrix attachment regions (S/MARs). Since only a limited number of putative S/MARs have been characterized so far, their role in genomic structure and function is not well understood. Thus, a more global analysis is necessary to answer a variety of questions such as: How are S/MARs distributed across the genome? Are S/MARs associated with different genomic features and are S/MARs typically AT-rich, as previously suggested? What is the nucleosomal organization at S/MAR sequences and do they define regions of accessible chromatin? Are S/MARs associated with specific epigenetic features such as certain histone modifications or DNA methylation? What role do S/MARs play in transcriptional regulation?

I have approached these questions by mapping the S/MARs on *Arabidopsis* chromosome 4 (chr4) using a high-resolution tiling array.
We find that *Arabidopsis* S/MARs can be divided into five clusters based on their location in relation to other previously identified genomic features and show that the S/MARs are a diverse group of elements with multiple functions within the genome. While many of the *Arabidopsis* chr4 S/MARs appear to define structural domains, others are significantly enriched at transcription start sites (TSS) of genes and have a highly significant association with genes that encode transcription factors (TFs). Our data, along with additional genomic information for S/MARs from other organisms, will help our understanding of the roles that different types of S/MARs play within the nucleus.
Diversity of Scaffold/Matrix Attachment Regions (S/MARs) in Arabidopsis is Revealed by Analysis of Sequence Characteristics, Nucleosome Occupancy, Epigenetic Marks, and Gene Expression.

by
Miguel Angel Flores Vergara

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Plant Biology

Raleigh, North Carolina

2011

APPROVED BY:

_______________________________  ________________________________
William F. Thompson            George C. Allen
Co-chair of Advisory Committee  Co-chair of Advisory Committee

_______________________________  ________________________________
Linda Hanley-Bowdoin            Dominique Robertson
DEDICATION

This thesis is dedicated to the memory of my Grandfather Angel Flores Patiño and Grandmother Carolina Vergara Sosa for the profound effect they had on my life.

For of him, through him, and to him, are all things: to whom be glory for ever.

Pro Rege.
BIOGRAPHY

Miguel Angel Flores Vergara was born February 26, 1975 in Chalcatzingo, Morelos, Mexico. He was raised by his grandparents and lived on a farm for the first eighteen years of his life. He received his elementary school education at General Mariano Matamoros school in Chalcatzingo, followed by high school education at Professor Manuel Noceda Barrios High School at Jonacatepec Morelos.

He received his Bachelor of Science in Agronomy from Metropolitan Autonomous University in Mexico City, Mexico in 1998. In June, 2002, he received a Master of Science degree in Plant Biotechnology from Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN) at Irapuato, Mexico, for studies on plant-geminivirus interactions. In September, 2002, he joined the laboratory of Dr. Dominique Robertson as technician in the Department of Plant Biology, North Carolina State University (NCSU).

Miguel was admitted to the Department of Plant Biology at NCSU in 2005, as a candidate for the degree of Doctor of Philosophy. He has undertaken his research under the guidance of Dr. William F. Thompson and Dr. George C. Allen.

Miguel married Yokiko Hiromoto on April 12, 2008. They plan to have children.
ACKNOWLEDGMENTS

Although it would be impossible to list all of the people who have influenced the work described here, I would like to acknowledge those who have had the greatest impact. I am especially grateful to my Ph.D. advisors Dr. George C. Allen and Dr. William F. Thompson for their support and untiring efforts as true teachers. I want to thank the members of my committee, Drs. Linda Hanley-Bowdoin and Dominique Robertson for their guidance and support. I would also like to acknowledge Dr. Steve Spiker for thoughtful discussions that were truly the motivation for much of this work. I would also like to acknowledge Dr. Pete Pascuzzi for his contribution in the analysis of the microarray data, processing of the figures and his comments and guidance on this work. I want to thank my friends and co-workers Yokiko Hiromoto, Mariana Franco, Trino Ascencio, Tae-Jin Lee, Sandeep Kumar, Randy Shultz, Emily Markham, Sergei Krasynanski, Roselyn Hatch, Donna Evans, Dolores Sowinski, Gerson Dominguez, and Shingo Nagaya for their encouragement.

Special thanks go out to my Mom Edith and Dad Alberto, and siblings Willy, Ale, and Beto for all their love, support and encouragement. I want to thank my uncles Migue, Milo, Rosalio, and Javier as well as my aunts Roge, Gloria, Chela, Duli, and Rocio for all their love. Finally I want to thank that one person who, more than anyone else, is responsible for my completing this degree. Yokiko, you have stood by me through everything with your love. Thank you.

This work was supported by a grant from the National Science Foundation.
TABLE OF CONTENTS

LIST OF TABLES ..................................................................................................................... vii
LIST OF FIGURES ................................................................................................................... viii
LIST OF ABBREVIATIONS ...................................................................................................... ix
LIST OF ABBREVIATIONS (CONTINUED) ............................................................................. x

CHAPTER 1 .............................................................................................................................. 1
High-Order Chromatin Organization ...................................................................................... 1
The Nuclear Matrix .................................................................................................................. 3
S/MAR Sequence Characteristics ......................................................................................... 6
The S/MAR Binding Assay ...................................................................................................... 11
S/MARs and Gene Transcription ......................................................................................... 14
S/MARs and DNA Replication ............................................................................................. 17
Mapped S/MARs in Metazoans (Large Scale) ....................................................................... 19
Mapped S/MARs in Plants ..................................................................................................... 22
Goals of this Thesis ............................................................................................................... 25
REFERENCES ........................................................................................................................... 29

CHAPTER 2 ............................................................................................................................ 49
INTRODUCTION ....................................................................................................................... 49
MATERIALS AND METHODS ................................................................................................. 53
Plant Materials and Growth Conditions .............................................................................. 53
Isolation of Nuclei .................................................................................................................. 53
Preparation of Nuclear Halos and Characterization of Proteins ........................................... 54
Analysis of nuclear matrix proteins by gel electrophoresis .................................................. 55
Isolation of S/MAR DNA ....................................................................................................... 55
Microarray Hybridization ..................................................................................................... 56
Exogenous S/MAR Binding Assay ....................................................................................... 58
Comparison of Experimental S/MARs with Predicted S/MARs ........................................... 60
K-means Clustering of S/MARs and Sequence Analysis ....................................................... 60
S/MARs and Transposable Elements (TEs) .......................................................................... 61
S/MARs and Chromatin Structure ....................................................................................... 62
Influence of S/MARs on Gene Expression .......................................................................... 64
Gene Ontology and Transcription Factor Enrichment in S/MAR Proximal Genes ............... 65
RESULTS ........................................................................................................................................ 66
Identification of S/MARs ........................................................................................................ 66
Exogenous S/MAR Binding Assays Validate Microarray Results ........................................ 67
S/MAR Clusters and Genomic Features .............................................................................. 67
AT Content of Clustered S/MARs ....................................................................................... 68
Comparing Experimental S/MARs and Predicted S/MARs ................................................... 70
S/MARs and TEs ..................................................................................................................... 71
Predicted Nucleosome Occupancy at S/MARs ................................................................... 72
S/MARs and Epigenetic Marks .............................................................................................. 74
S/MARs and Gene Expression .............................................................................................. 76
DISCUSSION ................................................................................................................................... 78
Identification of S/MARs on chr4 ........................................................................................ 78
# LIST OF TABLES

Table 1. Plant S/MAR Characteristics .................................................................................. 26
Table 2. Primer sequences for S/MAR regions ..................................................................... 119
Table 3. Primer sequences for non-S/MAR regions ............................................................... 120
Table 4. Results for K-means clustering of S/MARs .............................................................. 121
Table 5. Distribution of S/MARs along chr4 ......................................................................... 122
Table 6. Summary of the 24 selected putative S/MARs for exogenous binding assays.. 123
Table 7. Non-S/MAR fragments .......................................................................................... 124
Table 8. Proportion of predicted S/MARs that matched experimental S/MARs in each S/MAR cluster .............................................................................................................. 125
Table 9. Association of S/MARs with Transposon Element Superfamilies (TE-SFs). ... 126
Table 10. Analysis of S/MARs and chromatin structures ....................................................... 127
Table 11. Statistics for genes with TSS or TTS-associated S/MAR ........................................ 128
Table 12. Gene Ontology (GO) enrichment analysis for genes with a TSS-associated S/MAR ................................................................................................................................. 129
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>K-means clustering of S/MAR data</td>
<td>93</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Determination of S/MAR position relative to the TSS</td>
<td>94</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Histones are removed from Arabidopsis nuclei with 10 mM LIS</td>
<td>95</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Amplification scheme for the one-tube, single-reaction nested PCR method</td>
<td>96</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Quantification of matrix-attached (S/MAR) DNA and total genomic DNA (input)</td>
<td>97</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Distribution of S/MARs on Arabidopsis chromosome 4</td>
<td>98</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Schematic diagram showing the steps of the exogenous binding assay</td>
<td>99</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Chromosomal location of microarray-identified S/MARs and verification by</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>exogenous binding assays</td>
<td></td>
</tr>
<tr>
<td>Figure 9</td>
<td>Matrix binding competition of S/MAR fragments and increasing amounts of</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>sonicated E.coli DNA</td>
<td></td>
</tr>
<tr>
<td>Figure 10</td>
<td>Composite map of two 50-kb regions</td>
<td>102</td>
</tr>
<tr>
<td>Figure 11</td>
<td>NimbleGen microarray profile of S/MARs for the plastocyanin locus</td>
<td>103</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Association of S/MARs with various genomic features</td>
<td>104</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Partition of experimental S/MARs into five clusters according to location on</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>chromosome 4</td>
<td></td>
</tr>
<tr>
<td>Figure 14</td>
<td>AT content (percent) varies for S/MARs within a cluster</td>
<td>106</td>
</tr>
<tr>
<td>Figure 15</td>
<td>AT-rich k-mer content for S/MARs clusters A to E</td>
<td>107</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Experimental and predicted S/MARs show some overlap</td>
<td>108</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Transposon element superfamily (TE-SF) enrichment in S/MARs is strongly</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>correlated with the AT content</td>
<td></td>
</tr>
<tr>
<td>Figure 18</td>
<td>Predicted (probability) of nucleosome occupancy for each S/MAR cluster</td>
<td>110</td>
</tr>
<tr>
<td>Figure 19</td>
<td>DNA cytosine methylation (DNAmC) for each S/MAR cluster</td>
<td>111</td>
</tr>
<tr>
<td>Figure 20</td>
<td>S/MARs and H3K9me2</td>
<td>112</td>
</tr>
<tr>
<td>Figure 21</td>
<td>S/MARs and H3K4me1/2</td>
<td>113</td>
</tr>
<tr>
<td>Figure 22</td>
<td>S/MARs and H3K56ac</td>
<td>114</td>
</tr>
<tr>
<td>Figure 23</td>
<td>S/MARs are associated with TSSs of Genes</td>
<td>115</td>
</tr>
<tr>
<td>Figure 24</td>
<td>S/MARs are associated with the TSS of active Genes</td>
<td>116</td>
</tr>
<tr>
<td>Figure 25</td>
<td>Genes with TSSs-associated cluster C S/MARs are more likely to be expressed</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>than the average gene</td>
<td></td>
</tr>
<tr>
<td>Figure 26</td>
<td>A simplified model depicting the association of S/MARs with gene expression</td>
<td>118</td>
</tr>
</tbody>
</table>
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGRIS</td>
<td>Arabidopsis gene regulatory information server</td>
</tr>
<tr>
<td>ARS</td>
<td>autonomously replicating sequence</td>
</tr>
<tr>
<td>BrdU</td>
<td>bromodeoxyuridine</td>
</tr>
<tr>
<td>BUR</td>
<td>base-unpairing region</td>
</tr>
<tr>
<td>CGH</td>
<td>comparative genomic hybridization</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>Chr</td>
<td>chromosome</td>
</tr>
<tr>
<td>Col-0</td>
<td>Columbia-0</td>
</tr>
<tr>
<td>CUE</td>
<td>core-unpairing region</td>
</tr>
<tr>
<td>Cy3</td>
<td>cyanine 3</td>
</tr>
<tr>
<td>Cy5</td>
<td>cyanine 5</td>
</tr>
<tr>
<td>DAPI</td>
<td>4′,6-diamidino-2-phenylindole</td>
</tr>
<tr>
<td>DHFR</td>
<td>dihydropholate reductase</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP(s)</td>
<td>deoxyribonucleotide triphosphate(s)</td>
</tr>
<tr>
<td>DNase I</td>
<td>deoxyribonuclease I</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>ENCODE</td>
<td>the encyclopedia of DNA elements</td>
</tr>
<tr>
<td>eS/MAR(s)</td>
<td>experimental scaffold matrix attachment region(s)</td>
</tr>
<tr>
<td>ETT/ARF3</td>
<td>ettin/auxin response factor 3</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>FRET</td>
<td>fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>FISH</td>
<td>fluorescence in situ hybridization</td>
</tr>
<tr>
<td>GIK</td>
<td>giant killer</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>HAT(s)</td>
<td>histone acetyltransferase(s)</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid</td>
</tr>
<tr>
<td>HIB2</td>
<td>halo isolation buffer 2</td>
</tr>
<tr>
<td>hnRNA</td>
<td>heterogeneous nuclear ribonucleic acid</td>
</tr>
<tr>
<td>LBAR(s)</td>
<td>loop basement attachment region(s)</td>
</tr>
<tr>
<td>LINE(s)</td>
<td>long interspersed elements</td>
</tr>
<tr>
<td>LIS</td>
<td>3,5-lithium diiodosalicylate</td>
</tr>
<tr>
<td>LTR(s)</td>
<td>long terminal repeat(s)</td>
</tr>
<tr>
<td>MAR(s)</td>
<td>matrix attachment region(s)</td>
</tr>
<tr>
<td>MES</td>
<td>2-(N-morpholino)ethanesulfonic acid</td>
</tr>
<tr>
<td>MITE(s)</td>
<td>miniature inverted repeat transposable element(s)</td>
</tr>
<tr>
<td>NaCl</td>
<td>sodium chloride</td>
</tr>
<tr>
<td>NT-1</td>
<td>Nicotiana tabacum</td>
</tr>
<tr>
<td>PCNA</td>
<td>proliferating cell nuclear antigen</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS (Continued)

PMSF phenylmethanesulfonylfluoride
pS/MAR(s) predicted scaffold matrix attachment region(s)
Q-PCR quantitative polymerase chain reaction
RNA ribonucleic acid
RPA replication protein A
RT-PCR reverse transcription polymerase chain reaction
SAR(s) scaffold attachment region(s)
SATB1 special AT-rich sequence-binding protein-1
SDS sodium dodecyl sulfate
SIDD stress-induced duplex destabilization
S/MAR(s) scaffold/matrix attachment region(s)
SV40 simian virus 40
TAE tris-acetate-ethylenediaminetetraacetic acid
TAIR the Arabidopsis information resource
TE transposon element
TE-SF transposon element superfamily
TF(s) transcription factor(s)
TSS transcription start site
TTS transcription termination site
WGA whole genome amplification
CHAPTER 1

High-Order Chromatin Organization

The eukaryote nucleus contains the genomic DNA of the cell, which can vary from approximately 24 Mb to 350,000 Mb in diploid cells of different organisms. If the human genome’s 6,400 Mb were connected and measured end to end, it would have a length of approximately 2 meters (Davie, 1996). The human genomic DNA must be contained within a nucleus, which has a diameter of 7-10 \( \mu \text{m} \). To accomplish this feat the DNA is organized and packaged into a highly compact form. However, the packaging mechanisms must also be compatible with storage, controlled release, replication, and transmission of genetic information.

Three interconnected levels of chromatin structure have been described (Schalch et al., 2005; Goetze et al., 2007; Misteli, 2007). The first level of structure includes the linear (primary) arrangement of sequence elements that includes genes and regulatory sequences or stretches of repetitive DNA in the genome. The second structural level includes the complex of DNA that is associated with histones and other proteins to form chromatin. The nucleosome, which consists of \(~147\text{ bp}\) of DNA wrapped 1.7 times around the histone octamers, is the basic unit of chromatin (Fransz and de Jong, 2011). In vitro measurements have shown that nucleosomes are regularly spaced along the DNA fiber and separated by 10-100 bp of linker DNA (Luger and Richmond, 1998; Wong et al., 2007), forming a structure referred to as the 10-nm fiber (Lodén and Steensel, 2005), and commonly referred to as “beads on a string.” Under in vitro conditions, the nucleosomes that comprise the 10-nm fiber can be further compacted to form a 30-nm fiber (Finch and Klug, 1976;
Dorigo et al., 2004; Woodcock and Ghosh, 2010), or “solenoid,” which is stabilized by the linker histone H-1 (reviewed in Bassett et al., 2009). Whether the 10-nm or the 30-nm fiber can actually form under *in vivo* conditions is the subject of current debate (Eltsov et al., 2008; Maeshima et al., 2010; Fussner et al., 2011; Zimmer and Fabre, 2011). Alternatives to the original 30-nm solenoid model (Finch and Klug, 1976) have been proposed (reviewed in Fussner et al., 2011), including the two-start helix zigzag model (Woodcock et al., 1984; Horowitz et al., 1994), the cross-linker model (Williams and Langmore, 1991), and the supranucleosome model (Zentgraf and Franke, 1984). Fussner et al. (2011) support a model in which the 30-nm fiber is not required for compacting the DNA within the nucleus and that simple variations in the packing densities of the 10-nm fiber are adequate.

It is has been traditionally thought that the 30-nm fiber is further compacted into a series of loops, representing a third level of chromatin structure (Finch and Klug, 1976; Woodcock and Ghosh, 2010). Higher-order chromatin structure also includes a three-dimensional organization pattern in which some parts of the chromatin are packaged into more compact structures, often referred to as heterochromatin, while other regions ( euchromatin) remain less compact and more accessible to regulatory factors.

Laemmli and co-workers (Adolph et al., 1977; Marsden and Laemmli, 1979) originally proposed that mitotic chromosomes exist as a series of chromatin loops that are topologically constrained by interacting with a scaffold, or matrix, consisting of non-histone proteins (Berezney and Coffey, 1974). The actual existence of the nuclear matrix has been debated for many years (Pederson, 2000; Strukov and Belmont, 2009), and a random self-organization model has been presented as an alternative explanation of nuclear structure.
(Misteli, 2001). However, in contrast to random self-organization, recent data have shown that genes migrate to preassembled transcription sites upon activation (Osborne et al., 2004), a result which is most consistent with the existence of a nuclear substructure (reviewed in Chakalova et al., 2005).

In spite of Mistelli’s argument (2001), the current consensus is a hierarchy of packaging, leading to a loop organization. The estimated size of the loops ranges from 50-300 kb in a variety of organisms (Gasser and Laemmli, 1987; Gerdes et al., 1994). Apart from packing large amounts of DNA into a very small volume, and protecting the genetic information, compaction of genomic DNA into chromatin loops provides an additional level of functionality that may help to coordinate numerous nuclear processes, such as temporal and spatial control of thousands of genes, faithful DNA replication during cell division, and molecular mechanisms that are responsible for repairing damaged DNA (Demeret, 2001; Cremer et al., 2004; Goetze et al., 2007; Lanctot et al., 2007).

**The Nuclear Matrix**

Nuclei are intricately structured, and the subnuclear processes have an elaborate spatial organization. It has been proposed that the architecture of the nucleus is largely controlled by the organization of chromatin on a nuclear matrix (Berezney and Coffey, 1974). As noted above, there has been considerable controversy about the *in vivo* existence of the nuclear matrix (or scaffold, as it is sometimes called). This structure is defined operationally as a residual proteinaceous matrix that remains after selective extraction of
histones and DNA. This operational definition depends upon the specific histone extraction protocol used, and has been a primary cause of the controversy over whether or not a matrix exists in vivo. Biochemically, the nuclear matrix is defined as the insoluble material that remains after extraction of nuclei with 2 M NaCl (Berezney and Coffey, 1974), whereas the nuclear scaffold remains after extraction of nuclei with 10-25 mM lithium diiodosalicylate (LIS) (Mirkovitch et al., 1984). Although the operational definition has been generally recognized, morphological similarities of the structures visible in the isolated nuclear matrix compared to intact nuclei have been documented (reviewed in Nickerson, 2001). In early studies, Don Fawcett (1966) defined the nuclear matrix as the non-chromatin structures of the nucleus easily observed in unextracted cells under the electron microscope (Fawcett, 1966). Berezney and Coffey (1974) supported and expanded the original observation by finding a similar nuclear morphology when rat liver nuclear protein matrices were compared to the intact nuclei. Their data indicated that the structure of nuclear matrix exists in isolated nuclei, and it is not a result of the treatments with salt, detergents and enzymes. Alternative approaches have since been used to demonstrate the nature and existence of the nuclear matrix in situ, and to study specific aspects of the nuclear matrix (Goldfischer et al., 1981; Capco et al., 1984; Chaly et al., 1984; Fey et al., 1984; Nickerson et al., 1990). Despite the significant amount of literature describing the nuclear matrix both biochemically and microscopically, the existence of such network of nucleoprotein fibers still remains controversial (reviewed in Pederson, 1998; Hancock, 2000; Pederson, 2000).

The experimental foundations for nuclear matrix research has historically rested on four pillars: (1) the observation by electron microscopy of fibrogranular non-chromatin
structures in the nuclei of cultured cells and tissues; (2) the development of methods for isolating these non-chromatin structures that preserve their ultrastructure; (3) the observation that the chromatin is organized in loop domains attached at their bases to a non-chromatin structure; and (4) the discovery of spatially distinct functional domains within the nucleus that remain in place after the removal of soluble proteins and chromatin (reviewed in Nelson, 1986; Nickerson, 2001).

The nuclear matrix is a universal feature of the eukaryotic nuclei, since such residual nuclear structures have been isolated from a wide variety of metazoans and plant organisms (Berezney, 1995; Diaz de la Espina, 1996). The proteins comprising the matrix fraction represent between 10-25% of the total nuclear protein mass and comprise a subset of nonhistone proteins including lamina, components of the nuclear pore complex and a network of ribonucleoprotein filaments (He et al., 1990). These filaments can be visualized by electron microscopy and create a framework within the nucleus (Capco et al., 1982).

