Title
A Conserved Role for Intragenic DNA Methylation in Cell Context-Specific Gene Regulation

Permalink
https://escholarship.org/uc/item/3hj2g30t

Author
Maunakea, Alika Keolaokalani

Publication Date
2008

Peer reviewed|Thesis/dissertation
A Conserved Role for Intragenic DNA Methylation in Cell Context-Specific Gene Regulation

by

Aika Keolaokalani Maunakea

DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
Biomedical Sciences
in the
GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
Acknowledgements

Throughout my years of training as a graduate student in the University of California, San Francisco, I have been fortunate to encounter and work with several people that not only supported my endeavors in the research that has made the completion of this thesis possible, but also aided in my progression in a career in science. For this, I thank my thesis advisor Dr. Joseph Costello whose patient and outstanding mentorship provided me with the skills and the motivation essential towards completing my graduate education and whose encouragement inspired me to continue a career in academic research. I also wish to thank the former and current members of his lab, Drs. Maarit Tiirikainen, Giuseppe Zardo, Ying Cheng, Peter Jun, Jean Ching, Ben Cadieux, Chibo Hong, Raman Nagarajan, and Shaun Fouse for all of their unconditional support, encouragement, and advice. In addition to my thesis advisor, I am very grateful to the members of my thesis committee Drs. Renee Reijo-Pera and William Weiss for their insightful feedback and discussions regarding the work described in this dissertation as well as for their continued support of my career. I would also like to thank my collaborators, Drs. Nicole Sodir, Yu-Cheng Wang, Ksenya Shchors, Michael Wong, Vivi Heine, David Rowitch, Peter Laird, William Held, Vivianne Ding, and Nancy Goodman for providing technical support for some of the experiments described in this thesis and/or for enhancing my experiences and education in graduate school. I also thank the BMS program administrators, Lisa Magargal, Monique Piazza, and Kevin Luong, for helping to make life in graduate school a little bit easier and enjoyable.

Finally, I wish to thank my family and friends whose names are too long to list here, but whose ever-present love and support have made my achievements possible. In
particular, I wish to dedicate this work to my wife Mei Lin, daughter Natalie, mother Alexis, and parent-in-laws Guo Mei and Ge Hua – “me ka mahalo a aloha pau’ole.”

Part of this thesis was supported by an NIH NRSA-F31 fellowship award. This thesis has been completed in partial fulfillment of the requirements of the Biomedical Sciences Graduate Program at the University of California, San Francisco. Where stated, part of the text of this thesis is a reprint of the materials as it appears in Ching, T. and Maunakea, A. et al 2005 published in the journal Nature Genetics. The co-authors listed in this publication helped direct and supervise the research that, in part, forms the basis for the thesis. This work is comparable to a standard thesis.
A Conserved Role for Intragenic DNA Methylation in Cell Context-Specific Gene Regulation

Alika Keolaokalani Maunakea

Abstract

Although DNA methylation is commonly found in the bodies of genes, its biological significance is unclear. Using a novel method for high-resolution analyses of the DNA methylation status of CpG islands, we identified and validated many new intragenic CpG islands that were methylated in a tissue-specific manner in normal human tissues, one of which was in \textit{SHANK3}.

Through a comparative epigenomics approach, we discovered that the tissue-specific DNA methylation levels of the intragenic CpG islands of \textit{SHANK3} were evolutionarily conserved in humans and mice. In addition, the expression levels of \textit{SHANK3} were tissue-specific and evolutionarily conserved. \textit{SHANK3} encodes a structural protein in the postsynaptic densities of excitatory neurons and altered copy number, and presumably aberrant expression levels of \textit{SHANK3}, underlie the abnormal neurological phenotypes in patients afflicted with 22q13 deletion syndrome and autism spectrum disorders.

The known function of \textit{SHANK3}, its association with neurological diseases, and the persistence of \textit{SHANK3} tissue-specific CpG island methylation and gene expression patterns over 75 million years suggests that the intragenic DNA methylation might have a conserved function in gene expression. Because DNA methylation is known to influence the activity of promoter sequences, we searched for genetic and epigenetic evidence of \textit{in vivo} promoter activity embedded within \textit{SHANK3} in mouse and human
tissues. Using this integrated and cross-species approach, we identified two intragenic regions with promoter activity 

*in vitro*, are differentially methylated 

*in vivo*, and facilitate the tissue-specific transcription of novel and potentially protein-coding 

*SHANK3* transcripts. These correlative data suggest that the intragenic methylation negatively influences promoter activity. Through a series of *in vitro* and *in vivo* DNA methylation experiments, we directly demonstrated that methylation of the intragenic promoters precluded expression.

Additionally, we found that 68.2% of intragenic sequences exhibiting tissue-specific DNA methylation identified from a genome-wide screen harbored features of promoters more commonly associated with the 5’-end of genes. Altogether, these results support a role for intragenic DNA methylation in regulating the activity of alternate, intragenic promoters that may be generally applicable. Importantly, this study also provides insight into the transcriptional regulation of *SHANK3*, the level of which is critical for normal brain development.
# Table of Contents

Acknowledgements........................................................................................................................... iii

Abstract........................................................................................................................................... v

List of Tables ..................................................................................................................................... x

List of Figures................................................................................................................................... xi

1. Chapter I: Introduction ............................................................................................................... 1
   1.1 Motivation................................................................................................................................. 1
   1.2 Background ............................................................................................................................... 3
       1.2.1 Epigenetics and gene regulation......................................................................................... 3
       1.2.2 An abridged history of DNA methylation................................................................. 9
       1.2.3 The evolution and roles of DNA methylation in eukaryotaes ............................... 10
       1.2.4 The study of DNA methylation: Technologies and Findings ................................. 13
       1.2.5 Intragenic DNA methylation is ancient........................................................................ 18
       1.2.6 Proposed roles for intragenic DNA methylation......................................................... 19
       1.2.7 SHANK3 and neurological diseases ............................................................................. 21
   1.3 Objectives............................................................................................................................... 24

2. Chapter II: A novel genome-wide DNA methylation technology .............................................. 25
   2.1 Summary ................................................................................................................................. 25
   2.2 Introduction ............................................................................................................................. 26
   2.3 Methods .................................................................................................................................. 28
       2.3.1 Tissues and primary cells................................................................................................. 28
       2.3.2 Methylation array analysis ............................................................................................. 28
       2.3.3 Predicting fragment hybridization computationally....................................................... 29
       2.3.4 Methylation of genomic DNA in vitro........................................................................... 30
       2.3.5 Bisulfite treatment, PCR and sequencing................................................................. 31
       2.3.6 Real-time RT-PCR ......................................................................................................... 31
   2.4 Results ..................................................................................................................................... 32
2.5 Discussion & Conclusion................................................................................. 46

3. Chapter III: Evolutionary conservation & function of intragenic DNA methylation 49
   3.1 Summary .......................................................................................................... 49
   3.2 Introduction ...................................................................................................... 50
   3.3 Methods............................................................................................................ 52
      3.3.1 Normal tissues and cultured primary cells ........................................... 52
      3.3.2 Isolation of nucleic acids ...................................................................... 54
      3.3.3 Demethylation and deacetylation experiments ................................. 54
      3.3.4 Bisulfite treatment, PCR and sequencing ............................................. 55
      3.3.5 5'-Rapid amplification of cDNA ends ...................................................... 55
      3.3.6 Reverse transcription, standard and real-time reverse transcription PCR
                     ............................................................................................... 56
      3.3.7 Integration of promoter-associated features....................................... 56
      3.3.8 Cloning of SHANK3 ECRs, transfection, and promoter-reporter assays..
                     ................................................................................................ 57
      3.3.9 In vitro DNA methylation assay .......................................................... 58
   3.4 Results .............................................................................................................. 58
   3.5 Discussion & Conclusion ................................................................................ 91

4. Chapter IV: Conclusions............................................................................................. 99
   4.1 Summary of Findings....................................................................................... 99
      4.1.1 Genome-wide CpG island methylation ...................................................... 99
      4.1.2 Evolutionary conservation of tissue-specific DNA methylation .......... 100
      4.1.3 Function of tissue-specific intragenic DNA methylation ..................... 101
   4.2 Implications of Findings ................................................................................ 102
      4.2.1 Comparative genomics and epigenomics ............................................. 102
      4.2.2 5’-promoter versus intragenic DNA methylation ............................... 103
      4.2.3 Transcriptional regulation of SHANK3 .......... ................................. 105
      4.2.4 Role of the novel SHANK3 transcript variants ................................. 108
4.3 Limitations ..................................................................................................................... 109

4.3.1 Genome-wide CpG island DNA methylation approach................................. 109

4.3.2 Other potential regulatory regions within \textit{SHANK3}................................. 110

4.3.3 Demethylation experiments ........................................................................... 111

4.3.4 Tissue heterogeneity ...................................................................................... 112

4.4 Future Directions ................................................................................................. 113

References...................................................................................................................... 118

Appendix........................................................................................................................ 150

UCSF Library Release ................................................................................................. 153
List of Tables

Table 1: Regions of tissue-specific intragenic DNA methylation commonly harbor promoter-associated features. ........................................................................................................ 91
Table 2: Primer sequences and cycling conditions..................................................................... 150
Table 3: List of confirmed differentially methylated loci analyzed for promoter-associated features. ........................................................................................................ 151
Table 4: List of primers, their applications, and reaction conditions. ................................. 152
List of Figures

Figure 1: Cytosine methylation reaction............................................................................ 5
Figure 2: The master scaffolding protein SHANK3........................................................ 22
Figure 3: Detection of methylation differences using a BAC clone array. ................. 33
Figure 4: The log2 ratios of cohybridizations of NotI fragments do not exhibit intensity-dependent effects nor geographic location-specific effects....................... 35
Figure 5: Detection of methylation differences between cell types using NotI or BssHII. .................................................................................. 36
Figure 6: The log2 ratios of cohybridizations of BssHII fragments do not exhibit intensity-dependent effects nor geographic location-specific effects.......... 38
Figure 7: Validation of the array results by bisulfite-sequencing and gene expression analysis. ............................................................................................................... 40
Figure 8: Maps of the chromosomal loci corresponding to BAC clones and their loci that exhibit tissue-specific methylation. ............................................................ 41
Figure 9: Evolutionary conservation of tissue-specific CpG island methylation and gene expression in the SHANK3 gene......................................................... 43
Figure 10: The tissue-specific CpG island methylation and gene expression of SHANK3 is conserved in rat brain and PBL....................................................... 45
Figure 11: The tissue-specific DNA methylation patterns of SHANK3 CpG islands are evolutionarily conserved................................................................. 60
Figure 12: Characterization of SHANK3 DNA methylation and expression in diverse mouse tissue/cell types and during development................................. 62
Figure 13: Novel tissue-specific transcripts of SHANK3 initiate from differentially methylated internal promoters................................................................. 70
Figure 14: Internal SHANK3 promoter regions are differentially methylated and their respective transcripts are differentially expressed.................................. 74
Figure 15: The internal SHANK3 promoters are differentially methylated.................... 75
Figure 16: Evolutionary conservation of ECR32 methylation and expression of the novel SHANK3 transcripts in humans...................................................... 77
Figure 17: Dynamics of 32t expression and ECR32 promoter methylation during development..................................................................................................... 79
Figure 18: DNA methylation of the internal *SHANK3* promoters suppresses their activity. ................................................................. 82