A variety of specific proteins have been identified as part of the nuclear matrix including topoisomerase II (Earnshaw et al., 1985), components of spliceosomes (Bisotto et al., 1995), hyperphosphorylated large RNA polymerase II subunit (Mortillaro et al., 1996; Patturajan et al., 1998), transcription factors (Bidwell et al., 1993; Guo et al., 1995; Merriman et al., 1995; Nardozza et al., 1996), protein kinases (Berezney, 1995), proliferating cell nuclear antigen (PCNA) (Gerner and Sauer, 1999), and other proteins associated with DNA replication (van Driel et al., 1996; Anachkova et al., 2005). Recent work enumerates dozens of proteins identified as matrix components through biochemical isolation (Albrethsen et al., 2009).
Proteomic analysis of the *Arabidopsis thaliana* nuclear scaffold (matrix) extracted with 25 mM LIS identified homologues of nucleolar proteins Nop56, Nop58 (SAR-binding proteins), Nop140, fibrillarins, nucleolin, ribosomal components, a putative histone deacetylase, beta tubulins, and several unknown proteins (Calikowski et al., 2003). It is notable that several major proteins identified from the nuclear matrix of animal nuclei were absent from the plant nuclear matrix, such as topoisomerase, RNA polymerase II subunit, and lamins (Calikowski et al., 2003; Albrethsen et al., 2009). Although many nuclear matrix proteins have been identified and sequenced, less is known about where or how they assemble into structures. Future research with the combined application of biochemistry and cell biology will help to solve this enigma.

Numerous studies have provided evidence that the nuclear matrix is a highly dynamic structure involved with important biological functions such as chromatin organization, DNA replication, and heterogeneous nuclear RNA (hnRNA) synthesis and processing (Nelson, 1986; Davie, 1996). These discoveries indicate that many crucial nuclear events occur not in solution but rather in association with relatively insoluble structural components associated with the nuclear matrix.

**S/MAR Sequence Characteristics**

The chromatin loop model requires that the DNA sequences that are at the base of the loop must bind to the nuclear matrix. In identifying DNA sequences that bind to the nuclear matrix, two biochemical approaches have generally been used. In the high salt method
nuclear matrices are isolated by extraction of histones using 2 M NaCl, followed by digestion with DNase I. The DNase-resistant DNA fragments that remain bound to the nuclear matrices are operationally defined as matrix attachment regions (MARs) (Berezney and Coffey, 1974; Cockerill and Garrard, 1986). In the second method (low salt method) nuclei are extracted with a buffer containing 10 to 25 mM 3,5-lithium diiodosalicylate (LIS), which removes histones and other soluble proteins (Mirkovitch et al., 1984). Extracted nuclei are then digested with restriction endonucleases, and DNA fragments that remain bound are operationally defined as scaffold attachment regions (SARs).

Some authors use the term “scaffold” to refer to the matrix-like structures remaining after LIS extraction, reserving the term “matrix” for structures remaining after high salt extraction. However, the two methods yield generally similar results, so that a DNA fragment found in the matrix fraction following high salt extraction would usually also remain bound after LIS extraction. The nomenclature problem is further complicated by the fact that some authors use the same two terms, matrix and scaffold, to distinguish structures isolated from interphase nuclei and condensed mitotic chromosomes. DNA fragments isolated from interphase and mitotic nuclear matrices after nuclease treatment are known as MARs and SARs, respectively (Paulson and Laemmli, 1977; Mirkovitch et al., 1988; Berezney, 1995; Ottaviani et al., 2008b). However, for simplicity, I will designate all such DNA sequences as S/MARs, as originally suggested by Bode et al. (1996).

S/MARs have been found in all eukaryotes that have been examined, including yeast, animals, and plants (Amati and Gasser, 1988; Stief et al., 1989; Gasser et al., 1990; Hall et al., 1991; Breyne et al., 1992; Michalowski et al., 1999; Linnemann et al., 2009). Some of
these sequences have been shown to be capable of binding to nuclear matrices prepared from a variety of organisms (Mirkovitch et al., 1988; Chung et al., 1993; Dietz et al., 1994; van Drunen et al., 1997), indicating that at least one basic mechanism of S/MAR attachment to the nuclear matrix is evolutionarily conserved. However, some S/MAR sequences do show tissue specificity of binding, both in vitro (Levy-Wilson and Fortier, 1989; Bidwell et al., 1993), and in direct isolation (sometimes called in vivo\(^1\)) assays (Eivazova et al., 2006; Ottaviani et al., 2008b).

Indeed, S/MARs have been categorized as either being constitutive (permanent) or facultative (such as cell type specific) depending on their dynamic properties (Boulikas, 1993). Constitutive S/MARs are found in the matrix fraction of all types of cells, and may mark permanent domain boundaries, whereas the binding of facultative S/MAR appears to depend on the presence of proteins specific to a particular tissue or cell cycle phase. Some researchers suggest that loops formed by constitutive attachment sites may also contain facultative S/MAR elements that permit the attachment of particular genes to the nuclear matrix in appropriate conditions (Heng et al., 2004b). Alternatively, active genes may be transiently attached to the nuclear matrix via matrix-bound transcriptional complexes (Razin and Yarovaya, 1985; Iarovaia et al., 2005). However, it is important to be aware that due to technical limitations, all measurements of S/MAR binding are merely a snapshot of a very dynamic environment. Thus, the proposed S/MAR roles, whether constitutive or facultative,

\(^{1}\) As a note of refinement, we use the terms “direct isolation” or “endogenous assay” as an alternative to “in vivo” because even S/MARs isolated directly from matrix preparations might have become bound during the matrix preparation procedure and thus might not have been attached in vivo.
may not be easily classified through the study of a single time point, environment or cell type.

Reported sizes of S/MARs vary from several hundred base pairs to more than 10 kb. S/MARs are dispersed throughout eukaryotic genomes, having been identified in centromeric DNA (Strissel et al., 1996; Sumer et al., 2003), near enhancers in 5’ and 3’ flanking sequences (Mirkovitch et al., 1984; Small et al., 1985; Gasser and Laemmli, 1986a; Cockerill et al., 1987; Phi-Van, 1988; Zhong et al., 1999; Tikhonov et al., 2000), within genes (Käs and Chasin, 1987; Jarman and Higgs, 1988; Romig et al., 1994; Shaposhnikov et al., 2007), and in intergenic regions (Bode and Maass, 1988; Chernov et al., 2002; Glazko et al., 2003; Purbowasito et al., 2004 and references therein).

Analysis of S/MAR sequences has revealed a lack of extensive homology. Many groups have reported that most S/MARs are AT-rich (typically more than 70% A+T). However, it should be noted that a DNA fragment that is AT-rich might not behave as a S/MAR, suggesting that AT richness alone is not sufficient for binding.

Early studies indicated that a number of sequence motifs and features were commonly found in S/MARs including A-boxes (AATAAAYAAA) (Gasser and Laemmli, 1986a), T-boxes (TTWTWTTWTW) (Gasser and Laemmli, 1986a), a Drosophila topoisomerase II consensus sequence (GTNWAYATTNATNNR) (Gasser and Laemmli, 1986b), an autonomously replicating sequence (ARS) motif (WTTTATRTTTW) (Amati and Gasser, 1988), TG/CA-motifs, CT-rich stretches, kinked DNA, and curved DNA (Cockerill and Garrard, 1986; Nelson, 1986; Boulikas, 1993; Liebich et al., 2002).
While many sequence motifs and features have been associated with S/MARs, it remains unclear whether the presence of one or several such motifs is sufficient to establish a DNA fragment as a S/MAR (Michalowski et al., 1999; Liebich et al., 2002). A study by Van der Geest et al. (2004) used a subset of the “S/MAR sequence” motifs, known to be common to characterized S/MARs, to synthesize several S/MARs. When six of the synthetic S/MARs were tested using the *in vitro* matrix-binding assay, it was found that two of the S/MARs bound the nuclear scaffold (matrix) from maize leaves. The strength of matrix binding was similar that of the natural plant beta-phaseolin 3’ MAR, with strong specific association occurring in the presence of a large excess of *E. coli* competitor DNA (Van der Geest et al., 2004). A comparison of the “S/MAR sequence” motifs from the two strongly-binding synthetic S/MARs to the previously characterized 3’ beta-phaseolin S/MAR (shown in Table 1, Van der Geest et al., 2004) found no clear motifs in common.

Recent evidence suggests that an essential feature of S/MAR binding to the nuclear matrix is the ability of the DNA to melt and form single strands under relatively mild conditions (Bode et al., 1992; Benham et al., 1997; Bode et al., 2006). *In vitro* chemical reactivity studies have shown that S/MARs contain local regions with an inherent tendency to become single stranded, which can be displayed in the form of a stress-induced duplex destabilization profile (SIDD) (Benham et al., 1997). Depending on the salt conditions, strand separation initiates at the most easily destabilized site, the core-unpairing element (CUE), and can spread throughout a base-unpairing region (BUR) (Kohwi-Shigematsu and Kohwi, 1990; Benham et al., 1997). The likelihood of strand separation is probably
determined by the presence of (AT)n motifs and ATATT motifs (Kohwi-Shigematsu and Kohwi, 1990; Bode et al., 1992).

Varying distances between identified S/MARs suggest possible folding of chromosomal regions into loops of varying size. For instance, estimated chromatin loops, defined by the relative distance between S/MARs, ranged from 0.5 to 970 kb along human chromosome 16 (Linnemann et al., 2009). Similarly, the S/MAR mapping within the ENCODE regions of the human genome suggests loop sizes that range from 2.5 to 606 kb, with an average of 44 kb and median of 18 kb (Keaton et al., 2011). In plants, relatively small (3 to 10 kb) loops seem to represent a common pattern for the organization of plant chromatin (Breyne et al., 1992; van der Geest et al., 1994; Chinn and Comai, 1996; Avramova et al., 1998). In the Arabidopsis genome, for example, chromatin loops averaged 5 kb in the plastocyanin locus (van Drunen et al., 1997) and within an 80-kb region of chromosome 5 (Tachiki et al., 2009).

The S/MAR Binding Assay

Identification of a S/MAR requires an assay to examine whether a particular sequence can bind to the isolated nuclear matrix under in vitro conditions. Isolated nuclei are first treated with 10 to 25 mM LIS-containing buffers to remove histones. Removal of the packaging constraint imposed by the histones allows the DNA to unwind from the nuclear matrix core, resulting in a nuclear halo of extended DNA surrounding the matrix, or “residual nucleus.” These halos are easily visualized microscopically in DAPI-stained preparations.
Electron microscopy reveals that the loop bases remain attached to the nuclear matrix at the S/MARs (Paulson and Laemmli, 1977; Marsden and Laemmli, 1979). The halo of relaxed DNA can then be digested using restriction enzymes and separated from the nuclear matrix by centrifugation, leaving matrix attachment sites and their immediate flanking regions (Mirkovitch et al., 1984).

An in vitro binding assay, which is often referred to as an “exogenous” binding assay, is done by mixing labeled DNA fragments of interest with an excess of unlabeled competitor DNA (usually fragmented E. coli genomic DNA) to suppress nonspecific binding. The mixture is then incubated with the isolated nuclear matrices for binding. The insoluble nuclear matrices with bound S/MAR DNA are then separated from non-S/MAR DNA by centrifugation and the protein is removed using protease digestion. DNA fragments from the non-S/MAR (supernatant) and S/MAR-containing (pellet) fractions are then separated by electrophoresis on agarose gels and the labeled DNA is visualized. In my work, an infrared imaging system (odyssey, LI-COR) was used to visualize non-radioactive fluorescently labeled DNA, whereas previous studies have used \(^{32}\)P labeled DNA (Hall et al., 1991). If a particular labeled DNA fragment is present in the matrix-bound (pellet) fraction under appropriate conditions, then the fragment is considered a S/MAR. Conversely, a DNA fragment that corresponds to a putative loop region (non-S/MAR) remains in the supernatant fraction. The unbound DNA can also serve as a control to determine the quality of the matrix preparation. For example, if residual histones are present in the pellet fraction, non-specific binding can occur and non-S/MAR control DNA will also be found in the pellet (matrix) fraction.
A complementary method for S/MAR identification is the “endogenous” binding assay, which can be used to identify endogenous DNA fragments that remain bound to the matrix after the histones are removed. The initial steps of the endogenous assay are similar to the exogenous binding assay. Nuclei are isolated, the histones are removed with LIS, and the nuclear halos are digested with restriction enzymes. The nuclear matrix along with the endogenous bound S/MARs in the pellet are separated from the putative loop supernatant DNA by centrifugation, and the matrix proteins are digested with a protease, followed by extraction with phenol/chloroform/isoamyl alcohol. Specific sequences can then be monitored in the bound and unbound fractions by DNA gel-blot hybridization (Hall et al., 1991) or polymerase chain reaction (PCR) (Tachiki et al., 2009) assays. If a given sequence is found principally in the matrix-bound fraction, it is considered an “endogenous” S/MAR.

I have used a modification of the endogenous binding assay to identify S/MARs on chromosome 4 (chr4) from a suspension cell line of Arabidopsis thaliana. Arabidopsis S/MAR DNA (target DNA) and input sheared genomic DNA (reference DNA) samples were amplified, labeled with Cy5 and Cy3 fluorescent dyes, respectively, and hybridized to a custom-designed NimbleGen high-resolution tiling microarray covering chr4, as described in Chapter 2. Probe enrichment ratios were used to define genomic regions enriched in the fraction of DNA that remained bound to the matrix.

It is important to be cautious in interpreting these data. Although it is common in the literature to refer to S/MAR sequences isolated in this way as endogenous S/MARs and to assume that they represent in vivo attachments, one must remember that the procedure necessarily involves disrupting cells and nuclei, creating an opportunity for interactions to
occur during isolation that might not occur in vivo. For example, some DNA sequences may exist that have the potential to bind to the matrix but do not do so because of topological constraints in vivo. After restriction enzyme digestion, such sequences would no longer be topologically constrained and may bind to the nuclear matrix in a manner similar to the way sequences bind in the exogenous assay. While we do not know how prevalent such binding may be, it is possible that both the exogenous and endogenous binding assays may test in vitro binding. We have used a modified version of the endogenous binding assay to isolate and map S/MARs at a large genomic scale using a tiling array for Arabidopsis chr4, and the exogenous, in vitro binding assay of selected S/MAR regions to verify microarray results.

**S/MARs and Gene Transcription**

It has been proposed that tethering of chromatin to the nuclear matrix plays an important role in the transcription of a gene (Heng et al., 2004a; Eivazova et al., 2007; Ottaviani et al., 2008a). Certain S/MARs have also been proposed to delineate the boundaries of structural domains that allow for coordinate gene regulation by isolating the domains from the effects of the surrounding chromatin (Allen et al., 1996; Ishii et al., 2002). According to this model, a chromatin loop domain flanked by boundary S/MARs can resist the spread of heterochromatin (Byrd and Corces, 2003; Wei et al., 2005). In Drosophila melanogaster, S/MARs often flank genes or gene clusters and co-localize with gypsy, which is one of the most extensively studied chromatin insulator elements (Nabirochkin et al., 1998). When the nuclear matrix protein Su(Hw) binds to gypsy chromatin, loops are formed as demonstrated
by fluorescence in situ hybridization (FISH) and immunofluorescence analysis of *Drosophila* nuclear halos (Byrd and Corces, 2003). As expected, mutant forms of Su(Hw) that cannot bind to the nuclear matrix disrupt the chromatin loop structures (Byrd and Corces, 2003; Valenzuela and Kamakaka, 2006). This and other observations are consistent with a model in which S/MAR binding to the nuclear matrix topologically constrains the DNA into independent loop structures.

Transcriptionally active domains are characterized by increased general sensitivity to nucleases such as DNase I or micrococcal nuclease (reviewed in Berezney, 1995). This sensitivity indicates an open chromatin structure, which is important for allowing access to the transcriptional machinery. In contrast, transcriptionally inactive domains, which are relatively resistant to nuclease, have less access to the transcriptional machinery. The possibility of transcriptional domains and structural chromatin loops being one and the same is a basic concept of the loop domain model (Paulson, 1988; Gasser et al., 1990).

The loop domain model can also explain certain transgene expression results from a wide variety of eukaryotic organisms. When a cell is transformed, a transgene flanked by S/MAR sequences is often expressed at higher levels when compared with a transgene lacking the flanking S/MARs, suggesting that the flanking S/MARs may shield the transgene from the effect of the neighboring host chromatin (Grosveld et al., 1987; Stief et al., 1989; Halweg et al., 2005; Girod et al., 2007). The increases in expression are dependent upon the chromatin structure at the site of integration. Transgene expression from transformation procedures that do not require stable transgene integration, such as transient expression, is
not greatly affected by the presence of flanking S/MARs (Phi-Van et al., 1990; Klehr et al., 1991; Allen et al., 1996).

More recent studies have shown the effect of S/MARs on transgene expression may vary and appear more complex than what can be simply explained by the loop domain model (Breyne et al., 1992; Bonifer et al., 1994; Huber et al., 1994; De Bolle et al., 2003), (reviewed in Allen et al., 2000; Bode et al., 2000 and Thompson et al., 2007). It is likely that many additional factors are involved in controlling transgene expression, notably including post-transcriptional gene silencing (Matzke and Birchler, 2005; Thompson et al., 2007). Variation between experiments with different S/MAR elements may also reflect the existence of multiple classes of S/MARs, some of which may lack a loop domain boundary element function. This idea will be explored further in Chapter 2.

Some S/MARs are found at or near cis-acting regulatory elements of genes, which might position the genes at or close to the nuclear matrix, where higher concentrations of RNA polymerase and transcription factors may occur (Ciejk et al., 1983; Boulikas, 1994; Davie, 1996; Kimura et al., 1999; Keaton et al., 2011). This hypothesis is increasingly supported by experimental data showing that transcribing RNA polymerases are located at fixed sites on the nuclear matrix (Cook, 1999), and placing actively transcribed genes at the bases of chromatin loops close to matrix-attached RNA polymerases (Gerdes et al., 1994; Jackson et al., 1998; Papantonis et al., 2010). Many of the enzymes and complexes required for chromatin remodeling and gene regulation also appear to be associated with the nuclear matrix and may be present at the bases of chromatin loops (Hendzel et al., 1994; Reyes et al., 1997; Alfonso-Parra and Maggert, 2010). In addition Chua et al. (2000) found an AT-rich
transcriptional enhancer from the pea plastocyanin gene (PetE) that could bind to the nuclear matrix and was important for regulating expression through acetylation of H3 and H4 on the promoter and the nearby coding regions. These authors suggested that certain S/MARs mediate histone acetylation by recruiting histone acetyltransferases (HATs), which is consistent with the possibility that chromatin remodeling can occur on the nuclear matrix.

**S/MARs and DNA Replication**

The association of high-order chromatin with the nuclear matrix has long been proposed to play a role in organizing the genome for DNA replication (Pardoll et al., 1980; Vogelstein et al., 1980; Courbet et al., 2008). Indirect evidence supports the idea that each loop domain represents one individual replicon or individual replication unit, with the origin of replication localized within the loop domain and the ends of the replicon attached to the nuclear matrix at the bases of the loop, although not all the details are fully understood (Buongiorno-Nardelli et al., 1982; Hyrien et al., 1997; Anachkova et al., 2005). Experiments in which DNA synthesis was visualized by incorporation of BrdU (bromodeoxyuridine, a synthetic thymidine analog) showed that newly replicated DNA in nuclear halos is mainly localized at the base of the chromatin loops. When BrdU incorporation was followed by a thymidine chase, the location of the BrdU-labeled DNA moved away from the bases toward the loop periphery (Pardoll et al., 1980; Vogelstein et al., 1980; Gerdes et al., 1994). Additional data supporting the model that DNA replication occurred at the nuclear matrix was provided by Vaughn et al. (1990), who analyzed nuclear matrix-attached DNA using
two-dimensional gel electrophoresis and found enrichment for replication forks. Recently Courbet et al. (2008) used FISH and found that some known replication origins are located at or near the base of loop domains in *Xenopus* nuclei. Mesner et al. (2006) provided compelling evidence that DNA replication occurs at the nuclear matrix by using nuclear matrix-bound DNA from early S phase CHO cells as a purification step for newly replicated DNA. The nuclear matrix binding purification step increased the purity of the replication intermediates by ~25-fold. All of the fifteen randomly selected clones were validated by two-dimensional gel electrophoresis, which is the most stringent test of origin activity (Mesner et al., 2006). These results are consistent with the concept that DNA is replicated at the bases of the loops by DNA polymerase complexes bound to the nuclear matrix - possibly the fixed replication factories proposed by Cook’s group (Hozak et al., 1993). In addition, many proteins involved in DNA replication, such as DNA polymerases, primases, PCNA (proliferating cell nuclear antigen), RPA (Replication Protein A), and topoisomerases have been classified as matrix-bound (Earnshaw et al., 1985; Anachkova et al., 2005; Radichev et al., 2005). In plants, PCNA and BrdU pulse-labeled DNA co-localize as discrete foci in onion nuclei (Samaniego et al., 2001; Samaniego et al., 2002), which is consistent with the co-localization model for other eukaryotes (Hozak et al., 1993). However, no data are available from plant studies to address the question of whether or not replication is associated with the nuclear matrix.

Several lines of experimental evidence suggest a functional link between S/MARs and DNA replication origins. Indeed, some S/MARs have been found at or near to the chicken lysozyme DNA replication origin (Phi-Van, 1988), the DHFR origins β and β’
(Dijkwel and Hamlin, 1988), the lamin B2 origin (Lagarkova et al., 1998), and the c-myc origin (Girard-Reydet et al., 2004). Interestingly, the addition of S/MAR sequences to viral-based vectors, such as SV40 (Piechaczek et al., 1999), and non-viral vectors such as eukaryotic DNA minicircles (Broll et al., 2010) enhances their episomal replication. S/MARs are typically AT-rich (>70% AT) and have low melting energy, which would facilitate DNA strand separation required for initiation of DNA replication (Bode et al., 2006).

**Mapped S/MARs in Metazoans (Large Scale)**

Whole genome sequencing has greatly facilitated mapping and analysis of regulatory elements such as S/MARs on a genome scale. Large-scale S/MAR mapping has been done using exogenous binding assays and quantitative real-time polymerase chain reaction (Q-PCR), and more recently oligonucleotide tiling arrays. For example, Rollini et al. (1999) used exogenous binding assays to identify five S/MARs in the human serpin gene cluster (∼150 kb) at 14q32.1. Four of the S/MARs were in intergenic regions, and one S/MAR was within an intron, which suggested a role in regulation of transcription. Interestingly, sequence analysis revealed that all five S/MARs contained Long Interspersed Elements (LINEs) and Long Terminal Repeats (LTR). Chernov et al. used a similar exogenous binding assay to identify S/MARs within the 1-Mb multigenic FXYD5-COX7A1 region of human chromosome 19 (Chernov et al., 2002). A majority of the identified S/MARs (11 out of 16) were located in intergenic regions suggesting a structural role to delineate chromatin
domains, while the remaining five S/MARs were found within introns and were proposed to function in transcriptional regulation.

In mouse, a total of 52 S/MARs were mapped in the 1-Mb distal imprinted domain of chromosome 7 using exogenous binding assays (Purbowasito et al., 2004). The average S/MAR spacing was approximately 20 kb, but the spacing was unevenly distributed across the imprinted domain. Comparison of mapped S/MARs with genomic features revealed that 24 S/MARs were in introns, 25 S/MARS in intergenic regions, and 3 S/MARs were within genic regions. The AT content for the mapped S/MARs ranged from 55% to 71%, with an average of 62%. This average value was higher than the mean AT content for the whole mouse genome (58%), supporting the view that S/MARs are generally AT-rich. However, 7 of the 52 S/MARs also had a slightly lower AT content (55 to 57%) than the mean AT content for the whole mouse genome, supporting the idea that primary sequence is not the sole determinant for matrix binding.

Recently, Linnemann et al. (2008) used a NimbleGen comparative genomic hybridization (CGH) array system to survey SARs (LIS histone extraction) and MARs (2 M NaCl histone extraction) in HeLa cells. Their discussion focused primarily on chromosome 16, as data were similar for chromosomes 14, 15, 16, 17, and 18. A total of 1,016 SARs and 775 MARs were identified on chromosome 16, of which 403 were common between the LIS and 2 M NaCl extraction protocols. MARs were predominantly found in gene-poor intergenic regions, and genes that co-mapped with MARs mostly lacked transcripts. Conversely, SARs exhibited a more even distribution across the chromosome and tended to overlap with genes. Interestingly, only SARs located within 10 kb upstream of the 5’ ends of
genes significantly correlated with gene expression, whereas SARs that overlapped or occurred within genes did not. It was proposed that LIS extraction isolates inner nuclear matrix components, whereas NaCl extraction without cation/heat stabilization isolates peripheral nuclear matrix components (Linnemann and Krawetz, 2009). Taken together, the data suggested that different types of DNA attachments - revealed experimentally by either LIS vs NaCl extraction - work in concert with other factors to modulate gene expression in animal cells.

Recently Keaton et al. (2011) used tiling microarrays to identify 453 LIS-isolated SARs in HeLa nuclei across the 30 Mb of human genome sequence being studied by the ENCODE pilot project. Sequence analysis of the ENCODE SARs showed a higher median AT content (60.2%) than the median AT content of the regions analyzed (58.1%). The ENCODE SARs were predominantly found in genic areas, with 74% of the SARs residing within 5 kb of an annotated gene. Interestingly, the majority of these SARs were near an expressed gene. In addition, SARs were enriched for RNA polymerase II binding sites, acetylated histones H3 and H4, and methylation of histone H3K4, suggesting that SARs are associated with open chromatin conformation and active transcription.

Caution must be exercised when extending these results with animal cells to work with plant systems. Structural and functional studies of the plant nuclear matrix suggest that there are differences as well as similarities when compared to the animal nuclear matrix (Diaz de la Espina, 1996; Calikowski et al., 2003). One of the most striking differences is the apparent absence of bona fide lamins in plants, as plant genome sequences seem to lack homologues of this protein (Brandizzi et al., 2004), whereas lamins are a major structural
component of animal nuclei and the nuclear matrix (Albrethsen et al., 2009). Interestingly, lamin-like plant proteins have been identified immunologically in plants (McNulty and Saunders, 1992; Minguez and Moreno Diaz de la Espina, 1993), suggesting conservation of some epitopes in spite of extensive divergence of the nucleotide sequence. However, it remains unclear whether or not plant nuclear matrices contain functional equivalents of animal lamin proteins.

**Mapped S/MARs in Plants**

The presence and conservation of S/MARs in plants is well established, suggesting that high-order chromatin organization in plants does not differ greatly from other eukaryotes. Characteristics of plant S/MARs that have been studied to date are summarized in Table 1. The data are derived from mapped S/MARs within small genomic regions.

Most plant S/MARs have been isolated from DNA close to genes (reviewed in Holmes-Davis and Comai, 1998), although such distribution may be biased by the fact that most of the data come from studies of individual genes and their neighboring sequences. However a limited number of studies have mapped S/MARs over larger regions. For example, studies of chromatin organization along the Sh2/A1-homologous regions of rice (30 kb) and sorghum (50 kb) revealed the presence of 4 and 7 S/MARs, respectively (Avramova et al., 1998). The sequence characteristics of the rice and sorghum S/MARs showed a general enrichment in AT content, which was 70 to 80%. Despite the fact that the nucleotide sequences in S/MARs were not detectably conserved, the general organizational patterns of
S/MARs relative to the neighboring genes were maintained. All identified genes were contained in individual loops that were of comparable size for homologous genes. In addition, most of the S/MARs identified in the two genomic regions were associated with miniature inverted repeat transposable elements (MITEs), which are short AT-rich DNA sequences, have potential to form DNA secondary structure, and are flanked by inverted repeat sequences (Wessler et al., 1995). This observation raises a question related to evolution as to whether the MITEs carry S/MAR activity or MITEs tend to insert close to S/MARs, or perhaps both.

A subsequent study (Tikhonov et al., 2000) found that S/MARs were located at similar positions within the alcohol de-hydrogenase (adh1) regions of maize (225 kb) and sorghum (78 kb), leading to very similar maps when repetitive maize DNA was removed from the comparison. S/MARs frequently flanked individual genes, such as adh1. The S/MARs were divided as either “durable” or “unstable”, based on the degree of matrix association in the presence of competitor DNA. The durable S/MARs retained nuclear matrix association at the highest concentration (400 mg/mL) of E. coli competitor DNA and the majority were located in intergenic regions, consistent with a structural role in delineation of chromatin domains. The unstable S/MARs failed to bind the nuclear matrices at high competitor DNA concentrations, and were found mainly within introns. The authors suggested that S/MARs possess both chromatin domain-defining and regulatory roles. In addition, MITEs were often enriched in S/MAR sequences, as previously observed (Avramova et al., 1998).
Early *Arabidopsis* studies showed that the plastocyanin (16 kb) and the ATH1 (11.2 kb) loci are also organized in chromatin loops delineated by S/MARs (van Drunen et al., 1997). Using exogenous binding assays with rat nuclear matrices, the authors identified and characterized 3 S/MARs across each locus. A detailed analysis of the plastocyanin locus showed that the 3 S/MARs divided the locus into two small 5-kb chromatin domains, containing two genes each. The S/MAR sequences were AT-rich (>70%), contained A-box and T-box motifs, putative topoisomerase II binding sites, DNA unwinding sequences, and a MAR recognition signature (TAWAWWNNAWWRTAANNWWG) that is not found elsewhere in the region. Moreover, the 3 S/MARs were located within intergenic regions.

Recently, Tachiki et al. (2009) used a PCR-based assay to find 12 S/MARs within the intergenic regions of an 80-kb region of *Arabidopsis* chromosome 5. Typically the S/MARs flanked the ends of genes and had an interval of approximately 4.7 kb, with sizes ranging from 1 to 11 kb. The presence of S/MARs near 5’ ends of genes is consistent with the idea that S/MARs may influence promoter function for a subset of genes, but the authors did not find a correlation between S/MARs and endogenous gene expression.

One characteristic frequently mentioned for S/MARs is the presence of *Drosophila* topoisomerase II (Topo II) recognition motifs, and it is often assumed that the presence of Topo II recognition sites in MARs means that Topo II plays an important role in the nuclear matrix (Earnshaw et al., 1985; Yu and Moreno Diaz de la Espina, 1999; Hizume et al., 2007). Paul and Ferl (1998) treated *Arabidopsis* nuclei with the cytotoxic drugs VM26 and genistein, which create cleavable complexes at topoisomerase II sites, or, alternatively, limited exposure to DNase I. Both treatments produced median fragment sizes near 25 kb
(Paul and Ferl, 1998). The authors suggest that the 25 kb size represented structural S/MARs with very high affinity, which they termed Loop Basement Attachment Regions (LBARs). Paul and Ferl postulated that the higher median sizes of the LBAR intervals (loop), compared to the smaller S/MAR interval (loop) size, suggests that structural S/MARs are defined by the topoisomerase II recognition motifs, whereas other classes of S/MARs with lower binding affinity also exist. This interpretation is also consistent with the data of Tikhonov et al. (2000).

**Goals of this Thesis**

It has often been suggested that S/MARs organize chromatin domains and influence gene expression and DNA replication. Although there is evidence for each of these functions, a comprehensive understanding is lacking because the use of small-scale approaches has limited the number of examples that can be studied. A more global analysis is necessary to answer a variety of questions such as. How are S/MARs distributed across the genome? Are S/MARs associated with different genomic features and are S/MARs typically AT-rich as previously suggested? What is the nucleosomal organization at S/MAR sequences and do they define regions of accessible chromatin? Are S/MARs associated with specific epigenetic features such as certain histone modifications or DNA methylation? What role do S/MARs play in transcriptional regulation?

I have approached these questions by mapping the S/MARs on *Arabidopsis* chr4 using a high-resolution tiling array. The overall goals of my thesis are to advance our
understanding of the relationship between S/MARs and chromatin organization and to explore the possible role of S/MARs in endogenous gene expression and DNA replication in plant genomes.

Table 1. Plant S/MAR Characteristics.

<table>
<thead>
<tr>
<th>Source of S/MAR(s)</th>
<th>S/MAR Validation Assay</th>
<th>S/MAR Genomic Location</th>
<th>S/MAR Transgenic Studies</th>
<th>S/MAR Features</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean</td>
<td>Exogenous binding</td>
<td>Near the 3’end of the heat shock gene Gmhs17.6-L</td>
<td>Enhanced expression of the S/MAR-flanked transgene</td>
<td>395 bp in length; contains a topoisomerase II binding site.</td>
<td>(Schoffl et al., 1993)</td>
</tr>
<tr>
<td>Potato</td>
<td>Exogenous binding</td>
<td>Intron of ST-LS1 gene</td>
<td>NA</td>
<td>1.3 kb in length; 67% AT-rich; contains topoisomerase II binding sites.</td>
<td>(Mielke et al., 1990)</td>
</tr>
<tr>
<td>Tobacco</td>
<td>Exogenous and endogenous binding</td>
<td>Downstream of rb7-5A gene</td>
<td>Enhanced expression of the S/MAR-flanked transgene</td>
<td>1.1 kb in length; 73% AT-rich; contains topoisomerase II binding sites; AT-tracts of &gt;20 bp observed.</td>
<td>(Hall et al., 1991); (Allen et al., 1996)</td>
</tr>
<tr>
<td>Soybean</td>
<td>Exogenous binding</td>
<td>Upstream of lectin gene</td>
<td>Decreased expression of the S/MAR-flanked transgene</td>
<td>1.5 kb in length; contains a 520 bp region; 78% AT-rich.</td>
<td>(Breyne et al., 1992)</td>
</tr>
<tr>
<td>Pea</td>
<td>Exogenous binding</td>
<td>Downstream of plastocyanin gene</td>
<td>NA</td>
<td>540 bp in length; 77% AT-rich; contains topoisomerase II binding sites, A-box, T-box, and ARS-like motif.</td>
<td>(Slatter et al., 1991)</td>
</tr>
<tr>
<td>Maize</td>
<td>Exogenous and endogenous binding</td>
<td>Promoter region of Adh1-S gene</td>
<td>NA</td>
<td>688 bp in length; co-map with DNase I hypersensitive site.</td>
<td>(Avramova and Bennetzen, 1993)</td>
</tr>
</tbody>
</table>
Table 1. Continued.

<table>
<thead>
<tr>
<th>Common bean</th>
<th>Exogenous binding</th>
<th>Upstream region of β-phaseolin gene</th>
<th>Enhanced expression of the S/MAR-flanked transgene</th>
<th>Length (bp); AT-rich; Binding sites.</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common bean</td>
<td>Exogenous binding</td>
<td>Downstream region of β-phaseolin gene</td>
<td>Enhanced expression of the S/MAR-flanked transgene</td>
<td>1.2 kb; 72.5% AT-rich; A-box, T-box, and topoisomerase II binding sites.</td>
<td>(van der Geest et al., 1994)</td>
</tr>
<tr>
<td>Tomato</td>
<td>Exogenous binding</td>
<td>Upstream region of HSC80 gene</td>
<td>Enhanced expression of the S/MAR-flanked transgene</td>
<td>1.5 kb in length.</td>
<td>(Chinn and Comai, 1996)</td>
</tr>
<tr>
<td>Rice</td>
<td>Endogenous binding</td>
<td>NA</td>
<td>NA</td>
<td>504 bp; 44% AT-rich.</td>
<td>(Nomura et al., 1997)</td>
</tr>
<tr>
<td>Rice</td>
<td>Endogenous binding</td>
<td>NA</td>
<td>NA</td>
<td>366 bp; 64% AT-rich; A-box, T-box, and topoisomerase II binding sites.</td>
<td>(Nomura et al., 1997)</td>
</tr>
<tr>
<td>Tobacco</td>
<td>Exogenous binding</td>
<td>Upstream region of CHN50 chitinase gene</td>
<td>NA</td>
<td>699 bp; 75% AT-rich; A-box, T-box, and curved DNA sequence.</td>
<td>(Fukuda, 1999)</td>
</tr>
<tr>
<td>Tobacco</td>
<td>Exogenous binding</td>
<td>Upstream region of CHN50 chitinase gene</td>
<td>NA</td>
<td>850 bp; 74% AT-rich; topoisomerase II binding sites, curved DNA sequence.</td>
<td>(Fukuda, 1999)</td>
</tr>
</tbody>
</table>
Table 1. Continued.

<table>
<thead>
<tr>
<th>Plant</th>
<th>Binding</th>
<th>5’ flanking region</th>
<th>Expression of the S/MAR-flanked transgene</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea</td>
<td>Exogenous</td>
<td>vicilin gene</td>
<td>2.5 kb in length; 79% AT-rich; contains minimal S/MAR consensus sequence.</td>
<td>(Liu and Tabe, 1998)</td>
</tr>
<tr>
<td></td>
<td>binding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>Exogenous</td>
<td>NA</td>
<td>Enhanced expression of the S/MAR-flanked transgene</td>
<td>2.3 kb in length; contains ARS-like motifs.</td>
</tr>
<tr>
<td></td>
<td>binding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobacco</td>
<td>Exogenous</td>
<td>NA</td>
<td>Enhanced expression of the S/MAR-flanked transgene</td>
<td>1 kb in length; 62.8% AT-rich; contains two unwinding sites; T-box, and topoisomerase II binding site.</td>
</tr>
<tr>
<td></td>
<td>binding</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NA = Not Available.
REFERENCES


Klehr, D., Maass, K., and Bode, J. (1991). Scaffold-attached regions from the human interferon .beta. domain can be used to enhance the stable expression of genes under the control of various promoters. Biochemistry 30, 1264-1270.


CHAPTER 2

Diversity of Scaffold/Matrix Attachment Regions (S/MARs) in Arabidopsis is Revealed by Analysis of Sequence Characteristics, Nucleosome Occupancy, Epigenetic Marks, and Gene Expression.

Introduction

The genomes of eukaryotes are organized into chromatin to ensure accuracy of vital nuclear processes such as DNA replication, transcription, and repair (van Driel et al., 2003; Bloom and Joglekar, 2010; Naumova and Dekker, 2010; Fransz and de Jong, 2011). The basic component of chromatin is the nucleosome, which consists of ~146 base pairs (bp) of DNA wrapped twice around a histone octamer (Luger et al., 1997; Khorasanizadeh, 2004; Kornberg and Lorch, 2007). Nucleosomes cover most of the DNA, are separated from each other by 10-100 bp of linker DNA, and form a ‘beads-on-a-string’ filament with a diameter of ~ 10 nm, constituting the first level of chromatin organization (Luger et al., 1997; Woodcock and Ghosh, 2010). Most models postulate that the 10-nm fiber is then coiled to form a structure approximately 30 nm in diameter (Dorigo et al., 2004; Robinson et al., 2006; Woodcock and Ghosh, 2010; Fransz and de Jong, 2011). The precise spatial arrangement of the nucleosomes in this fiber is still unclear, and whether the 30-nm fiber can actually form under in vivo conditions is a subject of current debate (Eltsov et al., 2008; Maeshima et al., 2010; Fussner et al., 2011).
Early studies of electron micrographs of histone-depleted chromosomes suggested that 30-nm chromatin fibers form loops through interactions with a proteinaceous nuclear scaffold or matrix (Benyajati and Worcel, 1976; Paulson and Laemmli, 1977). To visualize the higher-order chromatin loops, it is important to remove the histones, which is normally done by extracting with either 2M NaCl or lithium diiodosalicylic acid (LIS). Histone removal extends the DNA, which can be seen to form loops when visualized by electron microscopy (Paulson and Laemmli, 1977). These loops are attached at their bases to a substructure referred to as the nuclear matrix or nuclear scaffold. Attachments occur at specific loci, which have been designated Matrix Attachment Regions (MARs) (Cockerill and Garrard, 1986), or Scaffold Attachment Regions (SARs) (Mirkovitch et al., 1984; Gasser and Laemmli, 1986b). It is common in the literature to refer to S/MARs, thereby to including both terminologies. I will follow that convention in this thesis.

S/MARs are operationally defined as (i) endogenous DNA fragments that co-purify with the nuclear matrix and remain bound after the unattached DNA is removed or (ii) exogenous DNA fragments that show specific binding to the nuclear matrix in the presence of excess non-specific competitor DNA (Cockerill and Garrard, 1986). Sequence analysis of S/MARs shows a lack of clear homology, except for being AT-rich (>70%). However, a DNA fragment that is AT-rich may not be a S/MAR, suggesting that secondary or even tertiary structure may be important. Early studies comparing a limited number of S/MARs found several sequence motifs and structural features thought to be important for matrix binding (Cockerill and Garrard, 1986; Nelson, 1986; Amati and Gasser, 1988; Boulikas, 1993; Liebich et al., 2002). While such features are shared among the small number of
S/MARs typically compared, it has become clear that S/MAR binding is complex and examination of many additional S/MAR sequences will be required to validate the generality of any proposed features (Michalowski et al., 1999).

Different nuclear functions have been attributed to S/MARs. For example, S/MARs are often thought to play a critical role in defining structural units of chromatin, functioning as boundary elements to define regions of condensed or open chromatin structure (Geyer, 1997; Holmes-Davis and Comai, 1998). Certain classes of S/MARs have also been implicated more directly in regulation of transcription, including stabilization of transgene expression (Stief et al., 1989; Phi-Van et al., 1990; Allen et al., 1993; Spiker and Thompson, 1996; Halweg et al., 2005). Mapping of S/MARs directly to or in the vicinity of origins of replication and the functioning of S/MARs as Autonomously Replicating Sequence (ARS) elements also suggests a role for S/MARs in DNA replication (Amati and Gasser, 1990; Vaughn et al., 1990; Mesner et al., 2003).

These reports suggested that S/MARs are involved in a variety of genomic functions, but a comprehensive understanding of their roles in plants is lacking because studies thus far have been limited to relatively small genomic regions (Tikhonov et al., 2000; Tachiki et al., 2009). The possibility that S/MARs have multiple roles within the nucleus suggested that different types of S/MARs exist, presumably with different properties and functions.

The advent of new tools for evaluating genome function has provided an opportunity to examine S/MARs on a genome-wide scale. Recently, S/MARs in the human genome were mapped by hybridizing matrix-bound DNA to microarrays containing probes for either human chromosomes 14 to 18 or the genomic regions studied in the ENCODE pilot project,
respectively (Linnemann et al., 2009; Keaton et al., 2011). The terms SAR and MAR are often considered interchangeable, as noted above. However, Keaton et al. (2011) used LIS to extract histones, and referred to the attachment regions thus obtained as SARs, while Linnemann et al. (2009) referred to attachment regions in LIS-extracted nuclei as SARs and those in salt-extracted nuclei as MARs. Linnemann et al. (2009) compared data from both isolation protocols, finding that 50% of identified S/MARs are detected by both protocols. However, they also showed that SARs are frequently located near the 5’ region of genes producing transcripts, whereas MARS are more often found within silenced genes.

Keaton et al. (2011) compared the SARs within the 30-Mb portion of the human genome studied by the ENCODE project. The ENCODE data appear to agree with the SAR results from Linnemann et al. (2009), as ENCODE SARs map near expressed genes, especially near transcription start sites (TSS). ENCODE SARs are also preferentially associated with RNA polymerase II binding regions, expressed genes, and early replicating regions. SARs were found less frequently in late replicating regions.

Here, we have used high-resolution tiling arrays to map S/MARs –SARs in the terminology of Linnemann et al. - on chromosome 4 (chr4) of the model plant *Arabidopsis thaliana*. We find that *Arabidopsis* S/MARs can be divided into five clusters based on their location in relation to other previously identified genomic features and show that the S/MARs are a diverse group of elements with multiple functions within the genome. While many of the *Arabidopsis* chr4 S/MARs appear to define structural domains, others are significantly enriched at transcription start sites (TSS) of genes and have a highly significant association with genes that encode transcription factors (TFs). Our data, along with
additional genomic information for S/MARs from other organisms, will help refine our ideas of the roles that different types of S/MARs play within the nucleus.

**Materials and Methods**

**Plant Materials and Growth Conditions**

The *Arabidopsis* cell line (Col-0, Columbia ecotype)(Calikowski et al., 2003), which was kindly provided by Dr. Iris Meier, was grown in Gamborg’s B5 basal medium with minor salt (Sigma 5893) supplemented with 1.1 mg/L 2,4-dichlorophenoxyacetic acid, 3 mM MES and 3% sucrose. The cells were grown on a rotary shaker at 160 rpm under constant light at 23ºC and subcultured every 7 days with a 1:10 (inoculum:fresh medium) dilution (Tanurdzic et al., 2008). Cells were harvested from a ‘7-d split culture’ by mixing 25 mL of fresh medium and 25 mL of the *Arabidopsis* culture at 7 days post subculture. The 7-d split culture was grown for 16 h and then collected for protoplast preparation (Lee et al., 2010).

**Isolation of Nuclei**

Protoplasts prepared from 200 mL of 7-d split cultures were suspended in 100 mL of nuclear isolation buffer as previously describe (Hall et al., 1991) with slight modifications. Nuclei isolation buffer contains 0.5 M of hexylene glycol, 20 mM HEPES pH 7.4, 20 mM KCl, 0.5 mM EDTA, 0.5% Triton X-100, 1% thiodiglycol, 0.05 mM spermine, 0.125 mM spermidine, 1 mM PMSF, and aprotinin at 2 µg/mL. This procedure gently lyses the plasma membrane and releases the nuclei. The nuclei were then incubated on ice for 5 min and
filtered through a tier of 100-µm, 50-µm and 30-µm nylon mesh to remove cellular debris and purified by centrifugation at 4°C in nuclear isolation buffer containing 15% Percoll at 600 x g for 10 min. The nuclear pellet was washed two times with nuclear isolation buffer without Triton X-100, adjusted to 50% glycerol, and stored at -80°C.