Figure 19: Genetic demethylation of the ECR32 promoter region in a lung sample correlates with increased expression levels of the associated 32t transcript. .......... 88

Figure 20: Region-specific differences of *SHANK3* DNA methylation in tissues of mice with hypomorphic alleles of *Dnmt1*. ................................................................. 90

Figure 21: Model of epigenetic regulation of *SHANK3* alternate, internal promoter usage. .............................................................................................................. 94
1. Chapter I: Introduction

1.1 Motivation

Early studies that measured 5-methylcytosine levels in genomic DNA showed tissue-specific differences in the overall levels of DNA methylation\(^1,2\). Determining the biological importance of this tissue-specificity has been an area of active interest, especially because DNA methylation is necessary for normal vertebrate development\(^3\). Significant controversy has also arisen regarding the potential role of tissue-specific methylation in gene expression, particularly during embryonic development\(^4\). In order to determine the distribution, extent, and site-specificity of normal tissue-specific DNA methylation, robust methods for genome-wide ‘profiling’ of DNA methylation are required. Studies using a 2D-gel approach with methylation-sensitive enzymes have confirmed that a proportion of single-copy genes harbor tissue-specific DNA methylation\(^5,6\), however this technique is limited to the analysis of a small fraction of the genome (<1\%) whose sequences are often unknown and is not easily expandable. We developed a novel approach that incorporated the similar single nucleotide resolution level offered by the use of methylation-sensitive enzymes but adapted to microarrays in order to evaluate DNA methylation throughout a greater fraction of a genome\(^7\).

Many approaches to map DNA methylation genome-wide have been published\(^8-16\). Two conclusions common among each of these studies are (1) contrary to the prevailing view that CpG islands remain methylation-free regardless of the expression state of the associated gene, a substantial fraction (2-8\%) of CpG islands are in fact differentially methylated in normal tissue/cell types,
and (2) DNA methylation within the body of genes, referred to as ‘intragenic’ DNA methylation, is far more common throughout the genomes of many species than previously appreciated. Although the methylation within CpG islands at the 5′-promoter regions of genes silences the expression of the underlying gene\textsuperscript{17}, the function of intragenic DNA methylation has been more controversial. From our analysis of genome-wide DNA methylation, we identified a gene called \textit{SHANK3} that contains three intragenic CpG islands all of which exhibit tissue-specific DNA methylation levels. To initially assess whether intragenic methylation may have a function, we used a comparative epigenetic approach and found that the intragenic DNA methylation patterns of \textit{SHANK3} are evolutionarily conserved. To our knowledge, the intragenic DNA methylation of \textit{SHANK3} is the first example of evolutionarily conserved tissue-specific DNA methylation that is not associated with imprinting or X-chromosome inactivation. We further investigated the intragenic DNA methylation of \textit{SHANK3} in order to uncover a role for the methylation and discovered that the methylation is involved in regulating the expression of novel \textit{SHANK3} transcripts. Furthermore, although the expression level of \textit{SHANK3} is critically important for normal brain development\textsuperscript{18,19}, how \textit{SHANK3} is transcriptionally regulated is unknown. Thus, while the main goal of this thesis is to investigate the role for the intragenic DNA methylation, this study also provides insight into the transcriptional regulation of \textit{SHANK3} in particular. To determine whether the role of intragenic DNA methylation in \textit{SHANK3} is more widely applicable, we also analyzed 35 additional sites of tissue-specific methylation that were identified by our collaborators.
1.2 Background

*Epigenomics is where genomics was 30 years ago, when everyone was working on part of the puzzle.* – Peter Jones

1.2.1 Epigenetics and gene regulation

Epigenetic processes may explain in part several of the non-Mendelian inheritance patterns in plant and animal systems. Coined in 1942 by C. H. Waddington, the word ‘epigenetics’ was used to describe “the interactions between genes and their products which bring phenotype into being” and has since been adapted to describe the study of mitotically and/or meiotically heritable changes in gene expression that are not caused by changes in DNA sequence. More recently, refining the definition of epigenetics to “the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states” has been proposed as to avoid constraints imposed by stringently requiring heritability, since cells can respond to various extracellular signals (e.g., paracrine, endocrine, and nutrient) by appropriately altering gene expression. Regardless of the precise definition, it is widely recognized that epigenetic modifications of chromatin structure interact with transcription factors to orchestrate gene expression programs that underlie the identity and function of a cell.

Epigenetic gene regulation requires molecular mechanisms that encode information beyond the DNA sequence. Three classes of such molecular mechanisms include DNA methylation, histone modifications, and DNA-binding proteins. In plants, RNA interference contributes to epigenetic regulation, however whether there is a conserved mammalian equivalent remains unclear.
All epigenetic mechanisms, directly or indirectly, influence the underlying chromatin structure and presumably affect access of the transcription machinery to their target DNA sequences.

The best characterized epigenetic mechanism is DNA methylation, which in mammals occurs almost exclusively within 5’-cytosine-guanine-3’ dinucleotides (CpGs), although CpNpG methylation has also been detected\(^{30}\). Specifically, the 5-carbon position of cytosine within a CpG is subjected to the enzymatic addition of a methyl-group to yield 5-methylcytosine (m\(^5\)C) (Figure 1A). DNA methylation influences gene expression by affecting the binding of methylation-sensitive DNA binding proteins and/or by interacting with various modifications of histone proteins that alter DNA accessibility\(^ {17,31}\) and is generally associated with gene silencing. DNA methylation is maintained through mitosis primarily by the DNA methyltransferase 1 (DNMT1) enzyme which associates with PCNA and the replication foci and has a significant preference for action on hemi-methylated DNA following DNA replication\(^ {32,33}\) (Figure 1B). This mechanism allows for perpetuation of the DNA methylation state in newly formed cells.
Figure 1: Cytosine methylation reaction.

(A) DNA methyltransferases (DNMT) utilize the universal methyl-donor S-adenosylmethionine (SAM) to create 5-methylcytosine. (B) Re-establishment of cytosine methylation upon DNA replication.

Since the first description of histone modifications by Allfrey et al in the early 1960s, histone proteins that package nuclear DNA are known to be subjected to a wide array of post-translational modifications on specific residues along their NH2-terminal ‘tails’ that project far outside of the nucleosome core. Such modifications include acetylation, methylation, phosphorylation, ubiquitination, and sumoylation, and potentially others. Since these modifications contribute to chromatin conformation, the potential information stored within different combinatorial patterns of these modifications led to the hypothesis of a ‘histone code’. The histone code hypothesis proposes that there
are specific combinations of modifications that dictate locus-specific transcriptional competence\textsuperscript{37}. One of the best understood histone modification is the reversible acetylation of lysine residues on histones H3 and H4\textsuperscript{38}. Acetylation is accomplished by specific histone acetyltransferases (HATs) that neutralizes the positive charge of the histone tail and facilitates access of transcription factors to the underlying DNA\textsuperscript{39-41}. While histone acetylation is clearly associated with transcriptional activation\textsuperscript{42-45}, the reverse reaction catalyzed by histone deacetylases (HDACs) increases the tail’s positive charge, lowers the transcription potential of the underlying DNA, and is associated with a transcriptionally repressed state\textsuperscript{46-49}. In contrast to acetylation, methylation of specific lysine residues on histones, such as H3-K9 or H3-K27, is associated with transcriptional repression\textsuperscript{37}. Histone methylation is carried out by specific methyltransferases\textsuperscript{50,51}. How histone methylation accomplishes gene repression is not fully understood, however it is known that methylation of certain lysine residues, such as H3-K9, acts as a docking site for heterochromatin protein 1 (HP1) which in turn recruits histone methyltransferases\textsuperscript{52,53}. Although histone phosphorylation, ubiquitylation, and sumoylation of specific serine, threonine, or lysine residues are involved in transcriptional regulation, modifications on other residues are associated with mitosis, DNA repair, and apoptosis\textsuperscript{35}. A more detailed description of these modifications and how they are accomplished are beyond the scope of this thesis. New evidence suggests the possibility that histone modifications could be mitotically heritable, though this is debated. In yeast for example, the heterochromatic states mediated by the interactions of
hypoacetylated histones and silent-information-regulator (SIR) proteins or H3-K9 methylation and the Swi6 chromodomain are maintained through cell division\textsuperscript{54}. In flies, H3-K27 and H3-K4 methylation, catalyzed by Polycomb-group (PcG) and trithorax-group (trxG) protein complexes, mediate mitotic inheritance of lineage-specific gene expression patterns\textsuperscript{55}. The extent to which other histone modifications are heritable remains unclear.

Another epigenetic mechanism is autoregulation by transcription factors. Gene expression is regulated by binding of transcription factors to their target sites within gene promoters. Those transcription factors that activate their own promoters can perpetuate their expression through cell division, such as been demonstrated for MyoD\textsuperscript{56}. When MyoD is transcriptionally activated during cell division, its protein product is partitioned to both daughter nuclei, and activates its own transcription as well as other target genes. Thus, autoregulatory transcription factors enable mitotic heritability of gene expression\textsuperscript{57}.