**Preparation of Nuclear Halos and Characterization of Proteins**

Histone removal from nuclei results in the formation of unpackaged DNA loops that can be visualized as a DNA halo, which remains attached to a core structure called residual nucleus, or nuclear scaffold or matrix (Paulson and Laemmli, 1977). Nuclear halos were prepared as previously described (Hall et al., 1991) with some minor modifications. Approximately 2 x 10^6 nuclei in 1 mL aliquots were stabilized for 15 min at 37°C with 1 mM CuSO_4 and the histones and other soluble proteins were removed by extraction with 2 mL of halo isolation buffer 2 (HIB2) (20 mM HEPES pH 7.4, 2 mM EDTA, pH 7.4, 0.1% digitonin, 0.5 mM PMSF, aprotinin at 2 µg/mL, 10 µM E-64, 100 mM lithium acetate) and 10 mM lithium 3,5-diiodosalicylate (LIS), established as an optimal concentration for removal of histones while leaving the matrix proteins intact in *Arabidopsis* cells (Figure 3). The nuclear halos were then centrifuged at 4°C, washed three times with digestion buffer (10 mM Tris-HCl pH 8.0, 20 mM KCl, 50 mM NaCl, 10 mM MgCl, 1 mM dithiothreitol, 0.1% digitonin, 1% thiodiglycol; 0.05 mM spermine, 0.125 mM spermidine, 0.5 mM PMSF, aprotinin at 2 µg/mL, 10 µM E-64, 1 mM phenantroline), and used immediately for either
isolation of MAR DNA, preparation of nuclear matrices, or characterization of the associated proteins.

**Analysis of nuclear matrix proteins by gel electrophoresis**

To determine the optimal LIS concentration, the nuclear halo proteins from nuclear halos prepared with various concentrations of LIS were solubilized in SDS sample buffer, resolved on precast 8-16% gradient SDS/polyacrylamide gels (Pierce), and stained with Coomassie Blue. Figure 3 illustrates that 10 and 20 mM of LIS selectively removed histones from *Arabidopsis* nuclei. However, we used 10 mM LIS in all *Arabidopsis* nuclear matrix preparations because higher concentrations may remove nonhistone proteins that are part of the nuclear matrix.

**Isolation of S/MAR DNA**

Nuclear halos from 4 x 10⁶ nuclei were resuspended in 700 µl of digestion buffer, containing 500 units of *Eco*RI and 500 units of *Hind*III and incubated at 37°C for 3 h to digest the nuclear halo DNA that was not attached to the nuclear matrix. The S/MAR DNA that remained bound to the nuclear matrix, was isolated by centrifugation (2000 x g for 10 min) and washed twice with digestion buffer. S/MAR DNA was purified by digesting the nuclear matrix proteins with an overnight incubation in lysis buffer (1% SDS, proteinase K at 500 µg/mL, 20 mM EDTA pH 8.0, 20 mM Tris-HCl pH 8.0), which was followed by phenol/chloroform/IAA extraction. The upper aqueous phase containing the S/MAR DNA
was mixed with 4µl of 150 µg/mL DNA co-precipitant GlycoBlue (Ambion) and the S/MARs were precipitated with 0.3 M sodium acetate and 2 volumes of cold ethanol. The S/MARs were centrifuged, and the resulting pellet was washed with 70% ethanol, dried for 10 min, and resuspended in RT-PCR grade water (Ambion).

**Microarray Hybridization**

Isolated S/MAR DNA (target DNA) and input sonication-sheared *Arabidopsis* genomic DNA (reference DNA) samples were amplified by adapting the standard protocol for whole-genome amplification (WGA) using the GenomePlex WGA1 kit (Sigma-Aldrich) as described (O'Geen et al., 2006). Briefly, the initial random fragmentation step of the WGA procedure was omitted, as the DNA was already been sonicated, and 10 ng each of target and reference samples were used for 14 cycles of linear amplification in the initial stage. Following the linear amplification step, 14 cycles of logarithmic amplification were carried out, starting with 10 ng of the linear amplification product. Amplified DNA was purified and concentrated to 200-250 ng/µL using a QIAquick PCR purification kit (QIAGEN). Target and reference amplified samples (1.5 µg) were labeled with Cy5 and Cy3 fluorescent dye-labeled 9mer respectively (TriLink Biotechnologies), and incubated with 100 units (exo-) Klenow fragment (New England Biolabs) and dNTP mix (10 mM each in Milli-Q water pH 5.0, New England Biolabs) for 3 hr at 37°C. Reactions (final volume 100 µl) were terminated by addition of 10 µl of 0.5 M EDTA pH 8.0, precipitated with 11.5 µl of 5 M NaCl and equal volume of isopropanol and resuspended in water. Then, 13 µg of the Cy5-labeled target and
13 µg of the Cy3-labeled reference were obtained from three labeling reactions, which were mixed, evaporated, and resuspended in 40 µl of NimbleGen Hybridization Buffer (Roche). The Cy-dye labeled target and reference samples were hybridized on a custom-designed NimbleGen tiling array for *Arabidopsis* chr4 in a water bath for 16 hr at 42°C. This experiment was composed of six microarrays representing three biological and three technical replicates. The arrays were washed using NimbleGen Wash Buffer System (Roche), dried by centrifugation, and coated with DyeSaver2 as described (Lee et al., 2010). Hybridized microarrays were scanned at 5-µm resolution using a Perkin Elmer ScanArray scanner and quantified using GenePix Pro software (version 6.01) as previously described (Lee, et al., 2010).

**Microarray Data Analysis**

All microarray normalization and analysis was performed with Limma and Bioconductor (Smyth, 2005; RDC, 2009). Microarray probes that map to the heterochromatic knob and pericentromere were filtered from the analysis before normalization due to their high level of repetitive DNA. For this experiment, we had three biological replicates each with two technical replicates. Microarrays were loess and quantile-normalized. On this array, each probe is present in duplicate, so the ratio averages of the duplicate probes were used to obtain a single value for each probe pair. Initially, a linear model was used to fit each biological replicate. The Pearson correlation for the probe ratios between the three biological replicates was greater than 0.78, so the biological replicates were treated as technical replicates for subsequent analyses. Final probe ratios were calculated in Limma using a
simple linear model on the six replicate arrays. Thus, the probe ratios used for mapping S/MARs are the result of twelve measurements, which included three biological replicates each with two technical replicates and duplicate probes on each array.

For peak finding, the normalized and averaged probe ratios were analyzed with the NimbleScan ChIP peak-finding function. This software allows the user to set two parameters, e.g. window size and minimum positive probe counts. NimbleScan uses a sliding threshold of the probe ratios to determine whether a probe is positive. We used window sizes of 300, 400, 500, 600 and 700 bp requiring three, four, five, five or six positive probes respectively. NimbleScan estimates a False Discovery Rate (FDR) for all peaks by a permutation process; peaks with an estimated FDR > 0.05 were removed. The remaining peaks for each set of parameters were then combined to arrive at our final set of 1358 putative S/MARs. As a control, we also performed peak-finding on negatively enriched probes, resulting in a set of 1497 negative peaks or Anti-S/MARs.

**Exogenous S/MAR Binding Assay**

S/MAR and non-S/MAR PCR fluorescent end-labeled fragments were obtained using the three-primer method described by Schuelke (2000) (Figure 4). The PCR mix contained 10-20 ng of template DNA, 2 mM MgCl₂, 200 µM each deoxynucleotide triphosphate, 0.2 µM forward primer, which included an M13 -21 sequence tail [TGTTAAAACGACGGCCAGT], 1 µM reverse primer, 0.05 µM fluorescently labeled M13 primer, Amplitaq 360 PCR reaction buffer, and one unit of Amplitaq DNA polymerase
The M13 primer was labeled using a fluorescent dye containing a 700 nm near-infrared dye (LI-COR) at the 5’-end. Unique forward and reverse primers were designed to flank the putative S/MAR region to produce PCR fragments that ranged from 300 bp to 1 kb. Forward and reverse primer sequences for S/MAR and non-S/MAR fragments are shown in Table 1 and 2 respectively. PCR was then used to produce a labeled fragments using the following conditions: initial denaturation at 94°C for 3 min; followed by 15 cycles of 30 s at 94°C, 30 s at 55°C, and 60 s at 72°C; followed by 25 cycles of 30 s at 94°C, 30 s at 50°C, and 60 s at 72°C; and a final extension time of 7 min at 72°C. End-labeled PCR fragments were then purified using a QIAquick PCR purification kit (Qiagen) and resuspended in RT-PCR grade water.

The fluorescently labeled fragments were used for in vitro binding assays using isolated NT-1 tobacco nuclear matrices as previously described (Hall et al., 1991; Allen et al., 1996; Michalowski et al., 1999). The nuclear halos were prepared as described above and digested with 500 units of EcoRI and HindIII at 37°C for 3 hr. The resulting nuclear matrices were pelleted and washed twice with binding buffer (70 mM NaCl, 20 mM HEPES pH 7.4, 20 mM KCl, 10 mM MgCl₂, 0.1% digitonin, 1% thiodiglycol, 0.2 mM PMSF, aprotinin at 5 µg/mL, 10 µM E-64, 1 mM phenanthroline). A typical binding assay contained approximately 6 x 10⁴ NT-1 nuclear matrices incubated at 37°C with frequent mixing for 3 hr with 10 ng of end-labeled PCR fragments and 10 µg of sonicated Escherichia coli genomic DNA. The nuclear matrices were centrifuged at 2000 x g for 10 min at 4°C, and the pellet, containing matrix-bound DNA fragments, and the supernatant, containing nonbinding fragments, were separated. The pellet fraction was washed once with digestion buffer without protease
inhibitors and incubated with 50 µL lysis buffer at 37°C for 16 hr. Equal fractions (usually 50%) of input DNA fragments (total), pellet and supernatant fractions were separated on a 1.5% agarose gel in TAE buffer. The gel was scanned by an infrared imagining system (Odyssey, LI-COR), immediately after electrophoresis to detect the 700-nm fluorescence.

**Comparison of Experimental S/MARs with Predicted S/MARs**

To determine the overlap between our S/MAR data set and the *in silico* predicted S/MARs (pS/MARs) for *Arabidopsis* (Rudd et al., 2004), we mapped the pS/MARs to the TAIR10 coordinate system. A lift-over was not available for this purpose so we used Bioconductor to map 50-bp tags from the start and end of each pS/MAR sequence to the TAIR10 chr4 sequence (Swarbreck et al., 2008). S/MARs and pS/MARs were designated as overlapping unless they were separated by a gap of 1 base or more. We removed all pS/MARs that mapped to the heterochromatic knob and pericentromere from this analysis.

**K-means Clustering of S/MARs and Sequence Analysis**

To develop a simple method of categorizing of S/MARs, we performed K-means clustering of S/MARs using the R Stats Package (RDC, 2009). We used three measurements for the genomic context of each S/MAR: gene content, exon content and TE content. These measurements were determined by the overlap of each S/MAR with these annotated genomic features (TAIR10) (Swarbreck et al., 2008) (Figure 12). We initially scanned the range of 2 to 25 clusters (Figure 1), but decided to use 5 clusters because it partitioned the S/MARs into
clusters that seemed biologically relevant. For the final clustering, we used the Hartigan-Wong algorithm, 20 maximum iterations and 1000 random starts, which is summarized in Table 4. We calculated the AT content of each S/MAR in Bioconductor. To compare the AT content profiles of the S/MARs clusters, we made a histogram for each S/MAR cluster, binning the AT content by 2 percent increments. For comparison, we used the mean AT content for Chr4 exclusive of the knob and pericentromere, regions where we could not map S/MARs.

To identify AT-rich kmers, we scanned each S/MAR with a 10-bp window incremented by 1 bp and flagged any window with 9 or more As and Ts. Overlapping windows were then merged, and the AT content for the resulting kmers was recalculated. As a quantitative measure of AT-rich kmer content, we divided the number of As and Ts in AT-rich kmers for each S/MAR by the total length of that S/MAR. As controls, we also calculated the AT-rich kmer content of 1400 randomly selected EcoRI/HindIII fragments and the 1497 Anti-S/MARs.

S/MARs and Transposable Elements (TEs)

To determine the significance of the association between S/MARs and TEs, we considered the TEs by their superfamily (TE-SF) according to TAIR10. We calculated the base pair overlap between each S/MAR and each TE-SF. If the association of the S/MARs with the TE-SFs is only random, then we expected that the TE-SF content of the S/MARs would be similar to the TE-SF content of Chr4 exclusive of the knob and pericentromere, and thus, the ratio of S/MAR TE-SF content to Chr4 TE-SF content or S/MAR TE-SF
enrichment should be equal to one. To test this null hypothesis, we used a permutation test to estimate a p value for our observed ratio. For the null dataset, we chose random start coordinates for 1358 regions but used the S/MAR lengths to assign end coordinates. We then calculated the overlap between these random S/MARs and the TE-SFs. The process was repeated 100,000 times to arrive at a null distribution for the ratio of S/MAR TE-SF content to Chr4 TE-SF content. The p value was then determined by the intersection of the observed ratio for the actual S/MARs with this null distribution, and a Bonferroni correction was used to adjust the p values for multiple testing.

To analyze the correlation between S/MAR enrichment in TE-SFs and the AT content of the TE-SFs, we first calculated the AT content of the TE-SFs in Bioconductor. We then fit a linear model of the S/MAR TE-SF enrichment as a function of TE-SF AT content. The resulting correlation was highly significant with an adjusted R-squared value of 0.5351 and a p value of 0.00034.

**S/MARs and Chromatin Structure**

We used two sources of data to analyze the chromatin structure of S/MARs and flanking sequences. A genome-wide predicted nucleosomal occupancy is available for Arabidopsis TAIR8 ([http://genie.weizmann.ac.il/software/nucleo_genomes.html](http://genie.weizmann.ac.il/software/nucleo_genomes.html)). To facilitate analysis with this large data set, we segmented Chr4 into 200 base pairs windows at the midpoints of our microarray probes. The average of the predicted nucleosomal occupancy for these windows was used in subsequent calculations, after adjusting to TAIR10 coordinates. ChIP-chip results for histone modifications and DNA methylation for our cell
cultures under the same culture conditions have been published (Tanurdzic et al., 2008). The microarray used for this analysis is lower resolution than our current platform, consisting of tiled PCR products with a mean length of approximately 1 kb.

For the analysis of S/MARs and flanking regions, we aligned all S/MARs at their midpoint and analyzed a region 5 kb upstream and 5 kb downstream of the S/MAR midpoints. This region was segmented into twenty 500-bp non-overlapping windows. Thus, the two center windows overlap the S/MAR. The mean nucleosomal occupancy of these windows was then calculated as described above. The mean nucleosomal occupancy of each window for each S/MAR cluster was calculated and the results were plotted. T-tests were performed to determine the statistical significance of the observed enrichment or depletion of mean predicted nucleosomal occupancy for the S/MARs (windows 10 and 11) as compared to the flanking region (windows 1-6 and 15-20). Windows 7-9 and 12-14 were excluded from this analysis because these regions display intermediate, transitioning values that may result from the inexact alignment of many S/MARs of different widths. Furthermore, the actual binding site of the nuclear matrix to the DNA may not lie at the midpoint of the S/MARs as defined by the microarray.

The process was similar for the histone modifications and DNA methylation. For these analyses, microarray results for each probe were considered only as qualitative, either the probe was considered enriched for the modification or not. We then calculated the mean number of enriched probes for each window, allowing the probes to overlap with more than one window, if appropriate. Statistical significance was determined on the mean number of enriched probes for the windows as above.
**Influence of S/MARs on Gene Expression**

For each S/MAR, we calculated the minimal distance to the closest gene transcription start site (TSS) or transcription termination site (TTS) as annotated in the TAIR10 genome release (Swarbreck et al., 2008). Histograms for the TSS and TTS distance for each S/MAR cluster were determined using 300-bp bins.

To determine whether the presence of a S/MAR can influence gene expression, we used two complementary analyses. First, the genes were grouped based on their S/MAR to TSS distance, with negative and positive values indicating upstream or downstream respectively after correcting for strand (Figure 2). All genes were then placed into bins in which distance values ranged from -1000 to +3000 into quintiles. Ideally, the number of genes in each bin would be the same, but not all genes are present on the ATH1 array so the number ranged from 160 to 192. The gene expression status was determined from published microarray results on the Affymetrix ATH1 array (Tanurdzic et al., 2008). For this analysis, we used the presence/absence calls (p<0.05) as determined by the MAS5 algorithm as implemented in the Affymetrix package for R (RDC, 2009). Gene activity was calculated for each bin as the fraction of genes in each bin that were expressed. Statistical significance was determined using the binomial test in the R stats package with the null mean as the fraction of all genes on Chr4 that are expressed. P values were adjusted by the Bonferroni correction for testing across the 5 bins, and 99% confidence intervals were calculated. As a control, we did a similar analysis using the S/MAR to gene TTS distance, analyzing the range -3000 to +1000 bp.
For the second analysis, we identified the most proximal gene for each S/MAR based on the S/MAR to gene TSS distance, which resulted in a set of 1216 genes for the 1358 S/MARs. The set does not contain 1358 genes because some genes are the most proximal neighbor to 2 or more S/MARs (Figure 2). These genes were grouped by the proximal S/MAR cluster and gene activity was analyzed as described above.

**Gene Ontology and Transcription Factor Enrichment in S/MAR Proximal Genes**

We performed gene ontology (GO) and transcription factor (TF) enrichment analysis using custom R scripts. For transcription factors, we used the transcription factors as annotated by AGRIS in addition to the GO categories (Yilmaz et al., 2011). The first step was to determine a reference gene set for chr4 exclusive of the heterochromatic knob and pericentromere. The S/MAR proximal genes are the 1216 genes determined above. Because we are limited to only Chr4 genes, we only analyzed GO categories that contained five or more S/MAR proximal genes. The Fisher Test was used to calculate p values, using the Bonferroni correction to adjust for multiple testing. Many of the enriched GO categories are associated with TFs. Therefore, we also determined which of the S/MAR proximal genes in these categories were also annotated by AGRIS as TFs.
Results

Identification of S/MARs

We found that 10 and 20 mM lithium diiodosalicylate (LIS) were optimal for extracting the histones from the nuclei (Figure 3) (Mirkovitch et al., 1984; Hall et al., 1991), however we decided to use 10 mM LIS as higher LIS concentrations remove other proteins that may be part of the nuclear matrix. Histone extraction unpackaged the genomic DNA that resulted in the formation of loops (nuclear halos) that remain bound to the nuclear matrix at the matrix attachment regions (S/MARs). The nuclear halos were digested with EcoRI and HindIII and centrifuged to separate S/MAR DNA (pellet) from loop DNA (supernatant) (Hall et al., 1991). Approximately 30% of the total nuclear halo DNA remained bound to the nuclear matrix pellet (Figure 5).

S/MAR DNA was hybridized to a custom-designed high-resolution NimbleGen tiling microarray for Arabidopsis thaliana Col-0 chr4. The heterochromatic Knob and pericentromeric regions were excluded from this analysis due to their high level of repetitive DNA. The four regions analyzed included the distal short arm (dsa), the proximal short arm (psa), the proximal long arm (pla), and the distal long arm (dla) (Figure 6A), which collectively cover 83% of chr4. Data from six microarrays representing three biological and three technical replicates were analyzed using a combination of the limma package and NimbleScan software to identify regions enriched for S/MAR DNA hybridization with false discovery rate (FDR) ≤ 0.05 compared to total genomic DNA. The correlation coefficient between biological replicates was above 0.78.
We identified and mapped a total of 1,358 S/MARs distributed along the chr4 dsa, psa, pla, and dla (Table 5). We found an average of one S/MAR every 11 kb. The distribution, while not uniform, was independent of either gene or transposable element (TE) density (Figure 6C and 6D, and Table 5). These results are similar to the LIS-based S/MAR mapping results for human (HeLa) chromosomes 14 to 18 (Linnemann et al., 2009).

**Exogenous S/MAR Binding Assays Validate Microarray Results**

The exogenous S/MAR binding assay is used to operationally define DNA fragments that show specific binding to isolated nuclear matrix in vitro (Figure 7) (Cockerill and Garrard, 1986). We used the exogenous S/MAR binding assay to test the veracity of our microarray results by monitoring the binding of 24 putative S/MARs on chr4 (Figure 8A and 8B, and Table 6) to nuclear matrices from the *Nicotiana tabacum* suspension cell line NT-1 (Hall et al., 1991). We did not use *Arabidopsis* nuclear matrices because chlorophyll contamination in these preparations caused a spurious fluorescent signal in the matrix fraction that interfered with quantification by the infrared imagining system. Figure 9 shows an example of our validation results for one of the *A. thaliana* chr4 S/MARs that was identified from our microarray analysis.

To determine the reliability of the statistical cut-offs used for the microarray results, 13 putative S/MARs, with FDRs between 0 and 0.028 and AT contents between 58 to 80%, were selected from the distal long arm. Despite the range of differences, all 13 S/MAR fragments bound to the nuclear matrices (Figure 8B, and Table 6). Furthermore, 8 non-S/MARs, which did not bind in the microarray experiments (hybridization signal less than
the average for the entire array), also did not bind in vitro (Figure 8B, and Table 7). Extending this analysis, we then tested eleven more putative S/MARs with higher FDRs (0.039 to 0.05). The 11 additional S/MARs had an AT content that ranged from 43 to 71% and represented all cytologically defined regions of chr4. A standard non-S/MAR fragment (corresponding to gene At4g19120) was included in each binding assay as an internal control. We found that 7 of these 11 S/MAR fragments bound to the tobacco nuclear matrix in vitro (Figure 8B, and Table 6). Together, 20 of the 24 S/MARs identified by the microarray analysis bound to tobacco nuclear matrices (Figure 8B, and Table 6); only the putative S/MARs with the higher FDRs and lower AT contents failed to bind.

In addition, 3 S/MARs from the Arabidopsis ATH1 (At4g32980) locus (Figure 10C), and 3 S/MARs from the plastocyanin locus (Figure 11) previously identified by exogenous binding assays (van Drunen et al., 1997), scored as positive S/MAR peaks in our microarrays. Taken together, we concluded that our microarray results show a high level of agreement with the exogenous S/MAR binding assay results (Figure 8B, and Table 6).

S/MAR Clusters and Genomic Features

S/MARs are present at both genic and intergenic regions of eukaryotic genomes (Mirkovitch et al., 1984; Cockerill and Garrard, 1986; Phi-Van et al., 1990; reviewed in Bode et al., 1996) and are important for maintenance of higher-order of chromatin structure (Paul and Ferl, 1998) and regulation of gene expression (Butaye et al., 2005; Thompson et al., 2007). To examine the relationship between S/MARs and genomic features, we created a database for computational analysis that combined our S/MAR map data, the most recent
Arabidopsis TAIR10 genome annotation (Swarbreck et al., 2008), the in silico S/MAR map for the Arabidopsis genome (Rudd et al., 2004), the nucleosome occupancy profile (Kaplan, et al., 2009), and gene expression and epigenetic data for our Arabidopsis thaliana Col-0 cell line (Tanurdzic et al., 2008). Inspection of this database showed that S/MARs are present at both genic and intergenic regions, are often but not always associated with TEs, and have a similar distribution in both heterochromatic (Figure 10B) and euchromatic (Figure 10C) regions.

Cluster analysis was used to determine whether the 1,358 experimental S/MARs could be partitioned based on the degree of association with various annotated genomic features (Figure 12). We developed five clusters using K-means clustering to group the S/MARs according to their degree of association with genes, exons, and TEs. S/MAR association with introns and unannotated regions, as well as the AT content, which is typically higher in S/MARs, were determined after the clustering (Figure 12 and Figure 13).

We initially scanned the range of 2 to 25 clusters (Figure 1), but we decided to use 5 clusters because additional clusters offered only incremental improvement, and the resulting 5 clusters seemed biologically relevant (Table 4). Figure 13 summarizes the partitioning of our experimental S/MARs into clusters according to their association with various genomic features. Red indicates the features with which the given S/MAR has the highest association, whereas blue indicates features least associated with that particular S/MAR. From the total 1,358 S/MARs identified, 424 are associated with unannotated regions (cluster A), 276 with TEs and high AT content (cluster B), 285 with boundaries between intergenic and genic
regions (cluster C), 231 with exons and low AT content (cluster D), and 142 with introns (cluster E).

**AT Content of Clustered S/MARs**

A comparison of the sequences for numerous previously identified S/MARs has shown that they are typically AT-rich (>70% AT) (Boulikas, 1993). Therefore, we examined the AT content of each S/MAR cluster. The S/MARs in clusters A, B, C and E were significantly enriched for AT content (Figures 13 and 14), with higher mean AT content (70, 73, 66, and 67% respectively) than the analyzed chromosomal regions (64%) (Table 5). In contrast, the S/MARs in cluster D had a mean AT content of 59% (Figures 13 and 14). These S/MARs are therefore slightly more GC-rich than the average for the regions we analyzed.

It is important to note that the AT content measurement includes all of the sequences showing enrichment on the microarray. While cluster D S/MARs are GC-rich overall, it is conceivable that they could contain AT-rich stretches sufficient to promote base-unpairing, which has been proposed as a basis for S/MAR binding to the nuclear matrix (Bode et al., 1992; Bode et al., 2006). Hence, we examined the S/MAR sequences for AT-rich k-mers (k=10) that could be considered potential regions for base unpairing. As controls, we also calculated the AT-rich k-mer content of 1400 randomly selected EcoRI/HindIII fragments and the 1497 “Anti-S/MARs” (microarray negative peaks). We found that the median AT-rich k-mer contents for S/MARs from cluster A, B, C, and E were higher than controls, whereas cluster D S/MARs was similar to or slightly lower than the controls (Figure 15). Importantly, a significant fraction (38%) of the GC-rich S/MARs in cluster D had no AT-rich
k-mers $\geq 10$ nt in length. Thus, the binding to the matrix of the putative S/MARs in cluster D must be explained in other ways.

Comparing Experimental S/MARs and Predicted S/MARs

An *in silico* analysis based on the SMARTTest program (Frisch et al., 2002), was previously used (Rudd et al., 2004) to predict S/MARs for the entire *Arabidopsis* genome. This analysis identified 2,892 S/MARs for the corresponding chromosomal regions of chr4. When we compared these *in silico* predicted S/MARs with our experimental S/MARs, 860 (63.3%) of the 1,358 experimental S/MARs matched the predicted S/MARs (Figure 16). However, these 860 S/MARs represent only 30% of the 2,892 *in silico* S/MARs (Figure 16).

To determine if certain S/MAR clusters were predominant in the Rudd et al. (2004) prediction, we compared the proportion of the predicted S/MARs that matched experimental S/MARs separately for each of the clusters. Strikingly, nearly all of the experimental S/MARs in cluster B matched predicted S/MARs (95%) (Table 8). Lower percentages were observed for the other clusters, with cluster A showing 73% matches, cluster C 59%, and cluster E 50%. Cluster D had the lowest proportion of matches (22%) (Table 8). Because the proportion of experimental S/MARs matching *in silico* predictions declines with decreasing AT content of the experimental S/MARs, these data suggested that the *in silico* algorithm relies too heavily on AT content.
S/MARs and TEs

Certain classes of TEs display nuclear matrix binding activity in *Drosophila* (Nabirochkin et al., 1998), human cells (Rollini et al., 1999), and plants (Avramova et al., 1998; Tikhonov et al., 2000). In addition, a survey for the presence of TEs in a collection of human experimental S/MARs found an enrichment of TE-derived sequences (Jordan et al., 2003). In *Arabidopsis*, the 276 S/MARs in cluster B were associated almost exclusively with TEs (Figure 10B and 10C, and Figure 13). An additional 271 S/MARs found in other clusters also had some association with TEs.

To determine whether S/MARs are preferentially associated with any of the eighteen specific TE Super Families (TE-SFs) annotated in *Arabidopsis* TAIR 10 (Swarbreck et al., 2008), we surveyed the amount TE sequences that were also present in the S/MAR sequences. We used this approach because the *Arabidopsis* TAIR 10 annotation includes many small fragments putatively derived from TEs as well as large intact TEs. S/MARs are fairly uniform in size and could include multiple TEs overlapping a single S/MAR or multiple S/MARs overlapping a single TE. We found that 547 (40.2%) of the 1,358 S/MARs contained TE sequence elements attributable to seventeen of the eighteen TE-SFs (Table 9). However, much of the TE sequence was not associated with S/MAR sequence (Table 7). Our data showed that S/MARs are significantly enriched in the TE-SFs *DNA, DNA/Harbinger, DNA/Mariner, DNA/MuDR,* and *RC/Helitron,* relative to their overall abundance for the corresponding chromosomal regions (Table 7). *LTR/Copia* is the only TE-SF depleted in S/MAR sequences (Table 7).
Since most S/MARs have an elevated AT content, could the association simply be due the coincidence that certain TE-SFs are also AT-rich? To test this hypothesis, we compared the enrichment of each TE-SF associated with a S/MAR as a function of the mean AT content for each of the TE-SFs using a simple linear model. From Figure 17, it appears that TE-SF enrichment with S/MARs is strongly correlated with the AT content of the TE-SF. However, it is not possible to make a clear conclusion as to whether the AT-richness of the S/MAR-associated TE-SFs is the cause or the effect of the biological activity associated with either element.

**Predicted Nucleosome Occupancy at S/MARs**

The nucleosomal organization of S/MARs is poorly understood (Bode et al., 1996). Many S/MARs sequences have DNA unwinding properties and are generally thought to have an increased likelihood of being nucleosome-free, especially in regions that interact with nuclear matrix proteins (Bode et al., 1996). Hence, we examined the probability of nucleosome occupancy at S/MARs using the nucleosome score model from Kaplan, et al (2009). For the analysis of MARs and flanking regions, we aligned all S/MARs at their midpoint and analyzed a region 5 kb upstream and 5 kb downstream of the S/MAR midpoints (see Materials and Methods).

Predicted mean nucleosome occupancy profiles and their relation to S/MAR position are shown separately for selected heterochromatic (Figure 10B) and euchromatic (Figure 10C) regions. Further analysis shows that the mean nucleosome occupancy patterns for S/MARs belonging to different clusters are distinct (Figure 17, and Table 10). S/MAR
sequences in clusters A, B, C, and E have lower probabilities of nucleosome occupancy, whereas S/MAR sequences in cluster D have a higher probability of nucleosome occupancy than flanking regions. These observations are consistent with previous reports showing that regions of high AT content exclude nucleosomes, whereas GC-rich regions promote nucleosome occupancy (Yuan et al., 2005a; Valouev et al., 2011).

**S/MARs and Epigenetic Marks**

We then examined the distribution of S/MARs in relation to selected epigenetic marks previously mapped on chr4 of our suspension culture cells grown under the same conditions (Tanurdzic et al., 2008). The epigenetic marks included DNA cytosine methylation (5mC), histone H3 lysine 9 dimethylation (H3K9me2), histone H3 lysine 4 di- and monomethylation (H3K4me1/2), and histone H3 lysine 56 acetylation (H3K56ac). S/MARs in each of the five clusters were scored for their association with each mark.

In *Arabidopsis*, DNA cytosine methylation (5mC) has been found in the coding region of genes in the euchromatic regions as well as in TEs and other repetitive sequences in heterochromatin regions (Zhang et al., 2006; Vaughn et al., 2007; Tanurdzic et al., 2008; Law and Jacobsen, 2010; Roudier et al., 2011). In contrast, we detected reduced 5mC in S/MAR clusters A, C, D, and E, whereas S/MAR cluster B showed no change relative to flanking regions (Figure 19, and Table 10).

The S/MAR clusters were then compared for the distribution of histone H3K9me2 (Figure 20), which is predominantly associated with heterochromatin in *Arabidopsis* (Baroux et al., 2011; Fransz and de Jong, 2011). We found that the H3K9me2 levels in the S/MARs in
clusters A, C, D, and E were relatively unchanged from the flanking chromatin regions, whereas a significant enrichment of H3K9me2 was found on S/MARs in cluster B (Figure 20, and Table 10). These data indicate that S/MARs that are associated with TEs (cluster B) are enriched for H3K9me2.

Methylation of H3K4 is mainly associated with euchromatin in both plants and animals specifically in gene coding regions that are either active or poised for transcription (Schübeler et al., 2004; Zhang et al., 2006; Barski et al., 2007; Tanurdzic et al., 2008). H3K4me3 has the strongest effect on gene expression in Arabidopsis (Zhang et al., 2006; Deal and Henikoff, 2011). We found that H3K4me1/2 is depleted in S/MAR clusters A, B, and C and to a lesser extent in clusters D and E relative to flanking regions (Figure 21, and Table 10).

Finally, we investigated the correlation between H3K56ac and S/MAR sequences in the different clusters. H3K56ac is associated with multiple biological processes that require an open chromatin conformation, including DNA replication, repair and transcription (Masumoto et al., 2005; Xu et al., 2005; Rufiange et al., 2007b; Li et al., 2008; Kaplan et al., 2009; Lee et al., 2010). H3K56ac is strongly enriched in gene promoter regions and early replicons, as well as the initiation zones of both early and late replicons in Arabidopsis, suggesting a role in transcription and DNA replication (Tanurdzic et al., 2008; Lee et al., 2010). We detected H3K56ac enrichment in S/MAR clusters C, D and E, whereas S/MAR clusters A and B showed no change relative to flanking regions (Figure 22, Table 10). These results indicated that S/MARs associated primarily with intergenic-genic boundaries, exons,
and introns are enriched for H3K56ac (Figures 13 and 22) and suggests that these regions have an open chromatin conformation.

**S/MARs and Gene Expression**

Comparing S/MAR positions with those of annotated genes, we found that S/MARs tend to be associated with the 5’ ends of genes (Figure 10B and 10C, and Figure 23A). The median distance between S/MAR midpoints and the nearest Transcription Start Site (TSS) is 480 bp, and 51% of the S/MAR midpoints are within a range of 500 bp upstream to 500 bp downstream of an annotated TSS, while only 17% of S/MAR midpoints are within 500 bp of a Transcription Termination Site (TTS) (Figure 23B). However, it is important to note that most of the chr4 genes are not associated with a S/MAR.

The clear tendency of S/MARs to be located in proximity to the TSS of a subset of genes suggested that S/MARs may influence promoter function for such genes. To test this hypothesis, we performed two complementary analyses. First, we grouped genes into several bins based on their S/MAR midpoint to TSS distance, with negative and positive values indicating upstream or downstream positions, respectively (Figure 2). We then determined whether the gene in question is transcribed, based on previously reported expression data for our cell line obtained from analysis with the ATH1 microarray platform (Tanurdzic, et al., 2008). Using the MAS5 algorithm from Affymetrix Micro Array Suite 5.0 (Hubbell et al., 2002), we determined the presence or absence (p<0.05) of detectable transcripts for each gene. We next calculated the fraction of active genes in each of the bins. As a control, we performed a similar analysis using the S/MAR to TTS distances, analyzing the range -3000
bp to +1000 bp. This analysis revealed two findings. First, genes with a TSS-associated S/MAR at -60 bp to +251 bp are more likely to be expressed than the average gene in the region analyzed (Figure 24A). Second, most of the active genes in this bin are associated with cluster C S/MARs, with smaller fractions associated with S/MARs in clusters D and E (Figure 24A). In contrast, when we applied a similar analysis to S/MAR-to-TTS distances we found no strong trend, with genes with TTS-associated S/MAR having a similar likelihood of expression as the average gene in the region analyzed (Figure 24B).

In the second analysis, we identified the closest gene to each S/MAR, using the distance from the S/MAR midpoint to the TSS (Figure 2). When a single TSS was associated with more than one S/MAR, the closest association was used. As a control, we used the same approach to determine the TTS closest to each S/MAR. Some S/MARs will not be associated with any TSS or TTS because they were positioned between other S/MARs and, as such, were excluded from both the TSS and the TTS analyses. We then grouped these genes by the cluster of the proximal S/MAR and analyzed the fraction of genes expressed in each group as described above. This analysis also showed that if the TSS of a gene is associated with a cluster C S/MAR, that gene is more likely to be expressed than the average gene (Figure 25A, and Table 11). Again, there is little correlation between gene activity and S/MARs position relative to the TTS (Figure 25B, and Table 11).

We used gene ontology (GO) enrichment analysis (Du et al., 2010) to investigate the biological function of genes with a TSS-associated S/MAR. This analysis revealed that the population of genes annotated as having sequence-specific transcription factor binding activity was significantly enriched genes with a TSS-associated S/MAR (Table 12). Of the
254 putative transcription factor (TF) genes identified by GO analysis on chr4, 113 (44%) have a TSS-associated S/MAR. To confirm that the GO-enriched genes were indeed transcription factors, we also checked their annotations in AGRIS (Yilmaz et al., 2011), which classifies TFs based on the presence of conserved domains. This analysis reduced the number of annotated TFs on chr4 to 197, of which 92 (47%) had a TSS-associated S/MAR (Table 12). Further, of the 92 TFs with a TSS-associated S/MAR, 31 (39%) of the 79 represented on the ATH1 array were expressed. This fraction is different than the fraction of all 215 genes with a TSS-associated S/MAR that contain expression data, of which 146 (67%) are expressed. These findings suggest that S/MARs are often found at or near TSS of genes of regulatory significance independent of their expression properties.

Discussion

Identification of S/MARs on chr4

We mapped 1358 S/MARs on Arabidopsis chr4 using high-resolution tiling arrays. Our analysis showed that Arabidopsis S/MARs can be divided into sub-groups based on their location in relation to other genomic features consistent with the idea that S/MARs are a complex group of elements with multiple functions. Earlier S/MARs mapping studies in plants were limited to smaller regions, ranging from an 80-kb region of A. thaliana chromosome 5 (Tachiki et al., 2009) to a 225-kb region from the maize adh1 locus (Tikhonov et al., 2000).
Previous studies used sequences from a limited set of characterized S/MARs to derive S/MAR prediction algorithms, which was then used to predict the global location of S/MARs for human chromosome 22 (Frisch et al., 2002) and the *Arabidopsis* genome (Rudd et al., 2004; Tetko et al., 2006). The predictive power of the existing algorithms is quite limited (Evans et al., 2007), especially for predicting GC-rich S/MARs. We found that the majority of the *Arabidopsis* chr4 S/MARs are indeed AT-rich with the exception of cluster D, which is slightly GC-rich compared to the average for the analyzed region (Figure 14). Thus, our experimental analysis allowed us to identify a subset of S/MARs that do not fit the parameters commonly used for S/MAR prediction.

Recent LIS-based S/MAR mapping studies for the human ENCODE regions (Keaton et al., 2011) and human chromosomes 14 to 18 from HeLa cells (Linnemann et al., 2009), showed that the average spacing between S/MARs is 44.2 and 88 kb, respectively. In *Arabidopsis*, small-scale S/MAR mapping studies (van Drunen et al., 1997; Tachiki et al., 2009) and genome-wide S/MAR prediction analysis (Rudd et al., 2004) suggested that the average interval between S/MARs is approximately 5 kb. In our study, we found that S/MARs on chr4 have an average spacing of 11 kb. However, we have observed shorter spacing between S/MARs in certain gene-rich regions similar to those studied by Van Drunen et al (1997) and Tachiki et al. (2009), and the average interval between S/MARs on *Arabidopsis* chr4 is obtained from both gene-poor and gene-rich regions. Since S/MAR spacing defines potential chromatin loops, the data of Keaton et al. (2011) and Linnemann et al. (2009) indicate the human genome is likely organized into larger chromatin loop domains than the *Arabidopsis* genome. Such differences may reflect differences in genome size or
gene density between the two genomes. However, in the *Arabidopsis* genome (excluding highly repetitive regions), our data showed that S/MARs are distributed along most of the length of chr4, and that their distribution along the entire chromosome is independent of either gene or TE density (Figure 6C and 6D).

**S/MAR Characterization**

Certain S/MARs have been proposed to play structural roles by anchoring sections of the genome as loop basements, or LBARs, which are distinguished from classical S/MARs (Paul and Ferl, 1998). Tikhonov et al. (2000) noted that S/MARs with higher binding affinity to the nuclear matrix (durable S/MARs) are usually found outside of gene domains and form the borders of chromatin loops, whereas the S/MARs with lower binding affinity are found within introns. Studies of other higher eukaryotes showed that S/MARs that are located near or within genes may play a transcriptional regulatory role (Anthony and Blaxter, 2007; Ng et al., 2009; Keaton et al., 2011). We found that the 424 cluster A S/MARs and 198 of the 276 cluster B S/MARs are located within unannotated intergenic regions (Figure 13) indicative of a potential role in delineating the chromatin loop domains as described above. Conversely, S/MARs in clusters C, D, and E are near or within genes and may function in other processes (Figure 13). This idea is supported by the observation that cluster C S/MARs, which are mainly associated with the TSS of a gene, are associated with an increased likelihood of gene expression (Figure 25A and Table 11). Our observation that S/MARs located at the TSS are associated with a greater probability that an endogenous gene will be expressed (Figures 23
A and 25) also extends previous work showing that a tobacco S/MAR can increase transgene expression (Allen et al., 1996; Abranches et al., 2005; Halweg et al., 2005).

**Arabidopsis S/MARs have Diverse Nucleotide Composition**

Previously characterized S/MARs have typically been AT-rich (Liebich et al., 2002). However, S/MARs with nucleotide compositions differing from those of the AT-rich S/MARs have also been reported (Boulikas, 1993; Liebich et al., 2002; Yusufzai and Felsenfeld, 2004). Our chr4 analyses reveals that AT-rich S/MARs are associated with unannotated regions, TEs, boundaries between intergenic and genic regions, and introns, whereas GC-rich S/MARs are associated with exonic regions (Figures 13 and 14).

While more study is necessary, it is tempting to speculate that the GC-rich exonic S/MARs may represent attachments in which transcriptionally active or poised chromatin interacts with components of the transcription machinery. This idea is consistent with models proposing that RNA polymerase II and general transcription factors are assembled at fixed transcription sites on the nuclear matrix (Xing and Lawrence, 1991; Kimura et al., 1999; Nickerson, 2001). Previous work by Gerdes, et al (1994) also directly demonstrated localization of active genes at chromatin loop bases by fluorescence *in situ* hybridization.

**GC-rich Arabidopsis S/MARs**

A clear consensus sequence for S/MARs has long been elusive, and S/MAR function is often considered depend less on primary sequence than on structural properties of S/MAR
DNA, such as the high unwinding and unpairing susceptibility of AT-rich sequences (Bode et al., 2003; Bode et al., 2006). The advent of high-throughput sequencing has provided new data that was not previously available and investigators have used existing or modified algorithms to predict S/MAR locations. A number of algorithms such as MAR-Finder, ChrClass, and SMARTtest emphasize features such as ARS, topoisomerase II binding sites, curved DNA, kinked DNA, SATB1 recognition sequences, MAR/SAR recognition signatures, but also focus strongly on AT-richness of the putative S/MAR DNA (reviewed in Evans et al., 2007).

However, it has been clearly documented that GC-rich S/MARs also exist (Boulikas, 1993; Yusufzai and Felsenfeld, 2004). Indeed, Goetze et al. (2005) showed that an algorithm that relies largely on AT-richness may miss S/MARs that have AT-rich patches but are GC-rich overall. Linnemann et al. (2007), who mapped S/MARs on the short arm of human chromosome 16 using a high-resolution CGH array, found a class of GC-rich S/MARs located in introns of gene-rich regions. Yusufzai and Felsenfeld (2004) showed that CTCF (CCCTC-binding factor) is a nuclear matrix protein that binds to a GC-rich S/MAR with insulator activity. In agreement with these observations, our study revealed a class of Arabidopsis S/MARs (Cluster D) that is GC-rich (Figure 14) and remains bound to the nuclear matrix following LIS extraction.

It is interesting to ask if GC-rich S/MARs differ in matrix affinity or other properties from AT-rich S/MARs. All our S/MARs were originally identified from an analysis of high-resolution array data. We then carried out a series of exogenous binding assays to validate the endogenous binding data from the array. Validations were initially done with a set of
S/MARs characterized by low FDRs (0 to 0.028) in the array analysis. These S/MARs had AT-contents ranging between 58 to 80%. We also included the strong-binding tobacco Rb7 S/MAR (Hall et al., 1991) as a positive control. A second set of validation assays focused on a set of S/MARs with higher FDRs (0.039 to 0.05), whose AT-contents ranged from 43 to 71%. Most of the S/MARs tested clearly showed specific binding in the exogenous assay (Figure 8). However, we were unable to detect in vitro binding activity for 4 S/MARs (out of 24 tested) (Figure 8). One of these S/MARs was from cluster E and had an AT-content of 66%, whereas the remaining three were from cluster D and had AT-contents of 43, 57 and 60% (Table 6). The cluster D S/MARs are GC-rich and located within exonic regions rich in H3K56ac (Figure 22), suggesting that they associate with the nuclear matrix by a different mechanism than the AT-rich S/MARs.

A closer examination of S/MAR sequences to identify runs of A and T residues (AT-rich kmers), which could be considered as potential regions for base unpairing and matrix binding (Bode et al., 1992), revealed the expected high AT-kmer content in AT-rich S/MARs (Figure 15), and a low AT-kmer content in GC-rich S/MARs (Figure 15). Importantly, a significant fraction (38%) of the GC-rich S/MARs in cluster D had no AT-kmers ≥ 10 nt in length, suggesting that GC-rich S/MARs might attach by a mechanism that does not involve DNA unwinding elements. These S/MARs are located within transcribed regions, so the possibility should be considered that they bind to the matrix as a byproduct of the transcription process, rather than by direct interaction.