All three mechanisms operate in an orchestrated, mutually reinforcing manner to maintain the epigenetic states of differenced cells\textsuperscript{57}. For instance, various histone modifications can promote regional CpG methylation which can then recruit methyl-binding domain proteins (MBDs) that simulate binding of histone modification enzymes such as HDACs and HMTs which bring about histone deacetylation and methylation that alters the local chromatin environment, thereby modulating transcription factor binding and transcription potential\textsuperscript{58-60}. These interconnections were revealed by biochemical analyses of histone modifying enzymes and MBDs, such as HDAC and MeCP2\textsuperscript{61,62}. 
Interacting with DNA methylation and histone modification enzymes, complexes of ATP-dependent chromatin remodeling proteins, such as SWI/SNF, are capable of inducing changes to the underlying configuration of chromatin to regulate gene expression\textsuperscript{63}. Specifically, the SWI/SNF complex appears to be required for the establishment of DNA methylation patterns in the genome by reorganizing histone-DNA contacts and provides transcriptional regulators access to the underlying DNA\textsuperscript{64}. How these interactions are temporally coordinated remains unclear, however results from chemical studies aimed at understanding the effects of histone modifying and DNA methylation enzymes offer some insight into how these mechanisms are linked. A common occurrence in transformed cells is the silencing of tumor suppressor genes mediated by aberrant hypermethylation of their promoter regions often overlapping with CpG-rich sequences called CpG islands\textsuperscript{65}. In certain instances, although treatment of these transformed cells with the potent HDAC inhibitor trichostatin A (TSA) fails to upregulate hypermethylated genes, induction of gene expression has been observed in combination with the DNA demethylating agent 5-azacytidine\textsuperscript{66}. This synergy illustrates the inter-relationship of distinct epigenetic modifications such as DNA methylation and histone acetylation in cancer cells. DNMT1 also complexes with HDAC1, Rb, and the transcriptional regulator E2F1 and functions to repress transcription of E2F-responsive promoters\textsuperscript{67}. This example establishes a molecular link between DNA methylation, histone modifications, and DNA-binding proteins.
1.2.2 An abridged history of DNA methylation

First recognized in 1948, 5-methylcytosine (m^5C) was identified as a minor base in DNA^{68,69}. The origin of this modification remained unknown until 1963 when specific DNA methyltransferases were isolated in bacteria^{70} and then a few years later in mammals^{71}. Although the biological significance of DNA methylation remained unknown in animals, early studies demonstrated tissue-specificity of overall 5-methylcytosine levels^{1,2,72}. In 1975 it was proposed that DNA methylation might be responsible for the stable maintenance of a particular gene expression pattern through cell division^{73,74}. This hypothesis predicted the existence of a maintenance DNA methyltransferase that preferentially methylates hemi-methylated sites. This prediction was later supported by experiments demonstrating the clonal inheritance of methylation patterns in mammalian cells^{75,76} and by the cloning of the first mammalian DNA methyltransferase, DNMT1, in 1988^{77}. Later studies showed that DNMT1 methylates hemi-methylated CpG dinucleotides with a higher efficiency than unmethylated DNA in vitro^{78} and localizes to DNA replication foci^{32}.

In the early 1990s, it was shown that genetic ablation of DNMT1 in mice resulted in misregulation of epigenetically silenced loci and embryonic lethality^{3}. Further genetic studies suggested that enzymes other than DNMT1 might be responsible for de novo methylation in vivo since DNMT1-deficient embryonic stem cells were still capable of methylating retroviral DNA de novo^{78}. The importance of DNA methylation in development was reinforced by genetic ablation of the de novo DNA methyltransferases DNMT3a and DNMT3b, which resulted in postnatal and prenatal lethality in mice, respectively^{79}. 
1.2.3 The evolution and roles of DNA methylation in eukaryotae

In eukaryotes from plants to humans, DNA methylation is found exclusively at cytosine residues. In mammals, cytosine methylation occurs within symmetrical CpG dinucleotides, whereas in fungi (e.g. *Neurospora crassa*) and plants (e.g. *Arabidopsis thaliana*), methylation is observed in various symmetrical and asymmetrical sequence contexts. Wide variation in the content of genomic cytosine methylation among distinct species of insects has been observed from low levels of methylation reported in the fruit fly *Drosophila melanogaster* to substantial methylation found in the honey bee *Apis mellifera*. However, not all eukaryotes harbor DNA methylation; the yeast *Saccharomyces cerevisiae* and the nematode worm *Caenorhabditis elegans* are completely devoid of cytosine methylation and their genomes lack recognizable DNMT-like genes.

Potential functions for the apparent evolutionarily conserved DNA methylation include silencing transposable elements, mediating developmental gene regulation, or reducing transcriptional noise. Indeed, DNA methylation in mammals is essential for embryonic development, differentiation, and cell cycle control. Additionally, DNA methylation plays critical roles in maintaining transcriptional silencing of genes on the inactive X-chromosome and imprinted genes. In mammals, genetic disruption of DNMTs caused apoptosis in embryos, widespread derepression of ectopic gene expression and transcriptional activation of transposable elements. Human diseases have been associated with abnormal DNA methylation patterns including cancer, ICF (immunodeficiency, centromeric instability, facial anomalies) syndrome, ATRX
(alpha-thalassemia, mental retardation) syndrome, and fragile X syndrome. In plants, DNA methylation deficiency resulted in spurious transcription initiation from cryptic sites, demonstrating a role for the methylation in reducing transcriptional noise. The central theme of these findings is that DNA methylation functions to maintain a repressed chromatin state, thereby stably silencing promoter activity. It is important to note, however, that DNA methylation is not always associated with gene silencing. Some studies have demonstrated that DNA methylation can augment expression of an imprinted gene by blocking the binding of repressor proteins to silencer elements within the gene. Thus, throughout eukaryotes, DNA methylation plays multiple and complex roles.

DNA methylation is dynamic during embryogenesis and gametogenesis in mammals. Active demethylation occurs in the male pronucleus by an as yet undefined mechanism. Following zygotic formation, the paternal and maternal chromosomes undergo passive demethylation that eliminates most, but not all, DNA methylation marks inherited from the gametes by the blastocyst stage. DNA methylation marks on imprinted genes are protected from demethylation, preserving parental imprints. After implantation, embryonic DNA methylation patterns are established through lineage-specific de novo methylation that begins in the inner cell mass of the blastocyst. In the primitive ectoderm that gives rise to the embryo proper, DNA methylation levels rapidly increase, whereas in the trophoblast, which gives rise to the placenta, and in the primitive endoderm, which gives rise to the yolk-sac membrane, genomic
DNA is undermethylated\textsuperscript{98}. A wave of DNA demethylation occurs again during gametogenesis. Around embryonic day 11.5-12.5, primordial germ cells (PGCs) undergo demethylation at imprinted loci that erases parental imprinting marks, but preserves methylation of other regions including transposable elements\textsuperscript{99,100}. During oogenesis, the maternal-specific genomic imprints are re-established through \textit{de novo} methylation activities of DNMT3a and DNMT3b, and an associated co-factor DNMT3l\textsuperscript{101-103}. Between the leptotene and pachytene stages of spermatogenesis, several factors appear to function to establish paternal-specific genomic imprints including histone hypoacetylation, Suv39h, DNMT3a, and DNMT3l\textsuperscript{102,104,105}. The role of these dynamic demethylation and remethylation processes during early development is not fully understood. However, it has been proposed that genome-wide demethylation during pre-implantation development may lead to chromatin decondensation and thus transcriptional activation of the zygotic genes essential for early development and/or may facilitate reprogramming of the genome through interacting with histone modifications and chromatin remodeling. In contrast, remethylation may be necessary to repress retrotransposons and establish a global gene silencing state required for embryonic development\textsuperscript{106}. Indeed, recent studies investigating the molecular mechanisms involved in nuclear reprogramming have suggested that correct DNA methylation patterns are required for successfully reprogramming differentiated adult cells into pluripotent embryonic stem cell like cells\textsuperscript{107,108}. 
Additional insights into the roles of DNA methylation in cell function and genome stability are being gleaned from new technologies for genome-wide DNA methylation mapping. These technologies produce a higher resolution view of the genomic distribution of DNA methylation (next Section).

1.2.4 The study of DNA methylation: Technologies and Findings

Largely based on evidence from genomic DNA digestion with methylation-sensitive enzymes, CpG methylation was found throughout the genomes of many eukaryotic organisms with the exception of cytosines that resided within CpG-rich sequences called CpG islands\textsuperscript{80,109}. Approximately half of all CpG islands genome-wide are found near the promoters of genes and are defined as being at least 200 bases long, having a G+C content of greater than 50\% and an observed to expected CpG ratio equal to or greater than 0.6\textsuperscript{110}. More recently derived CpG island criteria are slightly more stringent\textsuperscript{111}. The early studies of DNA methylation, however, are limited to the analysis of CpGs within the particular sequence motifs recognized by the restriction enzymes used. Thus, the genome-wide distribution of DNA methylation in a given tissue/cell type of an organism, the so-called ‘methylome’\textsuperscript{112}, remained poorly characterized. Recently, however, many novel approaches have been developed to map DNA methylation genome-wide, though none yet assesses methylation in all 28 million CpGs.

Four main approaches that allow detection of DNA methylation are digestion of DNA by methylation-sensitive or –insensitive restriction endonucleases\textsuperscript{113}, the chemical modification of DNA by sodium bisulfite\textsuperscript{114},
immunoprecipitation of 5-methylcytosine to separate unmethylated and methylated fractions of the genome\textsuperscript{115}, and enrichment of methylated DNA using DNA binding proteins specific to methylated DNA\textsuperscript{116}. All four approaches have been coupled to high-throughput technologies. The restriction enzyme approach yields methylation information of one site, the sodium bisulfite method yields nucleotide-specific data over a sequence of limited size (~100-bp to 400-bp), and the immunoprecipitation approach gives a general indication of methylation over a larger region without the single nucleotide resolution. Thus, these approaches are complementary and would potentially generate mutually reinforcing and comprehensive DNA methylation patterns.

We coupled restriction landmark genome scanning (RLGS) methods with high-throughput microarray technology to significantly improve genomic resolution and detection sensitivity of DNA methylation differences between normal tissue types specifically at CpG islands\textsuperscript{7}. The development of this technique as well as relevant findings are discussed in detail in Chapter II. Briefly, we have used BAC clone arrays that were hybridized with biotin-labeled DNA generated from end-filling of genomic DNA samples digested with the rare-cutting methylation-sensitive restriction enzymes \textit{BssHII} or \textit{NotI}. Approximately 89-94\% of these restriction enzyme recognition sites occur within CpG islands and so by sampling different normal human tissue/cell types, we were able to identify a number of differentially methylated CpG islands, including one within the \textit{SHANK3} gene (discussed in detail in \textit{Chapters II} & \textit{III}).
After the development of RLGS, one of the first methods to screen for differential methylation also utilized methylation-sensitive enzymes but in combination with CpG island microarrays\textsuperscript{117}. Another method that combined restriction enzyme digestion and microarray technology, known as HELP (\textit{HpaII} tiny fragment Enrichment by Ligation-mediated PCR)\textsuperscript{118}, utilizes differential hybridization of unmethylated genomic DNA enriched by \textit{HpaII} digestion and total genomic DNA digested with the methylation-insensitive isoschizomer \textit{MspI} onto a customized oligonucleotide array. Even though a larger number of fragments can be analyzed by both methods, detection of DNA methylation differences are still limited to the CpGs within the particular sequence motifs recognized by the specific restriction enzymes used.

To circumvent the limitations of restriction enzymes, a second approach for high-throughput DNA methylation analysis was developed using sodium bisulfite treatment of DNA. Sodium bisulfite chemically converts all unmethylated cytosines to uracil by hydrolytic deamination of the 5,6-dihydrocytosine-6-sulphonate product at the C4 position of the pyrimidine ring. However, a methyl group at the C5 position inhibits this reaction, protecting methylated cytosines from conversion\textsuperscript{119,120}. Upon PCR amplification using primers corresponding to the bisulfite-converted sequences of interest, the products are cloned and sequenced to evaluate CpG methylation throughout the amplified region\textsuperscript{121}. This ‘bisulfite-sequencing’ method has been widely used to map methylation of specific loci at single-nucleotide resolution. This method has been adapted to analyze DNA methylation in a large number of amplicons from
three human chromosomes, but its use for high-throughput methylation analysis is limited by the labor-intensive nature of this method. Recently, some applications of ‘bisulfite-sequencing’ have been developed to evaluate methylation globally. Two such methods include a shotgun-sequencing approach and a reduced representation shotgun sequencing approach. Bisulfite treatment of genomic DNA combined with ultra-high-throughput sequencing by a shotgun approach using Solexa sequencing technology allows sensitive measurement of cytosine methylation on a genome-wide scale within specific sequence contexts. So far, this has been successfully applied to the Arabidopsis thaliana genome and to the mouse genome. One major limitation of this approach is the very large amount of sequencing required in order to yield comprehensive and accurate DNA methylation data. To overcome this limitation, a large-scale random approach termed ‘reduced representation bisulfite sequencing’ (RRBS) has been devised. In RRBS, genomic DNA is first digested with specific infrequent cutting restriction enzymes such as BglII or MspI, fragments of a certain size range are selected, and these fragments are then ligated to adapters all before treating the DNA with bisulfite, which is followed by PCR amplification, cloning and sequencing of these fragments. Although this approach provides useful nucleotide-level quantitation of CpG methylation, the selection of fragments in a given size range explicitly excludes entire genome coverage.

A third approach that aims to evaluate DNA methylation genome-wide takes advantage of an anti-methylcytosine antibody that can be used to enrich for
methylated sequences within genomic DNA. This approach, termed methylated DNA immunoprecipitation (MeDIP), was used in a high-resolution analysis of methylated sequences coupled to a comparative genomic hybridization microarray in human samples\textsuperscript{115} and also to profile DNA methylation in the Arabidopsis genome\textsuperscript{13,14}.

A fourth strategy to evaluate DNA methylation is to first isolate and enrich for methylated DNA with an MBD domain fused to a human IgG\textsuperscript{123} or with MBD proteins bound to a sepharose matrix\textsuperscript{116} to hybridize onto an array. Limitations of these affinity approaches include the bias that CpG-rich sequences give higher enrichments than methylated CpG-poor sequences and the microarray-based detection cannot measure allelic DNA methylation or methylation of individual repetitive elements. Use of bisulfite-sequencing approaches overcomes these limitations.

Results common to all four main approaches allowed two unexpected discoveries: (1) in contrast to the long-held view that cytosines within CpG islands remain free of DNA methylation, studies of a large number of CpG island-containing promoters have observed that a significant fraction are in fact methylated in a tissue/cell type-specific manner where the methylation status is negatively correlated with the expression state of the associated gene, and (2) these large-scale studies have demonstrated that DNA methylation is targeted to gene bodies of actively transcribed genes in many divergent genomes including those from invertebrates, plants, and mammals, seemingly contradicting its role in gene silencing. The dynamic patterns of DNA methylation in promoters
reinforce the more than 30-year old hypothesis that DNA methylation has a role in development. Additionally, the surprising extent of DNA methylation within bodies of actively transcribed genes suggests a potentially novel mechanism for DNA methylation in influencing gene expression.

1.2.5 Intragenic DNA methylation is ancient

Intragenic or gene body DNA methylation has been observed in individual genes in invertebrate and vertebrate genomes, but the extent of intragenic methylation genome-wide has only recently become apparent using newer methylome technologies, as discussed above. Cross-species comparisons of the distribution of genomic methylation suggests that intragenic DNA methylation may be evolutionarily ancient.\(^{124}\)

Using methylation sensitive restriction enzymes, intragenic methylation was initially mapped in the invertebrate chordate Ciona intestinalis.\(^{125}\) Bisulfite-sequencing and computational analyses over the C. intestinalis genome showed that intragenic methylation occurs in approximately 60% of all genes.\(^{126}\) CpG methylation within the transcriptionally active esterase E4 gene in the aphid Myzus persicae was also observed.\(^{127,128}\) Intragenic DNA methylation was also observed in several honeybee genes.\(^{81}\) Taken together, intragenic DNA methylation seems to be a common feature of invertebrate genomes.

Similar to invertebrates, plants show substantial gene body methylation. Two independent studies using MeDIP coupled with microarrays demonstrated that approximately 33% of all genes in the flowering plant A. thaliana genome exhibit intragenic DNA methylation.\(^{13,14}\) Interestingly, the methylation was more
centrally located in the gene, being less common near the 5’ and 3’ end of the transcription units, similar to patterns observed in invertebrates.

In mammals, the majority of CpG cytosines are methylated in normal somatic tissues. One prominent fraction of the genome that remains largely unmethylated include CpG-rich sequences defined as CpG islands, which are present at the 5’-ends of approximately half of all genes. However, as mentioned in the previous section, 2-8% of CpG islands are normally differentially methylated, many of which appear to be intergenic or intragenic.

The similar distribution of DNA methylation at gene bodies throughout the genomes of organisms that diverged approximately 1.6 billion years ago implies that intragenic DNA methylation has a common ancestral function. Proposed models for the function of this non-promoter DNA methylation include inhibition of cryptic initiation, protection against mobile elements and maintenance of genome stability among other functions. These models could explain constitutive intragenic DNA methylation, but they do not explain the purpose for intragenic methylation that is tissue-specific.

1.2.6 Proposed roles for intragenic DNA methylation

A small fraction of CpG islands are methylated in a tissue-specific manner. However, CpG islands that are not associated with transcriptional start sites (those within or between genes) are significantly more likely to be methylated than those at gene promoters (7% versus 16%). Bisulfite-sequencing of 2,524 amplicons may in part explain the tendency that
differentially methylated regions are over-represented outside of known gene promoters, such as within genes\textsuperscript{8}. Importantly, some of the tissue-specific DNA methylation that occurs within gene bodies is evolutionarily conserved between humans and rodents\textsuperscript{7,8}. Together, these data strongly suggest a functional role for the tissue-specific intragenic DNA methylation.

Initial data from plants and mammals provide a possible link between intragenic DNA methylation and gene transcription. Deficiency of the DNA methyltransferase gene, \textit{MET1}, in \textit{A. thaliana} resulted in increased expression of genes genome-wide, particularly of those that normally exhibit intragenic DNA methylation\textsuperscript{14}. While the mechanism for the increased expression is unknown, intragenic methylated regions in \textit{A. thaliana} corresponded with regions of active elongation by RNA polymerase II (polII), whereas the 5’ and 3’ gene extremities lacking DNA methylation often harbored higher polII density suggesting a possible stall in elongation\textsuperscript{14}. These indirect results raises the possibility that intragenic DNA methylation interferes with polII elongation. Consistent with this proposed effect on elongation, intragenic DNA methylation of a chromosomally integrated transgene reduced the efficiency of transcriptional elongation through polII exclusion, and also diminished the overall mRNA level of the transgene\textsuperscript{135}. Additionally, the active X chromosome (X\textsubscript{a}) displays more allelic-specific methylation than that of the inactive X chromosome (X\textsubscript{i}) and the methylation specifically resides within gene bodies\textsuperscript{136}. On the X chromosome, there was no bias for the presence of methylation in repetitive elements embedded in genic amplicons. The intragenic methylation observed in the active X chromosome is
associated with elevated expression of genes\textsuperscript{137} and may hint at a functional role for the methylation in one aspect of dosage compensation in somatic cells\textsuperscript{138}. However, the question of how tissue-specific intragenic DNA methylation may be involved in gene regulation \textit{in vivo} remains to be addressed.

Using methylation-sensitive enzymes coupled with BAC microarrays, we identified differential intragenic DNA methylation in CpG islands embedded within the \textit{SHANK3} gene in human tissues. Comparing DNA methylation patterns of the orthologous sequences in analogous tissue types from mice and rats demonstrated that tissue-specific intragenic DNA methylation is evolutionarily conserved\textsuperscript{7}. Because of this conservation, we further investigated \textit{SHANK3} as a model to determine the role for intragenic DNA methylation in gene expression.

\subsection*{1.2.7 \textit{SHANK3} and neurological diseases}

The SHANK protein family, comprised of SHANK1, SHANK2 and SHANK3, is characterized by common protein interaction motifs or domains including ankryin repeats, followed by SH3- and PDZ-domains, a long proline-rich region, and a C-terminal SAM domain\textsuperscript{139}. SHANK proteins range in size from 120 kDa to more than 240 kDa\textsuperscript{139}. Western blot analyses indicate the presence of multiple isoforms of SHANK proteins in which individual parts of the domain structure may be missing due to alternative splicing and/or alternate promoter usage\textsuperscript{139-143}. In neurons, SHANK proteins including SHANK3 are localized almost exclusively to the postsynaptic densities (PSDs) of excitatory synapses. At the synapse, they interact with smaller adaptor proteins such as
PSD-95/SAP90, GRIP/ABP, and Homer/Vesl that anchor NMDA, AMPA, and metabotropic glutamate receptors, respectively⁠¹³⁹,¹⁴⁰,¹⁴⁴-¹⁵³ (Figure 2). Thus, the SHANK proteins are considered ‘master’ scaffolding molecules⁠¹⁵⁴.

**Figure 2: The master scaffolding protein SHANK3.**

Diagram of the postsynaptic density (PSD), a cytoskeleton specialization at neuronal synapses, comprised of L-glutamate neurotransmitter receptors, their molecular scaffolding molecules, cell adhesion molecules and a diverse set of other signaling proteins. SHANK3 is a master scaffolding protein thought to concentrate and organize neurotransmitter receptors to respond rapidly to neurotransmitter release in the synaptic cleft. Alterations in *SHANK3* or molecules of the intracellular cascade originating from SHANK3 are associated with cognitive and behavioral disorders. All known mutations identified within *SHANK3* in patients with autism spectrum disorders account for 1% of cases¹⁹,¹⁵⁵. Part of this figure was adapted with modifications from Sheng *et al*¹⁵⁶.