S/MARs from cluster D clearly bind to the nuclear matrix in the endogenous assay but their binding does not appear to be sufficiently robust for allow their validation in our
exogenous binding assay. However, it must be noted that the validation assays used nuclear matrices isolated from tobacco suspension cells, whereas our microarray results measured endogenous binding to *Arabidopsis* matrices from our suspension cell nuclei. Thus we cannot exclude the possibility that there are species-specific differences between the nuclear matrix from *Arabidopsis* and tobacco that preferentially affected binding of the GC-rich S/MARs in cluster D. This possibility is supported by a wealth of results showing that the binding activities of some S/MARs and the presence of certain proteins in the matrix can vary depending upon tissue type, environmental stimuli, or cell cycle stage (Boulikas, 1993; Ottaviani et al., 2008; Levy-Wilson and Fortier, 1989; Brotherton et al., 1991; Dickinson et al., 1992; Bidwell et al., 1993; Cai et al., 2003).

The tissue-specific expression of putative nuclear matrix proteins from plants is well documented (Chua et al., 2003; Ng et al., 2009; Gallavotti et al., 2011). The *Arabidopsis* flowering pathway is a model system for studying tissue-specific, epigenetic regulation in plants (Ng et al., 2009). *AGAMOUS*, a homeotic protein that regulates reproductive organ identity, controls the expression of *GIANT KILLER (GIK)*. *GIK* is a tissue-specific AT-hook protein that causes a rapid change in histone modification. Ng et al. (2009) found that *GIK* is a nuclear matrix protein that binds to a S/MAR located in the promoter of *ETTIN/AUXIN RESPONSE FACTOR 3 (ETT/ARF3)*, resulting in the control of gynoecium patterning. In this case, GIK binding to the promoter S/MAR of *ETT/ARF3* leads to repression. Ng et al. (2009) suggested that *GIK* might serve as a center for chromatin remodeling that occurs on the nuclear matrix for controlling target gene expression. While S/MARs are associated
frequently with increased gene expression, the GIK example shows that matrix binding of a S/MAR can also act to repress gene expression in *Arabidopsis*.

We also cannot rule out the possibility that GC-rich *Arabidopsis* S/MARs are artifacts that co-precipitate with the *Arabidopsis* nuclear matrices, and are not necessarily bound to the nuclear matrix before the S/MAR isolation. We are currently performing additional exogenous binding assays for GC-rich *Arabidopsis* S/MARs using *Arabidopsis* nuclear matrices to explore these possibilities.

**In silico vs. Experimental S/MARs**

Reliable computer prediction of S/MARs would contribute to our understanding of higher-order chromatin structure. We observed that about 63% of our experimental S/MARs matched the S/MARs predicted by SMARTest algorithm (Rudd et al., 2004)(Figure 16) and AT-richness was clearly a primary predictive factor. Accordingly, the AT-rich S/MARs associated with TEs (Table 8) showed the highest number of matches.

It is not surprising that the overlap between our experimental S/MAR population and those predicted by the SMARTest algorithm (Rudd et al., 2004) is comprised predominantly of AT-rich sequences. The original S/MAR library used as a training set for SMARTest was derived from a population of AT-rich S/MARs (Frisch et al., 2002). Thus, other experimental S/MARs divergent from AT-rich S/MARs are less likely to be detected by the algorithm. However, some S/MAR interactions with the nuclear matrix may change depending on the cell type and gene expression status, and might not be represented among our experimentally identified S/MARs.
S/MARs and AT-rich Transposon Element Superfamilies co-map

Previous studies have shown that some transposable elements (TEs) frequently co-map with S/MAR elements (Galliano et al., 1995; Avramova et al., 1998; Nabirochkin et al., 1998; Rollini et al., 1999; Tikhonov et al., 2000). We found that many of the chr4 S/MARs co-mapped with TE superfamilies (TE-SFs) (Table 9), which is exemplified most clearly by cluster B (Figure 13). However, not all S/MARs co-map with TE-SFs, and not all TE-SFs co-map with S/MARs. We found that S/MAR enrichment in TE-SFs is strongly correlated with AT-content (Figure 17). TE-SFs that were significantly enriched for S/MAR overlapping sequences include DNA, DNA/Harbinger, DNA/Mariner, DNA/MuDR, and RC/Helitron (Table 9). Conversely, the LTR/Copia retrotransposon superfamily was most depleted in S/MAR sequences (Table 9). MITEs have been found to be enriched in S/MARs in rice, sorghum, and maize (Avramova et al., 1998; Tikhonov et al., 2000), and it has also been reported that MITEs occur in the Arabidopsis genome (Santiago et al., 2002). However, we were unable to assess any possible association of MITEs with Arabidopsis S/MARs because MITEs are not annotated in the Arabidopsis TAIR10 database. In Drosophila, the gypsy transposon acts as a S/MAR element and has insulator activity (Nabirochkin et al., 1998; Byrd and Corces, 2003). However, our study shows while that gypsy superfamily TE sequences sometimes occur in Arabidopsis S/MARs, they are not enriched in that fraction (Table 9). Fragile sites from humans and mice have high AT-content and are enriched in both S/MARs and several transposon families, including DNA/mariner (Shiraishi et al., 2001). We found that the DNA/Mariner superfamily sequences are significantly enriched in Arabidopsis S/MARs (Table 9). However, the AT-richness of a transposon superfamily closely correlates
with the probability of S/MAR activity within that family (Figure 17), and it seems likely that matrix binding is more related to AT composition than to transposition potential. However, it is reasonable to suppose that AT-rich transposons may contribute to genome organization patterns by re-distributing matrix binding regions in different evolutionary lineages (Bennetzen, 2000; Tikhonov et al., 2001).

**Nucleosome Occupancy**

Chromatin structure regulates critical processes such as transcription and replication by controlling access to the chromatin. Accessibility is controlled primarily by the compaction of the nucleosomes. Less compact or nucleosome-free structures are typically found in genome regions that bind to regulatory molecules. Experimentally, the accessibility of a chromatin region can be measured by its accessibility to DNase I or micrococcal nuclease.

We have used an algorithm for predicting nucleosome density (Kaplan et al., 2009) to estimate the likelihood of nucleosomal occupancy along *Arabidopsis* chr4 for comparison with the distribution of our experimental S/MARs. The algorithm is primarily controlled by AT-content, and AT-rich regions are less likely to contain nucleosomes than regions with high GC-content as has been shown experimentally by many studies (Yuan et al., 2005b; Valouev et al., 2011).

Not surprisingly, the vast majority of the *Arabidopsis* S/MARs on chr4, including S/MARs in clusters A, B, C, and E, which have higher AT content than the analyzed chromosomal regions, are less likely to be occupied by nucleosomes when compared to...
flanking regions (Figure 18). However, cluster D S/MARs, which are GC-rich and found primarily in exonic regions, have greater predicted nucleosome occupancy (Figure 18). *Arabidopsis* genes tend to be more GC-rich than either intergenic or intronic regions (AGI, 2000), which is likely to explain this result. As noted above, it is possible that the matrix binding within a genic-exonic region is associated with a component of the transcription apparatus. Supporting this idea, previous studies showed that a hyperphosphorylated form of RNA polymerase II is part of the nuclear matrix of animal nuclei (Bisotto et al., 1995; Mortillaro et al., 1996) along with a variety of transcription factors (reviewed in Albrethsen, et al., 2009). However, Calikowski, et al. found no evidence of RNA polymerase II in the matrix fraction from *Arabidopsis* cells (Calikowski et al., 2003). LIS-purified scaffolds are thought to preserve replication complexes and disrupt transcriptional complexes (Bode et al., 1995), whereas 2 M NaCl disrupts replication complexes (Ma et al., 1999). However, we used a low concentration (10 mM) of LIS in our experiments. Because Calikowski et al. (2003) used a significantly higher concentration (25 mM), it is possible that their procedures removed some proteins that remain in our matrix preparations.

**S/MARs and Epigenetic Marks**

We previously mapped epigenetic features on *Arabidopsis* chromosome predominantly associated with different chromatin structures in *Arabidopsis* suspension cell line (Tanurdzic et al., 2008; Lee et al., 2010). In all of the matrix isolation experiments reported here, cells were grown under conditions identical to those previously described,
allowing us to make direct comparisons of our S/MAR and epigenetic data (Tanurdzic et al., 2008; Lee et al., 2010).

We found that the general pattern of histone modifications was very similar for all of the S/MAR clusters but there were some clear differences. For example, all five clusters showed a depletion of H3K4me1/2 relative to the overall average, although the levels of depletion varied (Figure 21, and Table 10). Very little is known about H3K4me1/2, although Zhang et al. (2009) have reported that H3K4me1/2 in Arabidopsis is associated with longer genes that are expressed at very low levels in highly tissue-specific manner.

All of the S/MAR clusters have levels of H3K9me2 very similar to the levels in flanking regions, with the exception of cluster B, which is enriched in H3K9me2, AT-rich and includes the majority of the TE-associated S/MARs. Our data are consistent with previous observations showing that H3K9me2 is found on heterochromatic TEs and is associated with small interfering RNAs (Fransz et al., 2003; Tanurdzic et al., 2008; Autran et al., 2011). Tanurdzic et al. (2008) found that H3K9me2 levels are greatly depleted in our cell line when compared to Arabidopsis leaves. The loss of H3K9me2 in our cell line is also accompanied by the reactivation of several TEs.

Most of the S/MAR clusters showed a decreased level in 5mC DNA. The one exception was cluster B S/MARS, which showed no change relative to flanking regions (Figure 19), even though the heterochromatic TEs in our cell line are hypomethylated when compared to Arabidopsis leaf DNA (Tanurdzic et al., 2008)

As noted previously, cluster D S/MARs are different from the S/MARs in other clusters. Cluster D S/MARs are characterized by higher levels of H3K56ac and potential
nucleosome occupancy. S/MAR clusters C and E also have elevated levels of H3K56ac but a reduced likelihood of nucleosome occupancy. H3K56ac has long been known to be a euchromatic mark (Roudier et al., 2011) and our previous studies of the epigenetic marks in the Arabidopsis cell line lead to a similar conclusion. Tanurdzic et al. (2008) found that H3K56ac was preferentially associated with promoter regions, and our recent study of DNA replication timing found H3K56ac enrichment in early replicons and within the initiation zones of both early and late replicons (Lee et al., 2010). In contrast to S/MAR clusters C, D and E, H3K56ac levels remain constant in S/MAR clusters A and B, which include S/MARs in unannotated regions and the TE-associated S/MARs.

Cairns (2009) refers to nucleosomes containing H3K56ac as “hot nucleosomes”, because the acetylation is thought to attract the ATP-dependent chromatin remodelers of the SWI/SNF class and make the nucleosome unstable and subject to ejection (Rufiange et al., 2007a). Interestingly, SWI/SNF has been reported to be associated with nuclear matrix proteins (Reyes et al., 1997; Euskirchen et al., 2011). According to this hypothesis, the nucleosomes within the S/MARs of clusters C, D and E would be subject to SWI/SNF remodeling and readily ejected (Henikoff, 2008). Such unstable nucleosomes may be required to initiate processes such as transcription and replication.

S/MARs and Gene Expression

Several studies have shown that S/MARs are sometimes found in genes, near or within enhancer or promoter regions (Gasser and Laemmli, 1986b; Jenuwein et al., 1997; Linnemann et al., 2009; Keaton et al., 2011). Studies of three developmentally regulated
genes in *Drosophila*, which include Sgs4, *fushi tarazu* and *Adh*, showed that S/MARs near the 5’end of these genes are required for enhanced expression (Gasser and Laemmli, 1986a). Linnemann et al. (2009) showed that S/MARs are frequently located near the 5’ region of transcribed genes on human chromosomes 14 to 18. In a more recent study, Keaton et al. (2011) compared S/MARs within the 30-Mb region of the human genome being studied by the ENCODE collaborators. In agreement with the S/MAR results from Linnemann et al. (2009), their data located the ENCODE S/MARs near expressed genes and in close proximity to transcription start sites (TSS). Keaton et al. (2011) also found that ENCODE S/MARs are preferentially associated with RNA polymerase II binding regions. In *Arabidopsis* chr4, we found that S/MARs are often associated with the 5’ ends of genes (Figure 10B and 10C, and Figure 23A), suggesting that they may influence promoter function.

We tested whether the chr4 S/MARs are located near TSS using two complementary analyses. Our data revealed that genes with a S/MAR located at -60 bp to +251 bp, relative to the TSS, were more likely to be expressed (Figure 24A and Figure 26). Most of the active genes in this category were associated with cluster C S/MARs, along with small fractions associated with S/MARs from clusters D or E (Figure 24A). We also showed that genes with a cluster C S/MAR associated with the TSS are more likely to be expressed than the average gene (Figure 25A, Figure 26, and Table 11). Our data are consistent with previous studies in other higher eukaryotes showing that S/MARs located near or within genes seem to play a transcriptional regulatory role (Anthony and Blaxter, 2007; Keaton et al., 2011), whereas S/MARs located in intergenic regions are more likely to define the matrix binding for loop domains (Spiker and Thompson, 1996; Ishii et al., 2002). However, we cannot rule out that
some of these S/MARs might be associated with intergenic transcription, as shown previously by several groups (Tanurdzic et al., 2008; Zheng et al., 2009; Kim et al., 2011). Although not all of the genes with a TSS-associated S/MAR are expressed in our suspension culture cells, it is likely that other factors dictate cell-specific expression patterns, and it is reasonable to suppose that S/MARs associated with an inactive gene may facilitate its transcriptional activation in response to developmental or environmental cues.

We found that chr4 genes associated with TSS S/MARs were significantly enriched for transcription factors (as annotated in the AGRIS database; Table 12), consistent with a prediction by Tetko et al. (2006) using in silico prediction of S/MARs in Arabidopsis, who anticipated transcription factors genes to contain a significant proportion of S/MARs. Van Drunen et al. (1997) identified three S/MARs upstream of the transcription factor ATH1 (At4g32980), one of which appeared to be in the 5’ proximal region (van Drunen et al., 1997). Our results show that the 5’ proximal ATH1 S/MAR is in cluster D and primarily exonic. ATH1 is an important light-regulated homeobox protein (Quaedvlieg et al., 1995). It is also interesting that while ATH1 is “light-regulated”, the mRNA is absent from our cell culture even though it is grown in continuous light. In plants, ATH1 expression is both light-responsive and dependent on AGAMOUS expression, which also not active in the cultured cells. Perhaps in the absence of AGAMOUS, the ATH1 S/MAR remains bound to the nuclear matrix and transcription is repressed.

Our data indicate that S/MARs may control gene expression by binding to the nuclear matrix and future studies are warranted to further understand a mechanism that is poorly understood.
Figure 1. K-means clustering of S/MAR data.
Putative S/MARs identified from the array data were arbitrarily grouped into between 2 and 25 clusters (x axis). The Y axis shows residual error (sum of squares) associated with each cluster. Five clusters partitioned the S/MAR data into groups that were biological relevant, with only incremental reductions in error thereafter.
Figure 2. Determination of S/MAR position relative to the TSS.
S/MAR (red segment) may influence the expression of multiple genes (green arrows) in this region. The distance (bp) from the S/MAR midpoint to the TSS of genes is shown. The +/- in the distance indicate if the S/MAR is positioned either upstream (-) or downstream (+) of the respective TSS. Black vertical lines indicate the TSS of genes. Red vertical line indicates midpoint of S/MAR.
Figure 3. Histones are removed from Arabidopsis nuclei with 10 mM LIS. A Coomassie-stained SDS/polyacrylamide gel analysis of proteins remaining after extracting nuclei with the LIS-containing buffer HIB2. LIS concentration (mM) is indicated above the respective lanes. From left to right: **Lane 1**, purified wheat histones; **Lane 2**, total nuclear proteins; **Lanes 3** to **7**, nuclear proteins after treatment with increasing LIS concentrations; **Lane 8**, purified wheat histones; **Lane 9**, molecular mass markers. Locations of wheat and *Arabidopsis* histones are indicated on the left, and masses (kDa) of the markers are shown on the right.
Figure 4. Amplification scheme for the one-tube, single-reaction nested PCR method. (A-C) The blue boxes indicate specific primers that flank S/MAR or non-S/MAR regions. The undulating gray box represents the universal M13(-21) sequence, and the red star the fluorescent IRD-700 label. (D) In the first PCR cycles, the forward primer with the M13(-21) tail is incorporated into the PCR products. (E) These products are then the target for the IRD-700-labeled universal M13(-21) primer, which is incorporated during subsequent cycles at lower annealing temperature of 50°C. (F) The final labeled product can be analyzed on a laser detection system.
Figure 5. Quantification of matrix-attached (S/MAR) DNA and total genomic DNA (input).
Nuclear haloes were digested with EcoR1 and HindIII, followed by low speed centrifugation. Total genomic DNA (DNA isolated from undigested nuclei) was used as reference. The DNA (ng) from total genomic DNA (red) and the matrix bound DNA (green) is shown on the Y-axis. Approximately 27% of the digested “halo” DNA remains bound to the nuclear matrix, representing an approximately 4-fold enrichment for matrix-associated DNA.
Figure 6. Distribution of S/MARs on *Arabidopsis* chromosome 4. (A) The gene-rich euchromatic distal short and distal long arms are shaded light blue while the heterochromatic knob and pericentromere rich in TEs are shaded black. The proximal portions of both the short and long arms have intermediate characteristics and are shaded dark blue. The regions shown do not include the repetitive telomere nor the nucleolar organizer regions. (B) AT content calculated in 1-kb non-overlapping windows. (C) Gene (green) and TE (yellow) density per 1 Mb. (D) S/MARs (red) per Mb.
Figure 7. Schematic diagram showing the steps of the exogenous binding assay. The binding of a particular DNA fragment was tested by incubation of end-labeled exogenous fragment with nuclear matrices. See text for additional details.
Figure 8. Chromosomal location of microarray-identified S/MARs and verification by exogenous binding assays.

(A) Locations (a to x) of the selected 24 putative S/MARs on Arabidopsis chr4. (B) Binding assay results for each of the 24 putative S/MARs (a to x). NT1 tobacco matrices were used for each in vitro binding assay. T=total input labeled DNA, P=pellet (bound fraction), S=supernatant or unbound fraction. Non-S/MAR serves as internal negative control. The asterisk (*) indicates putative S/MARs that failed to bind.
Figure 9. Matrix binding competition of S/MAR fragments and increasing amounts of sonicated *E. coli* DNA.
The total assay volume was 50 µL and contained ~6 x 10^4 NT-1 nuclear matrices. The top panel (Pellet) shows the labeled S/MAR or non-S/MAR remaining bound in the presence of 0, 1, 2.5, 5, 10, or 20 µg of *E. coli* competitor DNA. The bottom panel (Supernatant) shows the labeled S/MAR or non-S/MAR that was unbound in the presence of 0, 1, 2.5, 5, 10, or 20 µg of *E. coli* competitor DNA. The non-S/MAR DNA served as an internal negative control. The total DNA stained with ethidium bromide (EtBr) is shown below each binding assay for both the pellet and the supernatant fractions.
Figure 10. Composite map of two 50-kb regions.
(A) Schematic representation of *Arabidopsis* chr4. (B) Heterochromatin region. (C) Euchromatin region containing the *Arabidopsis* ATH1 locus. Tracks: NuOc=nucleosome occupancy prediction. pS/MARs= *in silico* predicted S/MARs. eS/MARs= experimental S/MARs. Array= microarray profile. Dashed lines indicate 5’ end of genes. (D) Two exogenous binding assays for S/MARs located at the chromosomal region marked by the vertical line. T= total input labeled DNA, P= pellet (bound fraction), S= supernatant (unbound fraction). The non-S/MAR serves as an internal negative control.
Figure 11. NimbleGen microarray profile of S/MARs for the plastocyanin locus. Positive S/MARs probes are above 0.0 whereas anti S/MAR probes are below the 0.0 log$_2$ ratio (black lines). Genes (green arrows) and S/MARs (purple bars) from previously identified S/MARs mapped at the plastocyanin locus by van Drunen, et al. (1997). The arrows indicate the direction of transcription for the respective genes.
Figure 12. Association of S/MARs with various genomic features.
S/MARs were clustered based on the degree of association (nucleotide overlap) with genomic features annotated in the *Arabidopsis* TAIR10 database, including genes, exons and TEs. S/MAR associations with introns and unannotated regions were calculated after the clustering.
Figure 13. Partition of experimental S/MARs into five clusters according to location on chromosome 4.
K-means clustering was used to group the 1,358 experimental S/MARs according to their degree of association with genes, exons, and TEs. Introns, intergenic (unannotated) regions, and the AT content, which is typically higher in S/MARs, were determined after the cluster analysis. Red indicates the features with which the given S/MAR has the highest association, whereas blue indicates features least associated with that particular S/MAR. The number of S/MARs within the respective cluster is shown to the right.
Figure 14. AT content (percent) varies for S/MARs within a cluster. The dashed line indicates the mean AT content for the 15.41 Mb of chr4 in which MARs were analyzed.
Figure 15. AT-rich k-mer content for S/MARs clusters A to E.
The quantitative measure of AT-rich k-mer (k=10) content was calculated by dividing the number of As or Ts in AT-rich k-mers for each S/MAR by the total length of the S/MAR. The AT-rich k-mer content was also measured in 1,400 randomly selected EcoRI/HindIII fragments and 1,497 anti-S/MARs serving as controls. Segment inside box plot indicates the median AT-rich k-mer content. The X axis shows the cluster (A to E), along with the control anti-S/MARs or Eco R1/HindIII fragments.
Figure 16. Experimental and predicted S/MARs show some overlap.
The 1358 experimental S/MARs (eS/MARs) compared with the 2892 in silico predicted S/MARs (pS/MARs) of Rudd et al. (2004) showed an overlap of 860 eS/MARs and pS/MARs.
Figure 17. Transposon element superfamily (TE-SF) enrichment in S/MARs is strongly correlated with the AT content. The TE-SF enrichment in S/MAR sequences is shown on the Y-axis whereas the TE-SF AT content (%) is shown on the X-axis. The enrichment for each TE-SF in S/MARs is ranked as significant enrichment (•) in proportion to its abundance in chr4 (●) or depleted (○). The line shows positive correlation.
Figure 18. Predicted (probability) of nucleosome occupancy for each S/MAR cluster. The experimental S/MARs for each cluster were aligned at their midpoint and analyzed 5 kb upstream and 5 kb downstream from the S/MAR midpoints. This 10-kb region was segmented into twenty 500-bp non-overlapping windows (scale bars in X axis). The two center windows overlap the S/MAR. The dashed line indicates the mean predicted nucleosomal occupancy for the corresponding chromosomal regions.
Figure 19. DNA cytosine methylation (DNAmC) for each S/MAR cluster.
The experimental S/MARs for each cluster were aligned at their midpoint and analyzed 5 kb upstream and 5 kb downstream from the S/MAR midpoints. This 10-kb region was segmented into twenty 500-bp non-overlapping windows (scale bars in x axis). The two center windows overlap the S/MAR. The dashed line indicates the mean number of enriched probes for the corresponding chromosomal regions. Fraction modified = fraction of cytosines methylated.
Figure 20. S/MARs and H3K9me2.
The experimental S/MARs for each cluster were aligned at their midpoint and analyzed 5 kb upstream and 5 kb downstream from the S/MAR midpoints. This 10-kb region was segmented into twenty 500 bp non-overlapping windows (scale bars in x axis). The two center windows overlap the S/MAR. The dashed line indicates the mean number of enriched probes for the corresponding chromosomal regions. Fraction modified= fraction of H3K9me2 modified.
The experimental S/MARs for each cluster were aligned at their midpoint and analyzed 5 kb upstream and 5 kb downstream from the S/MAR midpoints. This 10-kb region was segmented into twenty 500 bp non-overlapping windows (scale bars in x axis). The two center windows overlap the S/MAR. The dashed line indicates the mean number of enriched probes for the corresponding chromosomal regions. Fraction modified= fraction of H3K4me1/2 modified.