Prior to this dissertation work, how *SHANK* genes are transcriptionally regulated was not known. However, all *SHANK* genes appear to be co-expressed in neurons of the cortex and hippocampus. Interestingly, in cerebellum, the expression of *SHANK3* is restricted to granule cells, while *SHANK1* and
SHANK2 are found only in the Purkinje cell layer\textsuperscript{139,157}, suggesting that the SHANK proteins play distinct, non-redundant roles. Additionally, SHANK3 expression is not restricted to brain. In epithelial cells, SHANK3 is crucial for mediating sustained Erk-MAPK and PI3K signal transduction initiated from a tyrosine kinase receptor\textsuperscript{158}. Consistent with its known role in the PSD, SHANK proteins are important for normal brain function. Defects of SHANK1 or SHANK2 have not yet been reported in human diseases, however mice with a constitutive deletion of SHANK1 have been generated. Neurons from these mice harbor thinner PSDs and weaker basal synaptic transmission, both of which are associated with increased anxiety-related behavior and impaired contextual fear memory\textsuperscript{159}. Deficiencies of the SHANK3 gene, either by mutations or copy number alterations, appear to underlie the neurological abnormalities associated with autism or 22q13 deletion syndrome\textsuperscript{18,19,155}. Interestingly, a single copy gain of the SHANK3 locus was observed in a patient with Asperger syndrome, a mild form of autism\textsuperscript{19}. Coupled with clinical studies showing that haploinsufficiency of SHANK3 underlies the neurological defects observed in patients with 22q13 deletion syndrome\textsuperscript{18,160}, these results suggest that the level of expression of SHANK3 may be critical for normal brain development. However, how SHANK3 expression level is regulated is unknown. A recent study has indicated that SHANK3 expression levels are influenced by DNA methylation within the gene in primary cultured neurons\textsuperscript{161}. Determining how intragenic DNA methylation may influence SHANK3 expression is one of the objectives of this thesis.
1.3 Objectives

The initial objective of this dissertation was to develop and validate a novel method to evaluate DNA methylation genome-wide. We applied this method to determine the locations and extent of normal tissue-specific DNA methylation. In doing so, we discovered that tissue-specific intragenic DNA methylation patterns of certain genes are evolutionarily conserved. The expression level of one such gene, *SHANK3*, is critical for normal neurological function\textsuperscript{18,19}, however how the gene is transcriptionally regulated and whether DNA methylation is involved in transcription were unclear. Thus, the additional objectives of this dissertation have been to determine the functional significance for intragenic DNA methylation, using *SHANK3* as a model, and by doing so gain insight into how *SHANK3* might be transcriptionally regulated.
2. Chapter II: A novel genome-wide DNA methylation technology

2.1 Summary

CpG islands are present in one-half of all human and mouse genes and typically overlap with promoters and/or exons. We have developed a method for high-resolution analyses of the methylation status of CpG islands genome-wide, using arrays of BAC clones and the methylation sensitive restriction enzyme \textit{NotI}. We demonstrate the accuracy and specificity of the method, and by computationally mapping all \textit{NotI} sites, methylation events can be defined with single nucleotide precision throughout the genome. We also demonstrate the unique expandability of the array method using a different methylation sensitive restriction enzyme, \textit{BssHII}. We identified and validated new CpG island loci that are methylated in a tissue-specific manner in normal human tissues. The methylation status of the CpG islands is associated with gene expression for several genes, including \textit{SHANK3} which encodes a structural protein in neuronal postsynaptic densities, defects of which appear to underlie human 22q13 deletion syndrome. Furthermore, these patterns for \textit{SHANK3} are conserved in mice and rats. The known function of \textit{SHANK3}, its association with a disease phenotype, and the persistence of \textit{SHANK3} tissue-specific CpG island methylation and gene expression patterns over 75 million years suggests their functional importance.
2.2 Introduction

DNA methylation is essential for a properly functioning genome\textsuperscript{162-167}. DNA methylation involves transfer of a methyl-group to cytosine in a CpG dinucleotide via DNA methyltransferases that create \textit{de novo} (DNMT3a, 3b) or maintain (DNMT1) methylation patterns. CpGs are five times more abundant in CpG islands, relative to the majority of the genome\textsuperscript{129,168,169}. CpG islands are approximately 200 bp to several kb in length and nearly always encompass gene promoters and/or exons\textsuperscript{169-171}. CpG islands associated with imprinted genes\textsuperscript{172}, genes on the inactive X chromosome in females\textsuperscript{173}, and some tissue-specific genes can be differentially methylated at either 5' or intragenic CpG islands\textsuperscript{174-178}, whereas many CpG islands are thought to be unmethylated in all normal tissues. However, the methylation status of the vast majority of the estimated 30,000 CpG islands\textsuperscript{179-181} has not yet been completely assessed in multiple tissues, and their methylation status therefore remains poorly defined.

Array comparative genome hybridization is a quantitative, high-resolution method for measuring copy number across the genome\textsuperscript{182-185}. The arrays consist of thousands of BAC clones distributed across the genome, and are capable of accurately distinguishing single copy gain and loss in cancer, as well as copy number polymorphisms in normal tissue\textsuperscript{186}. Since many of the BAC clones contain CpG islands, we propose that these arrays are ideally suited for large-scale CpG island methylation analyses. Here we designed a method to use methylation-sensitive restriction enzymes and BAC clone arrays to determine the methylation status of CpG islands genome-wide in different tissues. The method should be
applicable to any organism for which BAC arrays have been developed, and to normal and disease tissue. The method is also uniquely expandable to a multitude of methylation-sensitive restriction enzyme sites and updated tiling path arrays. We further demonstrate the utility of the array by identifying new CpG island loci that are subject to tissue-specific methylation and gene expression in human tissue. Using a comparative epigenomics approach, we also show that the differential methylation and expression for one of the genes is conserved in rodents.
2.3 Methods

2.3.1 Tissues and primary cells

Normal brain samples were obtained from the Neurosurgery Tissue Bank at the University of California, San Francisco. Samples were obtained with informed consent and their use was approved by the Committee on Human Research at UCSF. Adult peripheral blood lymphocytes (PBL) were collected from healthy volunteers. Normal human primary adult keratinocytes and normal human fetal astrocytes were purchased from Cambrex and cultured for less than 3 passages. Mouse tissues were obtained from the FVBN, C57BL/6J and NIH/Ola strains. Mouse whole brain, purified primary astrocytes (postnatal day 10) and PBL were derived from normal FVBN. An additional brain sample from C57BL/6J was also analyzed and showed similar results to FVBN. Keratinocytes were isolated from the skin of normal newborn NIH/Ola pups by physical separation of the epidermal layer from whole skin. Rat brain and PBL were obtained from Sprague-Dawley rats.

2.3.2 Methylation array analysis

Genomic DNA (2 µg) from normal tissue samples was blocked in a 10 µl reaction by the addition of nucleotide analogues (αS-dGTP, αS-dCTP, ddATP, ddTTP) with DNA polymerase followed by heat inactivation (65°C, 30 min). The DNA was digested with 20U NotI (or BssHII) and 20U EcoRV for 4 hours, 37°C. The cleaved NotI (or BssHII) ends were filled in with biotin-dGTP and biotin-dCTP (NEN) using sequenase V2.0 (USB). The biotinylated fragments were separated by incubation with magnetic streptavidin beads (M-280, Dynal) at
25°C, 30 min. The supernatant was removed and the beads were then washed three times with B&W buffer (10mM Tris-HCl, pH 7.5, 1mM EDTA, 2M NaCl). The DNA fragments were released from the beads in 98% formamide containing 10mM EDTA at 65°C for 5 min. Amino-allyl-dUTP (Sigma) was incorporated into the purified fragments using the Bioprime DNA Labeling System (Life Technologies, Inc.). Free nucleotides were removed using PCR purification columns (Qiagen). Cy5 and Cy3 dye esters were coupled to amino-allyl-dUTP-labeled test and reference samples, respectively. Washing of slides was performed according to the established protocols for array CGH.\textsuperscript{182,183,185}

Images were quantified and summarized using UCSF SPOT and SPROC software with default parameters.\textsuperscript{188} The clones for which fewer than 2 replicate spots were retained or for which standard deviation of replicate spots exceeded 0.1 were removed from further analysis. In addition, the clones present in fewer than 50% of the samples were filtered out. The quality of the samples was assessed using tools available in R/Bioconductor freely available package “marray” \texttt{http://www.bioconductor.org/repository/devel/vignette/marray.pdf}. The conservative sample-specific threshold was calculated as described in the text.

### 2.3.3 Predicting fragment hybridization computationally

The UCSC July 2003 freeze of the human genome was used for our \textit{in silico} analysis. We downloaded the whole-chromosome assemblies, the BAC end sequences, CpG-islands, and NIH SNP annotations and analyzed them using a custom Perl program on a Windows XP workstation. The program simulates the array experiment and generates a list of restriction fragments using the enzyme
pairs across the genome and the BAC clones they overlap. Fragments residing on CpG islands are also identified and restriction sites residing on recognized polymorphisms are flagged. The program can be adjusted for any restriction enzyme combination and can process the entire human genome in under 30 minutes. We predict that 870 BACs on the existing array contain one or more NotI sites, and approximately 1700 contain BssHII sites. The Perl script is available from J. Costello, UCSF.

2.3.4 Methylation of genomic DNA in vitro

Two µg of total genomic DNA from normal human PBL was digested with 4U EcoRV (NEB) for 2 hours at 37°C followed by heat inactivation for 20 min at 80°C and phenol-chloroform-isoamyl (25:24:1) extraction. DNA was then precipitated with 3M NaOAc and ethanol, and the pellet was resuspended in 20 µl of water. The digested DNA was treated with 100U M.SssI (NEB) for 4 hours at 37°C according to the manufacturer’s instructions. The enzyme was heat inactivated for 20 min at 65°C, and DNA was purified and processed for array hybridization. The M.SssI-treated DNA was used as a test sample along with untreated PBL DNA as a reference. To ensure that appropriate labeling and hybridization of NotI selected fragments could occur in M.SssI-treated DNA, 5 ng of untreated DNA from a human BAC clone library (clones RP11-15M23 and RP11-23D23) was added to the M.SssI-methylated PBL DNA prior to NotI digestion and fragment selection. These NotI site-containing clones were verified by end-sequencing and correspond to chr7:40858466-40697704 (RP11-15M23) and chr7:870718-698268 (RP11-23D23) (July 2003 assembly, UCSC).
2.3.5 Bisulfite treatment, PCR and sequencing

Total genomic DNA was treated with sodium bisulfite as described\textsuperscript{190}. PCR was then carried out using primers listed in Table 2 (Appendix) and products were cloned into pCR2.1/TOPO (Invitrogen, Inc.). Single colonies were selected and inserts were sequenced using the ABI 3700 automated DNA sequencer.

2.3.6 Real-time RT-PCR

Expression of all genes analyzed from mouse, rat and human samples were measured by real-time RT-PCR using the Opticon2 Continuous Fluorescence Detector (MJ Research). Primers and probes were purchased from IDT and Applied Biosystems (Table 2; Appendix).
2.4 Results

We designed a method to use arrays of bacterial artificial chromosome (BAC) clones\textsuperscript{182,183,185} and methylation sensitive enzymes for analysis of CpG island methylation status (Fig. 3A). To assess the methylation status of CpG islands, we used \textit{NotI} sites as indicators, since its endonuclease activity is inhibited by methylation and 94\% of \textit{NotI} sites are within CpG islands\textsuperscript{191,192}. Using a custom Perl script, we estimate 2,500 \textit{NotI} sites can be analyzed using this prototype methylation array.

To prove that the signal intensities across the array are indeed from unmethylated \textit{NotI} fragments, we compared a self-self hybridization to a hybridization of peripheral blood lymphocyte (PBL) DNA that was first methylated \textit{in vitro} and cohybridized with untreated PBL DNA. Following \textit{in vitro} methylation, the sample was spiked with DNA from two unmethylated BAC clones which contain \textit{NotI} sites. The methylated genomic DNA-BAC mixture was then prepared as described in Fig. 3A and labeled with Cy3. For the self-self hybridization, equivalent amounts of Cy3 (red) and Cy5 (green) signal resulted in a yellow signal (Fig. 3B). In contrast, the \textit{in vitro} methylation of PBL DNA significantly reduced the Cy3 (red) signal for most spots, while the Cy5 (green) signal from untreated PBL DNA was maintained (Fig. 3C). The unmethylated \textit{NotI} fragments copurified from the two admixed BAC clones hybridized to their corresponding spots specifically (Fig. 3C). Consistent with the array image, the average log\textsubscript{2} ratio of \textit{NotI} site-containing BACs in the control self-self hybridization was -0.07, indicating equivalent signal from test and reference DNA.
In contrast, the average log₂ ratio of these BACs after in vitro methylation was significantly decreased (average log₂ ratio of −0.82, p<0.01, individual examples shown in Fig. 3D). As expected, the log₂ ratios of the purified BACs added to the Cy3-channel were significantly increased compared to that in the control self-self hybridization (Fig. 3D). This experiment was repeated and showed very similar results (Fig. 3D). Hybridization signals on the array are therefore predominately generated from unmethylated NotI fragments.

<table>
<thead>
<tr>
<th>Position</th>
<th>Clone ID</th>
<th>PBL</th>
<th>PBL-m + BAC</th>
<th>PBL</th>
<th>PBL-m + BAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>G14</td>
<td>RP11-15M23</td>
<td>-0.10</td>
<td>4.19</td>
<td>-0.20</td>
<td>3.84</td>
</tr>
<tr>
<td>L19</td>
<td>RP11-23D23</td>
<td>-0.22</td>
<td>2.67</td>
<td>0.08</td>
<td>3.64</td>
</tr>
<tr>
<td>B17</td>
<td>RP11-72M14</td>
<td>0.01</td>
<td>-1.98</td>
<td>-0.07</td>
<td>-2.40</td>
</tr>
<tr>
<td>F11</td>
<td>CTD-2100I14</td>
<td>-0.06</td>
<td>-1.44</td>
<td>-0.32</td>
<td>-1.78</td>
</tr>
<tr>
<td>F3</td>
<td>RP11-55M16</td>
<td>-0.21</td>
<td>-2.13</td>
<td>-0.14</td>
<td>-2.31</td>
</tr>
<tr>
<td>F9</td>
<td>RP11-192A12</td>
<td>-0.15</td>
<td>-1.42</td>
<td>-0.39</td>
<td>-1.81</td>
</tr>
<tr>
<td>G7</td>
<td>RP11-29P24</td>
<td>-0.09</td>
<td>-1.37</td>
<td>0.28</td>
<td>-1.49</td>
</tr>
<tr>
<td>H5</td>
<td>RP11-206I1</td>
<td>-0.01</td>
<td>-1.53</td>
<td>0.00</td>
<td>-1.62</td>
</tr>
<tr>
<td>I7</td>
<td>RP11-105N15</td>
<td>-0.19</td>
<td>-1.21</td>
<td>-0.09</td>
<td>-1.63</td>
</tr>
<tr>
<td>I9</td>
<td>CTD-2027I18</td>
<td>0.00</td>
<td>-1.74</td>
<td>-0.18</td>
<td>-2.64</td>
</tr>
<tr>
<td>K8</td>
<td>RMC20P117</td>
<td>0.06</td>
<td>-1.84</td>
<td>-0.15</td>
<td>-2.20</td>
</tr>
<tr>
<td>L15</td>
<td>RP11-282J24</td>
<td>-0.17</td>
<td>-1.78</td>
<td>-0.09</td>
<td>-2.01</td>
</tr>
</tbody>
</table>

Figure 3: Detection of methylation differences using a BAC clone array.

(A) The methylation array procedure. Genomic DNA is digested with EcoRV and a methylation-sensitive enzyme, such as NotI. The cleaved ends of unmethylated NotI fragments are filled in with biotin-dNTPs and isolated on streptavidin-coated magnetic beads, while the methylated fragments and most non-CpG-island DNA are eliminated. The purified fragments are hybridized to arrays of BAC clones printed in triplicate. The BAC arrays consist of 2413 unique BAC clones with an average interval of ~1 Mb across the human genome. (B) Hybridization signals on the array are derived from unmethylated DNA. Two array sections from a control PBL-PBL hybridization demonstrating equivalent amounts of Cy3 (red)
and Cy5 (green) hybridization signal. BACs without \textit{NotI} sites had little or no signal intensity, whereas most \textit{NotI} site-containing BACs exhibited signal. (C) The same array sections as in B from untreated PBL DNA labeled with Cy5 (green) and cohybridized with \textit{in vitro} methylated PBL DNA spiked with 5 ng of two unmethylated BAC DNAs (RP11-15M23 at position G14, and RP11-23D23 at L19, in PBL both of these sites are normally methylated; data not shown) and then labeled with Cy3 (red). The \textit{in vitro} methylation significantly reduces the Cy3 signal for most spots, while the Cy5 signal from untreated PBL DNA is maintained. (D) Average log$_{2}$ ratios of 10 \textit{NotI} site-containing BACs with the highest intensity in B and C (Experiment 1). This \textit{M.SssI} array experiment was repeated and showed very similar results (Experiment 2).

In order to evaluate the quality of the array data, we looked for the presence of spatial effects of a considerable size and for intensity dependence of the log$_{2}$ ratios. The log$_{2}$ ratios were normalized using global median shift to ease the across-array comparisons. Intensity dependence was evaluated by considering the shape of the lowess-smoothers within each subarray (Fig. 4). The geometric effects were evaluated by comparing print-tip specific log$_{2}$ ratio distributions within each array (Fig. 4). The geometric aspect of the array was judged to be satisfactory if the variability within individual subarrays was approximately constant and under 0.2, and if the inter-quartile range of each print-tip distribution contained zero. We accepted the array if none of the lowess lines showed considerable departure from a horizontal line and there was not significant geometric dependence. Taken together, Figure 4 demonstrates lack of an intensity-dependent effect and lack of strong spatial print-specific effects in the arrays.
We determined a conservative sample-specific cut-off for declaring a clone to be differentially methylated. For each sample we calculated the median absolute deviation (MAD) computed over the autosomal clones present in at least 50% of the samples. The second iteration of the MAD calculation was performed using only those clones that were within 2.5 MAD from zero. The conservative threshold for calling methylation differences was computed as the maximum of either 3
times MAD for a given sample or 0.5. Among repeated self-self hybridizations of acceptable quality, between 0 and 5 autosomal clones have exceeded the resulting threshold with the median of 1 (Fig. 5A). This is well within the typical range for the BAC arrays used in standard array CGH \cite{185}.

\textbf{Figure 5: Detection of methylation differences between cell types using NotI or BssHII.}

\textit{(A)} Log₂ ratios from an array hybridization of \textit{NotI} fragments independently isolated from two aliquots of a male PBL DNA (red dots). \textit{(B)} Log₂ ratios from the self-self hybridization as in \textit{A}, and a comparative hybridization between normal human astrocytes and PBL DNA (blue dots). \textit{(C)} Log₂ ratios from an array hybridization of \textit{BssHII} fragments independently isolated from two aliquots of a male PBL DNA (red dots). \textit{(D)} Log₂ ratios from the self-self hybridization of \textit{BssHII} fragments, and a comparative hybridization between normal human brain and PBL DNA (blue dots). Data are plotted as the mean log₂ ratio of triplicate spots for each clone, and only those log₂ ratios with a standard deviation less than 0.1 were considered valid. The BACs are ordered from left to right by their position in the genome starting at chromosome 1p (See also Fig. 6 for graphs demonstrating \textit{BssHII} array quality).
To test for tissue and cell type-specific methylation of CpG islands, a self-self hybridization (PBL) and a comparative hybridization (astrocyte versus PBL) were conducted (Fig. 5). In the self-self hybridization, the distribution of the log₂ ratios was narrow (Fig. 5A). In contrast, comparison of astrocyte DNA versus PBL DNA showed a significant subset of clones with log₂ ratios outside the sample threshold (Fig. 5B). These results suggest that there are differences in the methylation status of a significant number of NotI sites, possibly CpG islands, between the astrocytes and PBL. To extend these results to a larger set of CpG islands, we developed a protocol for using BssHII as the methylation-sensitive enzyme. We observed specific hybridization to BAC clones containing BssHII sites, many of which do not contain NotI sites. The results of cohybridizations of BssHII fragments from PBL vs. PBL and brain vs. PBL yielded overall results similar to the arrays hybridized with NotI fragments (Fig. 5C,D and Fig. 6). Together, the NotI and BssHII protocols will allow us to interrogate different CpG islands within a single BAC, or different CpGs within a single CpG island, depending on the restriction site distribution.
Figure 6: The log₂ ratios of cohybridizations of BssHII fragments do not exhibit intensity-dependent effects nor geographic location-specific effects.

M/A plots (top panels) and subarray plots (M/print tip, lower panels) of self-self hybridizations (plot 1 of each row, PBL versus PBL), and normal brain versus PBL (2 different individual brain DNAs are shown). The top plot of each pair shows the relationship between total intensity (log₂ of average of Cy3 and Cy5 signals, A-value) and log₂ ratio of the two signals (M-value). The M-values were shifted by their global median to ease the effect-size interpretation. The dotted horizontal curves are the lowess-fitted print-tip specific regressions. The bottom plot of each pair compares distributions of the M-values across the subarrays and displays the size of the print-tip specific effects, if any. Taken together the two figures demonstrate lack of an intensity-dependent effect and lack of strong spatial print-specific effects in the arrays. Note that variability across the subarrays is very similar. Thus, we employed global median normalization in our subsequent analysis of BssHII arrays.

To confirm that these differences were generated by changes in methylation, we performed 7 to 12 bisulfite-sequencing reactions for each of 25 data points on the arrays for loci that appear to be differentially methylated in a small panel of human tissues and primary cell cultures, including brain, fetal astrocytes, keratinocytes and PBL. As shown in Fig. 7 and Fig. 9, we found a clear association between the log₂ ratios from the array and the degree of methylation in
the \textit{Not}I sites or \textit{BssH}II site and adjacent CpGs. The \textit{Not}I site of RP11-218N6 is adjacent to a CpG island at the 5’ end of a gene of unknown function, C2orf32 (see Fig. 8 for maps of each locus). Bisulfite sequencing of the \textit{Not}I site and adjacent CpGs in each direction showed tissue-specific methylation (Fig. 7B). To determine if the differential methylation is associated with gene expression status, C2orf32 mRNA levels were assessed by real-time RT-PCR on the same set of tissues (Fig. 7B). The C2orf32 mRNA was undetectable in PBL and keratinocytes, where this region was methylated. The gene was expressed moderately to highly in astrocytes and brain, respectively, where the region remained unmethylated. Similar results were observed for another CpG rich region within the \textit{PRK}CZ gene (Fig. 7C), which contains two \textit{Not}I sites. A candidate from the \textit{BssH}II arrays, BAC clone RP11-96D22 exhibited a consistently positive log$_2$ ratio >0.5 in uncultured brain samples relative to PBL, suggesting more methylation in PBL relative to brain at this \textit{BssH}II site (Fig. 7D). The \textit{BssH}II site is within an intragenic CpG island associated with \textit{ARHG}EF17, a gene encoding a Rho guanine nucleotide exchange factor. Consistent with the array results, the \textit{ARHG}EF17 CpG island was unmethylated at nearly all CpGs (in two contiguous regions of bisulfite analysis) in brain where it is expressed and is densely methylated in PBL which do not express \textit{ARHG}EF17 (Fig. 7D). The CpG island is also unmethylated in astrocytes and keratinocytes which express low levels of \textit{ARHG}EF17. These data provide validation for the array method in detecting methylation differences. The results also identify new examples of tissue-specific methylation that are associated with specific gene expression states.
Figure 7: Validation of the array results by bisulfite-sequencing and gene expression analysis.