**Figure 21. S/MARs and H3K4me1/2.**
Figure 22. S/MARs and H3K56ac.
The experimental S/MARs for each cluster were aligned at their midpoint and analyzed 5 kb upstream and 5 kb downstream from the S/MAR midpoints. This 10-kb region was segmented into twenty 500 bp non-overlapping windows (scale bars in x axis). The two center windows overlap the S/MAR. The dashed line indicates the mean number of enriched probes for the corresponding chromosomal regions. Fraction modified= fraction of H3K56ac modified.
Figure 23. S/MARs are associated with TSSs of Genes. For each S/MAR, the minimal distance to the closest gene Transcriptional Start Site (TSS) (A), or Transcriptional Termination Site (TTS) (B) was calculated (see Figure 2), and grouped into data points.
Figure 24. S/MARs are associated with the TSS of active Genes.
Genes were grouped based on their S/MAR to Transcription Start Site (A), or Transcription Termination Site (B) distance, with negative and positive values indicating upstream or downstream respectively, and were then binned within a distance that ranged -1000 to +3000 into quintiles. Gene activity for each bin was then calculated as the fraction of genes in each bin that are expressed. The error bars indicate 95% confident intervals. Dashed line shows the average expression of gene regions. If the error bar does not overlap dashed line, then the p <0.05. Colors within bins show the proportion of each S/MAR cluster found in the active genes.
Figure 25. Genes with TSSs-associated cluster C S/MARs are more likely to be expressed than the average gene.

The gene most proximal to a S/MAR was identified based on the S/MAR to gene Transcription Start Site (A), or Transcription Termination Site (B) distance (see Figure 2), arriving at a set of 1216 genes for the 1358 MARs. The set does not contain 1358 genes because some genes are the most proximal neighbor to 2 or more S/MARs. Genes were then grouped by the proximal S/MAR cluster and analyzed for gene activity as the fraction of genes in each bin that are expressed. The error bars indicate 95% confident intervals. Dashed line shows the average expression of gene regions. If the error bar does not overlap dashed line, then the p <0.05.
Figure 26. A simplified model depicting the association of S/MARs with gene expression.
A gene with a S/MAR located at -60 bp to +251 bp (red rectangle), relative to the TSS, is more likely to be expressed than any gene in the loop domain. It has proposed that the transcription machinery is assembled at the site of S/MARs. Interaction of S/MARs with the nuclear matrix brings certain coding sequences near the transcription machinery, thus enabling specific genes to be regulated. Most of the active genes in this category were associated with cluster C, D, and E S/MARs (red rectangle), which are enriched in H3K56ac, an epigenetic mark associated with open chromatin conformation. Cluster A and B S/MARs (purple rectangle) have no effect on gene transcription.
Table 2. Primer sequences for S/MAR regions.

<table>
<thead>
<tr>
<th>ID_letter</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>5' gttgctgacgaagagcatc 3'</td>
<td>5' gaatctcggtagttcaagtc 3'</td>
</tr>
<tr>
<td>b</td>
<td>5' cccatgatgatcgggtac 3'</td>
<td>5' caattgcggcttctctac 3'</td>
</tr>
<tr>
<td>c</td>
<td>5' tcaacaaagatttgcttc 3'</td>
<td>5' gctcaagtaagctctct 3'</td>
</tr>
<tr>
<td>d</td>
<td>5' gactttaagaccacaaag 3'</td>
<td>5' gcaactgatacaaccaagatc 3'</td>
</tr>
<tr>
<td>e</td>
<td>5' caaacaccgacacctgatga 3'</td>
<td>5' gtcctcaatttctgcctaa 3'</td>
</tr>
<tr>
<td>f</td>
<td>5' ttgtaatgcgtctgacgtgc 3'</td>
<td>5' ccttcttggtctcaacgt 3'</td>
</tr>
<tr>
<td>g</td>
<td>5' gggttgagaacgatatagact 3'</td>
<td>5' cgagaagaaaaacgaggtg 3'</td>
</tr>
<tr>
<td>h</td>
<td>5'atatccagacacacctttaaag 3'</td>
<td>5' tagggtaacttaggcaaaag 3'</td>
</tr>
<tr>
<td>i</td>
<td>5' ttgacaagacactataatgggtg 3'</td>
<td>5' ctctctcaagattctcct 3'</td>
</tr>
<tr>
<td>j</td>
<td>5' tcagatcagtctccagcactc 3'</td>
<td>5' gtctccctagtgttagcagc 3'</td>
</tr>
<tr>
<td>k</td>
<td>5' caagttggaattctcgct 3'</td>
<td>5' cttggagaagctgcacattc 3'</td>
</tr>
<tr>
<td>l</td>
<td>5' tgaaggagtagtaaggtg 3'</td>
<td>5' agatggataatttgaaacc 3'</td>
</tr>
<tr>
<td>m</td>
<td>5' gtaccaaccaacactacaacag 3'</td>
<td>5' tgaagatgtatatgagatg 3'</td>
</tr>
<tr>
<td>n</td>
<td>5' ccaactataagtttgagat 3'</td>
<td>5' aagatgagaatctttcgt 3'</td>
</tr>
<tr>
<td>o</td>
<td>5' gctctctcacaaggtc 3'</td>
<td>5' caagggctctctcttcct 3'</td>
</tr>
<tr>
<td>p</td>
<td>5' gatccatactacttggaa 3'</td>
<td>5' ctacatcacaactaacctat 3'</td>
</tr>
<tr>
<td>q</td>
<td>5' tcagcagcatttgcggagc 3'</td>
<td>5' ggtctcatcaagcggtttatac 3'</td>
</tr>
<tr>
<td>r</td>
<td>5' acttaatttacaaacaggtt 3'</td>
<td>5' ctcaaccagttcatatctaa 3'</td>
</tr>
<tr>
<td>s</td>
<td>5' aactcacaacctcagcttgac 3'</td>
<td>5' tagaccactcacaactac 3'</td>
</tr>
<tr>
<td>t</td>
<td>5' ctggcctttatctcttc 3'</td>
<td>5' gctctcttctgttgaac 3'</td>
</tr>
<tr>
<td>u</td>
<td>5' caagcttagtcatatcagc 3'</td>
<td>5' atgtctcatcatactctacagc 3'</td>
</tr>
<tr>
<td>v</td>
<td>5' ccacaaacgagacgccaacta 3'</td>
<td>5' agttgccttagataattgaaaag 3'</td>
</tr>
<tr>
<td>w</td>
<td>5' tgtcttcgcaactaaggctc 3'</td>
<td>5' gttgaactctctctgatgtc 3'</td>
</tr>
<tr>
<td>x</td>
<td>5' ttctagactttctgtctct 3'</td>
<td>5' ttggactttttctcctccag 3'</td>
</tr>
</tbody>
</table>

ID letter according to Figure 8A, and 8B.
Table 3. Primer sequences for non-S/MAR regions.

<table>
<thead>
<tr>
<th>ID_letter</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>b</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>c</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>d</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>e</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>f</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>g</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>h</td>
<td>5'aggagggaaacaactccaaaca 3'</td>
<td>5'aacaattgtgtgtgcttca 3'</td>
</tr>
<tr>
<td>i</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>j</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>k</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>l</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>m</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>n</td>
<td>5'cacaacaacaagactggagatt 3'</td>
<td>5'tcgtgccacacccgcaagat 3'</td>
</tr>
<tr>
<td>o</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>p</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>q</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>r</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>s</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>t</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>u</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
<tr>
<td>v</td>
<td>5'tcagcccagttgtaactac 3'</td>
<td>5'agggttaaacatctccccaa 3'</td>
</tr>
<tr>
<td>w</td>
<td>5'tcagcccagttgtaactac 3'</td>
<td>5'agggttaaacatctccccaa 3'</td>
</tr>
<tr>
<td>x</td>
<td>5'agcaaaacggacttgaccttgctg 3'</td>
<td>5'agccacggataagatgcccacaaga 3'</td>
</tr>
</tbody>
</table>

ID letter according to Figure 8A, and 8B.
Table 4. Results for K-means clustering of S/MARs.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Exon</th>
<th>TE</th>
<th>Count</th>
<th>WithinSS</th>
<th>WithinSS.Count</th>
<th>Intron</th>
<th>MNO</th>
<th>AT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.93</td>
<td>7.38</td>
<td>424</td>
<td>105326.5</td>
<td>248.41</td>
<td>0.3</td>
<td>0.22</td>
<td>69.99</td>
</tr>
<tr>
<td>B</td>
<td>4.76</td>
<td>80.54</td>
<td>276</td>
<td>160962.9</td>
<td>583.19</td>
<td>0.98</td>
<td>0.17</td>
<td>72.9</td>
</tr>
<tr>
<td>C</td>
<td>36.82</td>
<td>7.11</td>
<td>285</td>
<td>158016</td>
<td>554.44</td>
<td>10.83</td>
<td>0.32</td>
<td>65.88</td>
</tr>
<tr>
<td>D</td>
<td>82.2</td>
<td>4.21</td>
<td>231</td>
<td>129215.4</td>
<td>559.37</td>
<td>7.73</td>
<td>0.47</td>
<td>58.68</td>
</tr>
<tr>
<td>E</td>
<td>28.84</td>
<td>14.21</td>
<td>142</td>
<td>162702.9</td>
<td>1145.79</td>
<td>64.93</td>
<td>0.29</td>
<td>67.11</td>
</tr>
</tbody>
</table>

Values in Exon, TE, and Intron columns indicate the percentage of overlap of each S/MAR cluster with these annotated genomic features.
Table 5. Distribution of S/MARs along chr4.

<table>
<thead>
<tr>
<th>REGIONS</th>
<th>low_coord</th>
<th>high_coord</th>
<th>length (bp)</th>
<th>AT (%)</th>
<th>S/MAR_count</th>
<th>Gene_density</th>
<th>TE_density</th>
<th>S/MAR_density</th>
</tr>
</thead>
<tbody>
<tr>
<td>distal short arm</td>
<td>1</td>
<td>1592652</td>
<td>1592652</td>
<td>63.85</td>
<td>141</td>
<td>284</td>
<td>254</td>
<td>89</td>
</tr>
<tr>
<td>Knob</td>
<td>1592653</td>
<td>2310327</td>
<td>717675</td>
<td>63.01</td>
<td>NA</td>
<td>238</td>
<td>499</td>
<td>NA</td>
</tr>
<tr>
<td>proximal short arm</td>
<td>2310328</td>
<td>2811478</td>
<td>501151</td>
<td>65.55</td>
<td>42</td>
<td>247</td>
<td>519</td>
<td>84</td>
</tr>
<tr>
<td>Pericentromere</td>
<td>2811479</td>
<td>5266603</td>
<td>2455125</td>
<td>61.68</td>
<td>NA</td>
<td>205</td>
<td>559</td>
<td>NA</td>
</tr>
<tr>
<td>proximal long arm</td>
<td>5266604</td>
<td>9200060</td>
<td>3933457</td>
<td>65.35</td>
<td>325</td>
<td>259</td>
<td>396</td>
<td>83</td>
</tr>
<tr>
<td>distal long arm</td>
<td>9200061</td>
<td>18585056</td>
<td>9384996</td>
<td>63.66</td>
<td>850</td>
<td>293</td>
<td>108</td>
<td>91</td>
</tr>
<tr>
<td>All Chr4</td>
<td>1</td>
<td>18585056</td>
<td>18585056</td>
<td>63.8</td>
<td>NA</td>
<td>270</td>
<td>267</td>
<td>NA</td>
</tr>
<tr>
<td>Chr4 -Het</td>
<td>NA</td>
<td>NA</td>
<td>15412256</td>
<td>64.17</td>
<td>1358</td>
<td>282</td>
<td>210</td>
<td>88</td>
</tr>
</tbody>
</table>

Chr4 –Het (chromosome 4 minus heterochromatin) refers to the four regions of chr4 analyzed in this study including distal short arm, proximal short arm, proximal long arm, and distal long arm. The term “density” is defined as the number of counts per 1 Mb. NA=Not Available
Table 6. Summary of the 24 selected putative S/MARs for exogenous binding assays.

<table>
<thead>
<tr>
<th>ID letter</th>
<th>low_coord</th>
<th>high_coord</th>
<th>Region</th>
<th>Binding Activity</th>
<th>AT (%)</th>
<th>Maximum FDR</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1402690</td>
<td>1403358</td>
<td>distal_short</td>
<td>1</td>
<td>62.6</td>
<td>0.046</td>
<td>A</td>
</tr>
<tr>
<td>b</td>
<td>2330028</td>
<td>2330957</td>
<td>proximal_short</td>
<td>1</td>
<td>69.67</td>
<td>0.045</td>
<td>A</td>
</tr>
<tr>
<td>c</td>
<td>2582325</td>
<td>2583363</td>
<td>proximal_short</td>
<td>1</td>
<td>60.44</td>
<td>0.042</td>
<td>A</td>
</tr>
<tr>
<td>d</td>
<td>5417054</td>
<td>5417687</td>
<td>proximal_long</td>
<td>1</td>
<td>68.61</td>
<td>0.049</td>
<td>C</td>
</tr>
<tr>
<td>e</td>
<td>6733925</td>
<td>6734668</td>
<td>proximal_long</td>
<td>1</td>
<td>65.86</td>
<td>0.044</td>
<td>C</td>
</tr>
<tr>
<td>f</td>
<td>7764581</td>
<td>7765483</td>
<td>proximal_long</td>
<td>1</td>
<td>71.44</td>
<td>0.05</td>
<td>C</td>
</tr>
<tr>
<td>g</td>
<td>8862492</td>
<td>8863245</td>
<td>proximal_long</td>
<td>0</td>
<td>60.47</td>
<td>0.049</td>
<td>D</td>
</tr>
<tr>
<td>h</td>
<td>9423805</td>
<td>9424472</td>
<td>distal_long</td>
<td>1</td>
<td>74.27</td>
<td>0.02</td>
<td>A</td>
</tr>
<tr>
<td>i</td>
<td>10384912</td>
<td>10386452</td>
<td>distal_long</td>
<td>1</td>
<td>71.51</td>
<td>0.019</td>
<td>E</td>
</tr>
<tr>
<td>j</td>
<td>10400248</td>
<td>10400614</td>
<td>distal_long</td>
<td>1</td>
<td>70.89</td>
<td>0.049</td>
<td>A</td>
</tr>
<tr>
<td>k</td>
<td>10459796</td>
<td>10460046</td>
<td>distal_long</td>
<td>1</td>
<td>70.96</td>
<td>0.005</td>
<td>A</td>
</tr>
<tr>
<td>l</td>
<td>10624741</td>
<td>10625438</td>
<td>distal_long</td>
<td>1</td>
<td>80.22</td>
<td>0.013</td>
<td>B</td>
</tr>
<tr>
<td>m</td>
<td>11756342</td>
<td>11757273</td>
<td>distal_long</td>
<td>1</td>
<td>66.2</td>
<td>0.02</td>
<td>B</td>
</tr>
<tr>
<td>n</td>
<td>13316540</td>
<td>13317132</td>
<td>distal_long</td>
<td>1</td>
<td>68.8</td>
<td>0.028</td>
<td>B</td>
</tr>
<tr>
<td>o</td>
<td>13680445</td>
<td>13680710</td>
<td>distal_long</td>
<td>0</td>
<td>43.23</td>
<td>0.049</td>
<td>D</td>
</tr>
<tr>
<td>p</td>
<td>14770905</td>
<td>14779654</td>
<td>distal_long</td>
<td>1</td>
<td>71.02</td>
<td>0.019</td>
<td>A</td>
</tr>
<tr>
<td>q</td>
<td>14990285</td>
<td>14990626</td>
<td>distal_long</td>
<td>0</td>
<td>57.89</td>
<td>0.049</td>
<td>D</td>
</tr>
<tr>
<td>r</td>
<td>15292245</td>
<td>15293078</td>
<td>distal_long</td>
<td>1</td>
<td>70.98</td>
<td>0.015</td>
<td>A</td>
</tr>
<tr>
<td>s</td>
<td>15705843</td>
<td>15705843</td>
<td>distal_long</td>
<td>1</td>
<td>70.11</td>
<td>0.009</td>
<td>A</td>
</tr>
<tr>
<td>t</td>
<td>15917323</td>
<td>15917279</td>
<td>distal_long</td>
<td>1</td>
<td>70.31</td>
<td>0.01</td>
<td>D</td>
</tr>
<tr>
<td>u</td>
<td>16566916</td>
<td>16560010</td>
<td>distal_long</td>
<td>1</td>
<td>56.29</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>v</td>
<td>17642749</td>
<td>17643318</td>
<td>distal_long</td>
<td>1</td>
<td>73.33</td>
<td>0.019</td>
<td>B</td>
</tr>
<tr>
<td>w</td>
<td>18184357</td>
<td>18185216</td>
<td>distal_long</td>
<td>1</td>
<td>70.73</td>
<td>0.005</td>
<td>A</td>
</tr>
<tr>
<td>x</td>
<td>18311832</td>
<td>18312288</td>
<td>distal_long</td>
<td>0</td>
<td>66.06</td>
<td>0.039</td>
<td>E</td>
</tr>
</tbody>
</table>

ID letter according to Figure 8A, and 8B. Binding activity, 1 = bound to nuclear matrix, 0 = failed to bind nuclear matrix.
Table 7. Non-S/MAR fragments.

<table>
<thead>
<tr>
<th>ID_letter</th>
<th>Non-S/MAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>At4g19120</td>
</tr>
<tr>
<td>b</td>
<td>At4g19120</td>
</tr>
<tr>
<td>c</td>
<td>At4g19120</td>
</tr>
<tr>
<td>d</td>
<td>At4g19120</td>
</tr>
<tr>
<td>e</td>
<td>At4g19120</td>
</tr>
<tr>
<td>f</td>
<td>At4g19120</td>
</tr>
<tr>
<td>g</td>
<td>At4g19120</td>
</tr>
<tr>
<td>h</td>
<td>At4g16760</td>
</tr>
<tr>
<td>i</td>
<td>At4g19120</td>
</tr>
<tr>
<td>j</td>
<td>At4g19120</td>
</tr>
<tr>
<td>k</td>
<td>At4g19120</td>
</tr>
<tr>
<td>l</td>
<td>At4g19490</td>
</tr>
<tr>
<td>m</td>
<td>At4g19120</td>
</tr>
<tr>
<td>n</td>
<td>At4g26330</td>
</tr>
<tr>
<td>o</td>
<td>At4g19120</td>
</tr>
<tr>
<td>p</td>
<td>At4g19120</td>
</tr>
<tr>
<td>q</td>
<td>At4g19120</td>
</tr>
<tr>
<td>r</td>
<td>At4g31540</td>
</tr>
<tr>
<td>s</td>
<td>At4g19120</td>
</tr>
<tr>
<td>t</td>
<td>At4g19120</td>
</tr>
<tr>
<td>u</td>
<td>At4g19120</td>
</tr>
<tr>
<td>v</td>
<td>Intergenic region</td>
</tr>
<tr>
<td>w</td>
<td>At4g39030</td>
</tr>
<tr>
<td>x</td>
<td>At4g19120</td>
</tr>
</tbody>
</table>

Seven non-S/MARs were selected from regions that did not bind in the microarray experiments. Exogenous binding assays confirmed their inability to bind to the nuclear matrix. A non-S/MAR derived from At4g19120 locus was used as standard internal control in subsequent exogenous binding assays. ID letter according to Figure 8A, and 8B. Coordinate for non-S/MAR intergenic region: 17840000..17842500.
Table 8. Proportion of predicted S/MARs that matched experimental S/MARs in each S/MAR cluster.

<table>
<thead>
<tr>
<th></th>
<th>S/MAR Cluster</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>pS/MAR NO</td>
<td>114</td>
<td>15</td>
<td>117</td>
<td>181</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>pS/MAR YES</td>
<td>310</td>
<td>261</td>
<td>168</td>
<td>50</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>S/MAR Cluster</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>pS/MAR NO</td>
<td>0.26</td>
<td>0.05</td>
<td>0.41</td>
<td>0.78</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>pS/MAR YES</td>
<td>0.73</td>
<td>0.94</td>
<td>0.58</td>
<td>0.21</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>
Table 9. Association of S/MARs with Transposon Element Superfamilies (TE-SFs).

<table>
<thead>
<tr>
<th>TE Superfamily</th>
<th>TE-SF Count</th>
<th>S/MAR Count</th>
<th>S/MAR TE Count</th>
<th>P value</th>
<th>TE AT Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>121</td>
<td>29</td>
<td>31</td>
<td>0.0058</td>
<td>74.97</td>
</tr>
<tr>
<td>DNA/En-Spm</td>
<td>46</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>62.48</td>
</tr>
<tr>
<td>DNA/Harbinger</td>
<td>39</td>
<td>15</td>
<td>15</td>
<td>0.0011</td>
<td>67.83</td>
</tr>
<tr>
<td>DNA/HAT</td>
<td>131</td>
<td>20</td>
<td>20</td>
<td>0.49</td>
<td>66.81</td>
</tr>
<tr>
<td>DNA/Mariner</td>
<td>22</td>
<td>9</td>
<td>9</td>
<td>0.0004</td>
<td>80.6</td>
</tr>
<tr>
<td>DNA/MuDR</td>
<td>607</td>
<td>158</td>
<td>201</td>
<td>0</td>
<td>67.53</td>
</tr>
<tr>
<td>DNA/Pogo</td>
<td>58</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>79.52</td>
</tr>
<tr>
<td>DNA/Tc1</td>
<td>15</td>
<td>5</td>
<td>5</td>
<td>0.059</td>
<td>81.23</td>
</tr>
<tr>
<td>LINE?</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>65.51</td>
</tr>
<tr>
<td>LINE/L1</td>
<td>140</td>
<td>12</td>
<td>13</td>
<td>0.25</td>
<td>59.02</td>
</tr>
<tr>
<td>LTR/Copia</td>
<td>185</td>
<td>17</td>
<td>17</td>
<td>0</td>
<td>61.51</td>
</tr>
<tr>
<td>LTR/Gypsy</td>
<td>130</td>
<td>8</td>
<td>10</td>
<td>1</td>
<td>57.01</td>
</tr>
<tr>
<td>RathE1_cons</td>
<td>27</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>63.31</td>
</tr>
<tr>
<td>RathE2_cons</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>63.75</td>
</tr>
<tr>
<td>RathE3_cons</td>
<td>17</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>64.96</td>
</tr>
<tr>
<td>RC/Helitron</td>
<td>1653</td>
<td>357</td>
<td>546</td>
<td>0</td>
<td>76.46</td>
</tr>
<tr>
<td>SINE</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>61.56</td>
</tr>
<tr>
<td>Unassigned</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>59.36</td>
</tr>
</tbody>
</table>

Significant enrichment or depletion of TE-SFs in S/MAR sequences is highlighted in yellow and blue, respectively. Base pair (bp).
Table 10. Analysis of S/MARs and chromatin structures.