Log2 ratios from the array correlate well with the methylation status of the NotI (A-C) and BssHII (D) sites, and adjacent CpGs. For diagrams of each locus, see Figure 8. HUGO designations for the associated genes are given in parenthesis next to the BAC names. (A) Methylation analysis in and around the NotI site of BAC clone RP11-60M20 located on chromosome 2q14.2. (B) Methylation analysis in and around the NotI of BAC clone RP11-218N6 located on chromosome 2p14, and quantitative gene expression analysis of the associated gene C2orf32. (C) Methylation analysis in and around the NotI of BAC clone RP11-82D16 located on chromosome 1p36.33, and quantitative gene expression analysis of the associated gene PRKCZ. We assessed methylation in both NotI sites in this local region by performing two PCRs on contiguous regions (I and II). (D) Methylation analysis in and around the BssHII site of BAC clone RP11-96D22 located on chromosome 11q13.4, and quantitative gene expression analysis of the associated gene ARHGEF17. We assessed methylation across 53 CpGs by performing two PCRs on contiguous regions (I and II). Each row of circles represents a summary of the methylation status of cytosines of CpGs after bisulfite treatment and sequencing (n, number of independent clones sequenced). Open
circles indicate that 0-33% clones were methylated at the CpG; gray circles, 34-66% methylated; filled black circles, 67-100% methylated. The total methylation % was calculated considering the methylation status of all CpGs in all sequence reactions for the given sample. The relative mRNA levels of each gene relative to control GUSB were determined by real-time RT-PCR, performed in triplicate (mean +/- sd). ND, not detectable after 45 PCR cycles. The expression analysis was repeated two additional times and showed similar results. NA, not analyzed.

Figure 8: Maps of the chromosomal loci corresponding to BAC clones and their loci that exhibit tissue-specific methylation.

A-D here corresponds to Figure 7 A-D. In A, a gene or CpG island corresponding to this locus has not been reported. In B-D, the arrows indicate the putative transcription start site of the associated gene. Note that for C and D, the CpG island is intragenic, whereas it is 5’ in B.

Another candidate for differential CpG island methylation, CTA-799F10, exhibited high log2 ratios in brain, astrocytes and keratinocytes relative to PBL, suggesting there is more methylation in PBL. We identified a NotI site-containing CpG island in this region that is within the SHANK3 gene160,194 (Fig. 9A), which encodes a structural protein in neuronal postsynaptic densities. Haploinsufficiency
via deletion or intragenic translocation of SHANK3 is thought to underlie the severe neurologic symptoms of 22q13 deletion syndrome in humans\textsuperscript{160,195,196}. A chromosomal breakpoint in one patient occurs precisely within this CpG island\textsuperscript{18,160}. Consistent with the array results, bisulfite-sequencing demonstrated that this CpG island is extensively methylated in PBL, whereas methylation in other tissues was low or absent (Fig. 9B). SHANK3 is expressed in brain tissue where it is primarily unmethylated, while it is not detectable in PBL where the CpG island is methylated. Thus, the unmethylated state appears to be permissive for SHANK3 expression while methylation of the CpG island appears repressive. The methylation and expression status of SHANK3, and the association of defects in SHANK3 with a severe disease phenotype in humans suggest that the tissue-specific methylation status of this CpG island is of some functional importance.
Identification of gene regulatory elements is a major goal of current research and has been significantly advanced at the DNA sequence level through comparative genomics. Here we reasoned that a comparative epigenomics
approach could help determine if the differentially methylated CpG island of *SHANK3* is conserved between divergent species and whether the methylation status has a functional impact on gene expression. We therefore assessed the methylation status and mRNA levels of *SHANK3* in analogous mouse tissues (Fig. 9C,D). *SHANK3* was densely methylated in PBL, while less or no methylation was observed in brain, astrocytes and keratinocytes, very similar to the pattern observed in corresponding human tissues. The methylation and expression pattern in PBL and brain was very similar in both C57BL/6J and FVBN mice. In rat PBL, *SHANK3* was similarly densely methylated, while in rat brain it was partially methylated (Fig. 10). The tissue-specific expression of *SHANK3* was also conserved in the mouse (Fig. 9D) and rat (Fig. 10). The apparent conservation of tissue-specific CpG island methylation and gene expression patterns for *SHANK3* over approximately 75 million years suggests that they are of some functional importance.

**A**  Chr. 7q34

![Diagram showing methylation and expression pattern](image)

**B**  Rat *SHANK3*

<table>
<thead>
<tr>
<th>Tissue/Cell Type</th>
<th>n</th>
<th>I (5)</th>
<th>II (10)</th>
<th>III (15)</th>
<th>IV (20)</th>
<th>V (25)</th>
<th>VI (30)</th>
<th>Total Methylation %</th>
<th>Expression %</th>
</tr>
</thead>
<tbody>
<tr>
<td>brain-1 male</td>
<td>9</td>
<td>96.7%</td>
<td>71.6%</td>
<td>64.0%</td>
<td></td>
<td></td>
<td></td>
<td>77.1</td>
<td>12.9 ± 1.0</td>
</tr>
<tr>
<td>brain-1 female</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.2</td>
<td>24.9 ± 1.5</td>
</tr>
<tr>
<td>brain-2 male</td>
<td>10</td>
<td>92.0%</td>
<td>65.6%</td>
<td>74.3%</td>
<td></td>
<td></td>
<td></td>
<td>84.0</td>
<td>32.7 ± 4.3</td>
</tr>
<tr>
<td>brain-2 female</td>
<td>9</td>
<td>90.0%</td>
<td>87.8%</td>
<td>69.2%</td>
<td></td>
<td></td>
<td></td>
<td>81.4</td>
<td>39.2 ± 6.2</td>
</tr>
<tr>
<td>PBL male</td>
<td>8</td>
<td>100.0%</td>
<td>97.1%</td>
<td>96.0%</td>
<td></td>
<td></td>
<td></td>
<td>96.0</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td>PBL female</td>
<td>8</td>
<td>100.0%</td>
<td>100.0%</td>
<td>96.0%</td>
<td></td>
<td></td>
<td></td>
<td>99.2</td>
<td>0.0 ± 0.0</td>
</tr>
</tbody>
</table>
Figure 10: The tissue-specific CpG island methylation and gene expression of SHANK3 is conserved in rat brain and PBL.

DNA and RNA were obtained from brain samples (brain 1 is mixed gray and white matter, brain 2 is predominantly gray matter) and PBL from one male and one female rat. Each row of circles represents a summary of the methylation status of cytosines of CpGs after bisulfite treatment and sequencing (n, number of independent clones sequenced). Open circles indicate that 0-33% clones were methylated at the CpG; grey circles, 34-66% methylated; filled black circles, 67-100% methylated. The region was artificially divided into three equal zones. Zone methylation as well as total % methylation levels are given as the percent of methylated CpGs over the total number of CpGs analyzed. While both the brain and PBL were densely methylated in the 5’ portion of the CpG island, the 3’ zone was densely methylated in PBL, but was partially methylated in rat brain, as it is in the mouse and human. Though all three organisms exhibit partial methylation in the brain, the level of partial methylation is variable. The relative mRNA levels of rat SHANK3 relative to control GAPDH were determined by real-time RT-PCR, performed in triplicate (mean +/- sd). Relative SHANK3 expression levels in rat brain and PBL are similar to those in human and mouse. This experiment was repeated and showed similar results.
2.5 Discussion & Conclusion

Our novel method allows methylation analysis across the genome with single nucleotide precision. Since even generic BAC arrays can be used for the methylation array method, it is applicable to all species\textsuperscript{184}, and can provide comprehensive methylation analysis using complete tiling path arrays\textsuperscript{187} and alternative methylation-sensitive enzyme combinations. As with RLGS and other genome scale methylation screens\textsuperscript{197,198}, one possible limitation of our BAC array method is the presence of SNPs which might create or eliminate a restriction site, leading to a false positive or negative. To address this, we have written a Perl script that maps all known SNPs to the restriction sites being tested on the array. Thus, we will know \textit{a priori} if a common polymorphism might influence the array results. This limitation will be irrelevant when comparing multiple samples from one individual or from a panel of genetically similar organisms.

We applied the logic of comparative genomics to our epigenomic data from the array and discovered that the tissue-specific pattern of CpG island methylation of \textit{SHANK3} is very similar in mouse, human and rat tissues. Two different mouse strains also yielded similar methylation and gene expression results, further supporting our conclusions. Conservation suggests function, which can be tested by further in-depth understanding of other \textit{cis}- and \textit{trans}-acting factors that regulate the \textit{SHANK3} gene (\textit{Chapter III}). Certainly the presence or absence of relevant transcription factors, modifications to other portions of the \textit{SHANK3} gene or the histones bound to it, as well as structural changes in positioning of \textit{cis}-acting elements within each cell type might also play a role. We also tested \textit{ARHGEF17}
in mice and found a lack of conservation of the CpG island methylation (data not shown). Together, this data suggests that \textit{SHANK3} methylation patterns may be evolutionarily conserved in a gene-specific manner, rather than simply being one part of mechanism that is entirely conserved. This also does not necessarily imply that \textit{ARHGEF17} methylation is non-functional in humans, particularly if the pattern is conserved across a large population. Conservation of protein (histones) methylation and acetylation patterns between mice and humans also has been reported recently\textsuperscript{199}.