<table>
<thead>
<tr>
<th>S/MAR cluster</th>
<th>Predicted Nucleosome Occupancy</th>
<th>DNAmC</th>
<th>H3K9me2</th>
<th>H3K4me2/1</th>
<th>H3K56ac</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>df</td>
<td>P value</td>
<td>confint_low</td>
<td>confint_high</td>
</tr>
<tr>
<td>A</td>
<td>-31.71</td>
<td>1465.99</td>
<td>1.73E-16</td>
<td>-0.14</td>
<td>-0.12</td>
</tr>
<tr>
<td>B</td>
<td>-34.16</td>
<td>907.15</td>
<td>4.21E-165</td>
<td>-0.2</td>
<td>-0.18</td>
</tr>
<tr>
<td>C</td>
<td>-11.93</td>
<td>764.64</td>
<td>3.10E-30</td>
<td>-0.09</td>
<td>-0.06</td>
</tr>
<tr>
<td>D</td>
<td>8.31</td>
<td>625.34</td>
<td>5.73E-16</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>E</td>
<td>-10</td>
<td>426.3</td>
<td>2.49E-21</td>
<td>-0.09</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

Center and flanks refer to windows 10 and 11, and windows 7-9 and 12-14, respectively in Figures 18 to 22. T-tests were performed to determine the statistical significance of the observed enrichment or depletion of mean predicted nucleosomal occupancy for the S/MARs (center) as compared to the flanking region (flanks).
Table 11. Statistics for genes with TSS or TTS-associated S/MAR.

<table>
<thead>
<tr>
<th>S/MAR cluster</th>
<th>Gene TSS on</th>
<th>total</th>
<th>P value</th>
<th>estimate</th>
<th>confint_low</th>
<th>confint_high</th>
<th>null</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>130</td>
<td>266</td>
<td>0.2</td>
<td>0.48</td>
<td>0.4</td>
<td>0.56</td>
<td>0.55</td>
</tr>
<tr>
<td>B</td>
<td>56</td>
<td>142</td>
<td>0.0009</td>
<td>0.39</td>
<td>0.29</td>
<td>0.5</td>
<td>0.55</td>
</tr>
<tr>
<td>C</td>
<td>146</td>
<td>215</td>
<td>0.0007</td>
<td>0.67</td>
<td>0.59</td>
<td>0.75</td>
<td>0.55</td>
</tr>
<tr>
<td>D</td>
<td>100</td>
<td>167</td>
<td>1</td>
<td>0.59</td>
<td>0.49</td>
<td>0.69</td>
<td>0.55</td>
</tr>
<tr>
<td>E</td>
<td>59</td>
<td>102</td>
<td>1</td>
<td>0.57</td>
<td>0.44</td>
<td>0.7</td>
<td>0.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S/MAR cluster</th>
<th>Gene TTS on</th>
<th>total</th>
<th>P value</th>
<th>estimate</th>
<th>confint_low</th>
<th>confint_high</th>
<th>null</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>124</td>
<td>241</td>
<td>1</td>
<td>0.51</td>
<td>0.42</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td>B</td>
<td>54</td>
<td>124</td>
<td>0.055</td>
<td>0.43</td>
<td>0.32</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>C</td>
<td>111</td>
<td>192</td>
<td>1</td>
<td>0.57</td>
<td>0.48</td>
<td>0.66</td>
<td>0.55</td>
</tr>
<tr>
<td>D</td>
<td>89</td>
<td>155</td>
<td>1</td>
<td>0.57</td>
<td>0.46</td>
<td>0.67</td>
<td>0.55</td>
</tr>
<tr>
<td>E</td>
<td>52</td>
<td>92</td>
<td>1</td>
<td>0.56</td>
<td>0.42</td>
<td>0.69</td>
<td>0.55</td>
</tr>
</tbody>
</table>

The gene expression status was determined from published microarray results on the Affymetrix ATH1 array (Tanurdzic, et al.2008). For this analysis, the presence/absence calls (p<0.05) were used as determined by the MAS5 algorithm as implemented in the affy package for R (Smyth, 2005). Gene activity was calculated as the fraction of genes in each S/MAR cluster that are expressed. Statistical significance was determined using the binomial test in the R stats package with the null mean as the fraction of all genes on chr4 that are expressed. P values were adjusted by the Bonferroni correction for testing across the five S/MAR clusters, and 99% confidence intervals were calculated. Blue and yellow indicate genes with TSS-associated S/MARs are less and more likely to be expressed than the average gene, respectively.
Table 12. Gene Ontology (GO) enrichment analysis for genes with a TSS-associated S/MAR.

<table>
<thead>
<tr>
<th>GO term</th>
<th>GO_ID</th>
<th>aspect</th>
<th>Gene count</th>
<th>S/MAR count</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequence specific DNA binding transcription factor activity</td>
<td>GO:0003700</td>
<td>F</td>
<td>254</td>
<td>113</td>
<td>4.00E-07</td>
</tr>
<tr>
<td>regulation of transcription, DNA-dependent</td>
<td>GO:0006355</td>
<td>P</td>
<td>122</td>
<td>56</td>
<td>1.12E-03</td>
</tr>
<tr>
<td>nucleus</td>
<td>GO:0005634</td>
<td>C</td>
<td>311</td>
<td>114</td>
<td>2.72E-02</td>
</tr>
<tr>
<td>plasma membrane</td>
<td>GO:0005886</td>
<td>C</td>
<td>246</td>
<td>93</td>
<td>2.93E-02</td>
</tr>
<tr>
<td>DNA binding</td>
<td>GO:0003677</td>
<td>F</td>
<td>189</td>
<td>74</td>
<td>3.56E-02</td>
</tr>
<tr>
<td>transporter activity</td>
<td>GO:0005215</td>
<td>F</td>
<td>34</td>
<td>19</td>
<td>4.33E-02</td>
</tr>
<tr>
<td>AGRIS TfdB</td>
<td>NA</td>
<td>NA</td>
<td>252</td>
<td>108</td>
<td>1.03E-05</td>
</tr>
</tbody>
</table>

Aspect: (F) molecular function; (P) biological process; (c) cellular component. TFdB= AGRIS Transcription Factor data Base, which contains information on approximately 1,770 transcription factors (TFs). These TFs are grouped into 50 families, based on the presence of conserved domains.
REFERENCES


CHAPTER 3

Future Perspectives

Development of a Support Vector Machine to Predict S/MARs

It is thought that S/MARs play major roles in the spatial and functional organization of higher-order chromatin structure. To date, the only definition of a S/MAR element is operational and determined by binding to the nuclear matrix. While many sequence motifs and features have been associated with S/MARs, it remains unclear whether the presence of one or several such motifs is sufficient to establish a DNA fragment as a S/MAR (Michalowski et al., 1999; Liebich et al., 2002). For instance, classic S/MARs tend to be AT-rich, but not all AT-rich DNA fragments are S/MARs.

In silico prediction of S/MARs has been a challenging problem for more than a decade (Singh et al., 1997). Several S/MAR prediction algorithms such as MAR-Finder (also known as MARWIZ), ChrClass, and SMARTest have been developed for S/MAR prediction. All of the algorithms emphasize common features found in known S/MARs such as origin of replication, topoisomerase II binding sites, curved DNA, kinked DNA, SATB1 recognition sequence, MAR/SAR recognition signature, but also focus strongly on AT-richness of the putative S/MAR DNA (reviewed in Evans et al., 2007). However, all these features have been proven to be insufficient to identify S/MARs with good precision (Platts et al., 2006; Evans et al., 2007). An algorithm that is accurate for predicting S/MARs in various genomes could provide a window into the intricacies of higher-order chromatin structure and nuclear architecture.
Advanced computing techniques have enabled the use of iterative learning processes to analyze the massive amounts of new genomics information being produced. For example, iterative learning systems, termed Support Vector Machines or SVMs (Vapnik, 1995) have been applied to a number of tasks that involved searching for a particular pattern within sequences and for the construction of a prediction model (Noble, 2004). Increasingly SVMs are being used to address biological questions that require analysis of complex data sets, such as transcription factor binding sites (Bauer et al., 2010), improved DNA sequencing (Erlich et al., 2008), predicting RNA secondary structure (Gorodkin et al., 2010), and classification of Origin Recognition Complex (ORC)-associated regions (MacAlpine et al., 2010). While several algorithms have been developed for S/MARs prediction, only limited data have been available to support algorithm design, and, to date, SVM technology has not been applied to the problem.

Now that a larger dataset of experimentally determined S/MAR sequences is available, it is possible to test the hypothesis that the use of SVM can improve S/MAR prediction accuracy. A successful S/MAR SVM should both increase accuracy of S/MAR prediction and reduce the false discovery rate (performance). The identified S/MARs in *Arabidopsis* chr4 represent a large and diverse group of elements that seems to have multiple functions within the genome. Both the size of this group and the fact that we can identify subgroups within it should be very useful in developing better prediction tools.

Our project is an extension of our ongoing collaboration with Matt Vaughn (TACC, Univ. of Texas) to develop an SVM using our new S/MAR data from *Arabidopsis thaliana* chr4. The SVM will be trained using the *Arabidopsis* chr4 S/MAR dataset and used to predict
S/MARs for the entire *Arabidopsis* genome. The SVM S/MAR prediction will be tested by sequencing the S/MARs of the *Arabidopsis* genome.

The Vaughn group built the S/MARs classifier using an initial training set consisting of chr4 S/MARs and anti-S/MARs provided by Pete Pascuzzi from our group using microarray data I generated. Sequence algorithms for the initial S/MAR (and anti-S/MARs) are currently being developed. The SVM will be used to “trained” on S/MARs from a portion of the experimental dataset and then tested for its ability to predict the remainder. Modifications will be continued until the prediction accuracy cannot be further improved. If the SVM can accurately predict the experimental chr4 S/MARs, it will then be used to predict the S/MARs for the entire *Arabidopsis* genome. These predictions can then be tested experimentally. The generality of the predictions might eventually also be tested in different plant species, and perhaps in animal genomes as well.

Concurrently, we will isolate S/MARs from our *Arabidopsis* nuclei as described in Chapter 2 and the S/MARs will be sequenced to obtain data for the entire genome. The SVM S/MARs predicted will be compared with the experimental S/MARs (see next proposal below) and modifications to the SVM will be made as necessary. If successful, the use of SVM may also allow S/MAR prediction in additional plant and potentially animal species. Such comparisons will provide data on how similar or dissimilar S/MAR sequences are between closely and distantly related species.
Are the Diversity and Functions of Chromosome 4 S/MARs Conserved Globally Throughout the Arabidopsis Genome?

I have examined the distribution of S/MARs on Arabidopsis chr4 using high-resolution tiling microarrays. Experimentally, we will use the same approach to identify Arabidopsis thaliana S/MARs genome-wide. Therefore, our LIS isolation procedure will again be used, but instead of microarrays we will rely on high-throughput sequencing (Mardis, 2008) to characterize sequences that remain bound to the matrix. Genome-wide S/MAR identification will allow us to determine whether the characteristics of the five clusters and epigenetic characteristics identified for the chr4 S/MARs are consistent throughout the genome, and will provide a larger dataset for additional analysis.

The experimental design will be identical to what was used for chr4. Three biological replications will be done for each experiment. Putative S/MARs, identified by sequencing, will then be tested using the exogenous binding assay to validate the results. These data will provide a whole genome S/MAR map that can be compared to the results of other experiments to increase our understanding of S/MAR interactions with other elements throughout the genome. As noted above, our experimental S/MAR data will also be compared with SVM predictions to improve our understanding of sequence requirements for matrix binding and for possible use in other species.

Identification of Constitutive and Facultative S/MARs from Arabidopsis: S/MAR Mapping in Arabidopsis Dark-grown Suspension Cell Cultures

It has been proposed that S/MARs bind to the nuclear matrix in a constitutive (permanent) or facultative manner (cell type specific or stimulus-induced), depending on
their dynamic properties (Cockerill and Garrard, 1986; Cook, 1989; Boulikas, 1993; Paul and Ferl, 1998; Tikhonov et al., 2000). Although we found considerable diversity in our Arabidopsis chr4 S/MARs, we used a single cell line grown under a single set of conditions, so we do not presently have a basis for distinguishing between constitutive or facultative S/MARs.

Currently, the largest genomic region (4 Mb) examined for differential S/MAR binding is the major histocompatibility complex (MHC) locus from humans (Ottaviani et al., 2008). Ottaviani et al. (2008b) found differential S/MAR binding at the (MHC) locus following treatment of human fibroblast cells with interferon-gamma (IFNG). The authors identified constitutive, cell-specific, and IFNG-dependent S/MARs by comparing S/MAR profiles from untreated fibroblast, IFNG-treated fibroblasts and B-lymphoblastoid cells.

Recent cytological studies of chromosomes have shown that light causes large-scale reorganization of chromatin in Arabidopsis plants (Tessadori et al., 2007; Tessadori et al., 2009; Zanten et al., 2010; Fransz and de Jong, 2011). For example, Zanten et al. (2010) found that a decrease in the light intensity induced a large-scale heterochromatin decondensation in interphase nuclei of Arabidopsis leaves. Moreover, chromatin decondensation was a reversible process. When 3-week old plants were transferred to low light levels for 4 days the percentage of nuclei with normal heterochromatin content decreased from 80 to 10%. After returning the plants to normal light conditions the decondensation process was reversed, with 80% of the nuclei again showing a normal heterochromatin phenotype (Zanten et al., 2010).

Light, which has major effects on plant gene expression, also alters specific histone
modifications (Guo et al., 2008). Chromatin immunoprecipitation of H3K4me3, H3K9ac, H3K9me2 and H3K27me3 from *Arabidopsis* seedlings growing in continuous white light or in the dark showed different histone modification patterns in various light-regulated genes. For example, H3K9ac, which is highly correlated with transcriptional activation, was enriched in promoters of light-responsive genes in white light-grown seedlings, whereas H3K9me2 and H3K27me3, which are associated with heterochromatin, were enriched in the promoter regions of dark-grown seedlings. Taken together, these results suggest that plant cell growth in light or darkness can be used to induce changes in both chromatin organization and in gene expression and should provide a convenient system with which to investigate differential S/MAR binding.

It would be interesting to test the hypothesis that differential S/MAR binding occurs on *Arabidopsis* chr4 in cells growing under different environmental conditions known to affect chromatin structure. The *Arabidopsis* cell line used in the S/MAR mapping experiments can be grown in constant light or in heterotrophic dark conditions. Two lines of the *Arabidopsis* culture could thus be grown in light or dark conditions, respectively. Nuclear matrices prepared from the respective cell lines would be used to isolate the S/MARs. RNA be purified from the same cell lines would be analyzed by RNA sequencing (Mardis, 2008) to characterize gene expression patterns under the two conditions. The differences in mRNA abundance for specific *Arabidopsis* genes produced in either light or dark would then be compared to the location of the S/MARs from cells of the same biological replicate to examine potential correlations between changes in gene expression and S/MAR attachment to the matrix.
S/MAR binding in this experiment would be identified using deep sequencing of matrix-bound DNA fragments, and the results for light- and dark-grown cell lines compared to determine whether specific S/MAR binding was altered as a result of the light or dark growth. A constitutive S/MAR would be defined as being bound to the nuclear matrix, regardless of light or dark. Conversely, a facultative S/MAR would be defined as one for which the binding varies depending upon growth conditions.

These projects bring together a unique combination of experimental and bioinformatics studies of S/MARs, which will help to refine our ideas of the roles that different types of S/MARs play within the nucleus. I expect my work will help to open new strategies for high and stable transgene expression research in plants.

**Future Directions of the S/MAR Field**

The nuclear matrix has long been proposed as a central control center for the organization and control of nuclear processes (Jackson and Cook, 1985; Berezney, 1995; Nickerson, 2001). Unfortunately, studies of higher-order structure and nuclear architecture have been extremely problematic due to the major technological challenges required to preserve the biological structure without creating artifacts (Pederson, 1998). To address this question, future studies will necessary to determine whether the nuclear matrix is a true biological entity. While preliminary proteomics studies have provided a glimpse of the abundant proteins that remain following histone extraction, it remains an open question as to whether these proteins are part of a dynamic or static matrix, or possibly due to isolation artifacts (Jackson et al., 1990; Jack and Eggert, 1992; Pederson, 1998). While several studies
have identified nuclear matrix proteins of different organisms (Earnshaw et al., 1985; Bisotto et al., 1995; Mortillaro et al., 1996; Patturajan et al., 1998; Albrethsen et al., 2009), transcription factors (Bidwell et al., 1993; Berezney, 1995; Guo et al., 1995; Merriman et al., 1995; Nardozza et al., 1996), PCNA (Gerner and Sauermann, 1999), and other proteins associated with DNA replication (van Driel et al., 1996; Anachkova et al., 2005), all studies are limited to histone-extracted nuclei and very little is known about what occurs in the living cell.

To avoid possible artifacts, in vivo studies are necessary to allow a comparison of the nuclear localization and protein-protein interactions. In order to do such a study, it is first necessary to identify the nuclear matrix proteins. Calikowski et al. (2003) did a proteomic analysis of the A. thaliana nuclear scaffold (matrix) following removal of the histones with 25 mM LIS. They identified several proteins including homologues of nucleolar proteins Nop56, Nop58 (SAR-binding proteins), Nop140, fibrillarins, nucleolin, ribosomal components, a putative histone deacetylase, beta tubulins, and several unknown proteins (Calikowski et al., 2003). Identification of the A. thaliana nuclear matrix proteins is only the first step and subsequent studies are now required to determine whether such proteins assemble into a matrix structure in living cells. For example, it may be possible to use fluorescent fusion proteins and techniques such as FRET (Fluorescence Resonance Energy Transfer) to assess interactions of candidate proteins, and high resolution confocal or convolution microscopy of nuclei containing with such labeled proteins may allow more critical analysis of matrix structures in vivo. Future research with the combined application of biochemistry and cell biology to understand the protein complexes in living cells will help
to solve this enigma (Neumann et al., 2009; Rohila et al., 2009; Mravec et al., 2011; Simon et al., 2011; Wanke et al., 2011).

Previous studies have identified a number of proteins that bind to S/MARs (Dickinson et al., 1992; Dworetzky et al., 1992; Dickinson and Kohwi-Shigematsu, 1995; Liu et al., 1997). Chromatin immunoprecipitation (ChIP) with antibodies to the known *Arabidopsis* S/MAR-binding homologues could be used to identify whether the proteins interact with specific types of S/MARs, or simply AT-rich DNA. For example, High Mobility Group (HMG Y/I) proteins promote an open chromatin conformation and have been shown to be associated with S/MARs (Pedersen et al., 1991). In contrast, proteins such as histone H1 that are important for chromatin condensation may be associated with S/MARs in heterochromatic regions (Kas et al., 1993; Zhao et al., 1993).

In plants, only a few proteins are known to interact with S/MARs including filament-like protein (MFP1) (Gindullis and Meier, 1999), MAR binding protein 1 (MARBP-1) (Hatton and Gray, 1999), AT hook-containing MAR binding protein 1 (AHM1) (Morisawa et al., 2000), plant AT-rich sequence- and zinc-binding protein 1 (PLATZ1) (Nagano et al., 2001), AT-hook motif nuclear localized protein 1 (AHL1) (Fujimoto et al., 2004), and GIANT KILLER (GIK) (Ng et al., 2009). The ChIP identification of the DNA sequences bound by these proteins can be done using either array-based chromatin immunoprecipitation (ChIP-chip), or high throughput genomic sequencing (ChIP-seq). If S/MARs bound to a specific protein, are the S/MARs related? Furthermore, if a specific protein associated with a specific S/MAR (*in vivo*) it would be possible to use mutants, to learn the biological role that specific S/MAR-protein interactions may play.
What does the map of S/MARs across *Arabidopsis* chr4 tell us about chromatin architecture, and is it consistent with the chromatin loop domain model? The commonly accepted loop domain model shows that S/MARs are located at the bases of the chromatin loops. However, each S/MAR may not be the base of individual loops. Some S/MARs may not be bound at a given time or in a given cell type, and others may be bound to a separate part of the matrix from the loop base. In addition, successive, closely juxtaposed S/MARs may be placed together at the base of a chromatin loop by their common attachment to a particular set of nuclear matrix proteins. Certain S/MARs may be located within a loop rather than at the base, and S/MARs may anchor chromatin folds rather than loops. These hypotheses can be tested using biochemical approaches such as chromosome conformation capture (3C), which is a powerful tool for studying the *in vivo* three-dimensional organization of chromatin, since it identifies physical associations with distant loci throughout the genome (Dekker et al., 2002). A 3C analysis of the human *Ifng* locus indicates that certain S/MARs are either associated with a linear or looped chromatin conformation, indicating that not all S/MARs mediate the organization of chromatin into loop structures (Eivazova et al., 2007). In this view, certain S/MARs of *Arabidopsis* chr4 may have regulatory roles and would not necessarily be located at the loop base. Future studies to determine the three-dimensional chromatin architecture by 3C analysis will make it possible to determine which S/MARs mediate loop structure and which S/MARs have a regulatory role. Any possible overlapping between these processes could be distinguished by determining the frequency of spatial interactions that a given S/MAR may have with genes in two different plant tissues. If the S/MAR 3C interaction with genes is limited in both tissues, then, that S/MAR may have a
structural role defining chromatin loops. If the S/MAR 3C interacts with genes only in one tissue, then, that S/MAR may have both structural and regulatory roles. S/MAR 3C interaction with genes must be followed by changes in gene transcription in order to assign a regulatory role. Such 3C studies will increase our understanding of how higher-order chromatin structure and genome function are interrelated. These results will allow us to ask new questions, leading to new discoveries in this poorly understood area of biology.