Despite the general knowledge that methylation patterns are tissue-specific, the genomic location and function of most of these sites is unknown. Early studies suggested substantial differences in the total 5-methylcytosine content of different tissues\textsuperscript{200}. Our results show that the intragenic CpG islands of \textit{SHANK3} and \textit{ARHGEF17} exhibit tissue specific methylation and gene expression. The mechanism by which intragenic methylation might modulate levels of gene expression is unclear, though a recent report suggests that intragenic methylation may result in exclusion of RNA polymerase II, reducing transcription elongation efficiency\textsuperscript{201}. Our results extend recent reports from the human epigenome project\textsuperscript{176}, RLGS\textsuperscript{175} and elsewhere\textsuperscript{178,202} that CpG islands can be methylated in specific tissues. Nevertheless, a substantial fraction of the genome remains to be explored in any tissue, and very few genes have been studied for methylation in more than a hand full of normal tissues, or specific cell types\textsuperscript{178}. The method we report here should be very useful for further identification of such sites, including 5’ and intragenic CpG islands, as well as non-island sequences. By determining
the distribution of other methylation sensitive restriction enzymes in relation to genes, CpG islands, or repeat sequences, as we have for NotI and BssHII, we can examine new sets of CpG islands or bias more towards non-island sequences. Our approach is applicable to many important questions including normal epigenetic variation in the population, epigenetic reprogramming in cloned organisms, tissue specific methylation, imprinting, and aberrant methylation in cancer and other diseases.

**NOTE:** A version of this Chapter has been published⁷.
3. Chapter III: Evolutionary conservation & function of intragenic DNA methylation

3.1 Summary

Although DNA methylation is commonly found in gene bodies, its biological significance is unclear. Using the autism gene SHANK3 as a model, we show that intragenic methylation is tissue- and cell type-specific, but also differs significantly within a single cell type from distinct brain regions. The intragenic methylation is also gene region–specific and evolutionarily conserved. Because DNA methylation is known to influence the activity of 5’-promoter sequences, we searched for genetic and epigenetic evidence of promoter activity embedded within SHANK3. Using this integrated and cross-species approach, we identified two intragenic regions that have promoter activity that is blocked by methylation in vitro, that are differentially methylated in vivo, and which drive transcription of novel and potentially protein-coding SHANK3 transcripts. The constitutively unmethylated 5’-CpG island promoter and the differentially methylated intragenic promoters of SHANK3 are coordinately regulated, but by distinct epigenetic mechanisms. In 15/22 additional genes (68.2%) having tissue-specific intragenic methylation, the differentially methylated sites coincided precisely with multiple features of promoters. Altogether, these results support an evolutionarily conserved, cell context-specific role for intragenic DNA methylation in regulating the activity of alternate promoters. We suggest that intragenic DNA methylation has a greater role in gene regulation than methylation of 5’-CpG island promoters.
3.2 Introduction

DNA methylation is essential to embryonic development\textsuperscript{3}, differentiation\textsuperscript{83}, cell cycle control\textsuperscript{84,85}, and maintenance of genome stability\textsuperscript{132-134}, playing critical roles in maintaining transcriptional silencing of genes on the inactive X-chromosome, imprinted genes, and a few tissue-specific and/or developmentally-regulated genes from their 5’-promoters\textsuperscript{8,9,11,80,115,203-213}. However, the majority of 5’-promoters of somatic genes, including those associated with CpG islands, is thought to be unmethylated in all normal tissues. This suggests that, although important for normal cellular functions, DNA methylation plays a limited role in gene regulation\textsuperscript{4}.

In contrast to 5’-promoters, DNA methylation is common within gene bodies\textsuperscript{8,10,13,14,115,214}. Such intragenic methylation is found in organisms that diverged as much as 1.6 billion years ago, implying that it has a common ancestral function\textsuperscript{124}. Proposed functions for DNA methylation outside of gene promoters (e.g. intragenic regions) include inhibition of cryptic, unproductive initiation of transcription\textsuperscript{14,215} and protection against deleterious effects of mobile genetic elements\textsuperscript{12,86,131-133}. Because these functions are important for all tissue/cell types, they do not readily explain the purpose for intragenic DNA methylation that is tissue- or cell type-specific, which has now been widely observed in mammals\textsuperscript{5-8,10,11,16,115,216}. Further, at certain loci, the tissue-specificity of intragenic DNA methylation is evolutionarily conserved\textsuperscript{7,8}. This suggests that tissue-specific intragenic methylation has a role distinct from those functions that require constitutive methylation.
From early correlative studies, conflicting views arose about how intragenic DNA methylation might influence transcription, or indeed whether intragenic methylation has a function\textsuperscript{217}. These studies were based largely on correlations between methylation and gene expression across different tissue types\textsuperscript{218-221}, the interpretation of which is confounded by the fact that there are multiple factors that regulate transcription which also differ among different cellular contexts and cell types. A recent study that controlled for cell type has demonstrated that intragenic DNA methylation of a chromosomally integrated transgene in cancer cells diminished the levels of expression by reducing the efficiency of transcriptional elongation\textsuperscript{135}. Based in part on this important finding, correlative data from endogenous genes in plants has been similarly interpreted to indicate a role for intragenic methylation in influencing transcriptional elongation, or reducing transcriptional noise\textsuperscript{14}. However, these studies do not consider where within the gene that expression is measured, which is important to determine how intragenic methylation and gene expression may be related.

Intragenic DNA methylation also could influence gene expression by affecting the activity of promoters embedded within gene bodies, just as it does to promoters at the 5’-end of genes. Two observations support this hypothesis: (1) genome-wide mRNA ‘cap-trapping’ experiments in mammals suggest the existence of multiple sites of transcription initiation within the bodies of most protein-coding genes\textsuperscript{222-225}, implying that alternate internal promoters are common throughout the genome and (2) intragenic tissue-specific DNA methylation often coincides with non-coding regions that are highly conserved\textsuperscript{8}, implying that the
methylation might influence cis-acting regulatory elements. However, a functional relationship between intragenic DNA methylation and usage of alternate promoters embedded in gene bodies has not been established.

To determine if intragenic DNA methylation is functional, and how it might work, we investigated in-depth the evolutionarily conserved tissue-specific intragenic DNA methylation we previously reported within one intragenic CpG island in \textit{SHANK3}\textsuperscript{7}, the product of which functions as a master scaffolding protein in neurons\textsuperscript{156} (Fig. 2) and as a signal transduction protein in epithelial cells\textsuperscript{158}. In addition to its function, \textit{SHANK3} levels are important for normal brain development – loss or gain of one copy of \textit{SHANK3} underlies subsets of 22q13 deletion syndrome and autism\textsuperscript{18,19}. However, how \textit{SHANK3} is transcriptionally regulated remains unknown. By integrating comparative genomics and comparative epigenomics approaches, our study provides novel insight into the function and specificity of intragenic methylation in \textit{SHANK3} and likely many other genes.

3.3 Methods

3.3.1 Normal tissues and cultured primary cells

Normal human brain samples were provided from the Neurosurgery Tissue Bank at the University of California San Francisco (UCSF) and we collected adult peripheral blood lymphocytes (PBL) from healthy volunteers. All samples were obtained with informed consent, and their use was approved by the Committee on Human Research at UCSF. Normal human primary adult keratinocytes and normal human fetal astrocytes were purchased from Cambrex
and were cultured for fewer than three passages. Normal human ES cells (HSF6) were kindly provided by Mary Firpo while at UCSF. Mouse whole brain, cerebella, hippocampi, lung, pancreas, heart, PBL, and sperm were isolated from normal 8-week old C57BL/6J mice. Keratinocytes from the skin of normal newborn NIH/Ola pups were isolated by physical separation of the epidermal layer from whole skin. In addition to adult stages, brain and lung tissues were derived from mice at pre- and post-natal developmental time points where indicated in the text. Dnmt1 hypomorphic mice: 129/Sv Dnmt1R hypomorphic mice (Dnmt1<sup>R/+</sup>) were crossed to 129/Sv Dnmt1N hypomorphic mice (Dnmt1<sup>N/+</sup>) to generate ‘NR’ (Dnmt1<sup>N/R</sup>) mice. 129/Sv Dnmt1<sup>+/+</sup> littermates were used as wild-type (‘WT’) controls. DNA was extracted from tail biopsies of 3-week old mice by the standard method of proteinase K digestion as previously described<sup>226</sup>. The genotypes of the Dnmt1 alleles were determined by a multiplex PCR assay as described<sup>227,228</sup>. These mice were then euthanized at 15 weeks of age and tissue samples were dissected and snap frozen in liquid nitrogen. Blood was collected for isolation of PBL. All animal experiments were conducted in compliance with a protocol approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Southern California. Astrocyte monolayers were derived from the postmortem cerebral cortex and hippocampus of postnatal day 7 C57BL/6J mice. The cerebral cortex dissection was preformed in such a way as to exclude all cells of the ventricular or subependymal region. Primary cultures were generated by mincing the tissue and incubating it with papain enzyme, after which cells were filtered through a 70 µm cell strainer. The
resulting cell suspensions were seeded on laminin coated plates in DMEM/F12 medium containing 10% (vol/vol) FCS supplemented with 2mM glutamine and allowed to grow to confluence. Primary cerebellar granule neural progenitors were isolated from postnatal day 5 C57BL/6J mice and neurons were derived from *in vitro* differentiation as described\textsuperscript{229}. Mouse ES cells (from C57BL/6J blastocysts) were kindly provided by Miguel Ramalho-Santos (UCSF). All tissue samples were homogenized for isolation of nucleic acids. All cultured cells were collected by trypsinization using 0.25% trypsin-EDTA and washed before cell lysis.

### 3.3.2 Isolation of nucleic acids

Genomic DNA was isolated as described\textsuperscript{226} by lysing the samples with 50 mM Tris-HCl (pH 8.0)-1 mM EDTA-0.5% sodium dodecyl sulfate-1 mg of proteinase K/ml overnight at 55°C followed by isopropanol precipitation. The DNA was pelleted, washed with 70% ethanol, and resuspended in 10 mM Tris-HCl-1 mM EDTA (pH 7.6). Total mRNA was isolated using Trizol (Invitrogen) as specified by the manufacturer.

### 3.3.3 Demethylation and deacetylation experiments

Primary mouse astrocytes were seeded at $1 \times 10^5$ cells per well of a six-well plate, incubated for 24 hours in Dulbecco's Modified Eagle's Medium (DMEM) high glucose with 10% serum, and then supplemented with fresh media containing 5-aza-2'-deoxycytidine (5azadC; 1 or 5 µM; Sigma-Aldrich) for 72 hours or trichostatin A (TSA; 100 ng/ml; Sigma-Aldrich) for 12 hours. For the combination treatment, 1 or 5 µM 5azadC was present for 72 hours and TSA was
added for the last 12 hours. The media containing drugs were changed every 24 hours.

3.3.4 Bisulfite treatment, PCR and sequencing

We treated total genomic DNA with sodium bisulfite for 4 hours as described\textsuperscript{230}, carried out PCR of the indicated \textit{SHANK3} regions using primers listed in Table 4 (Appendix), and cloned products into pCR2.1/TOPO (Invitrogen). We selected a specified number of individual colonies and sequenced inserts using the ABI 3700 automated DNA sequencer.

3.3.5 5’-Rapid amplification of cDNA ends

Total RNA from brain and lung of normal 8-week old C57BL/6J mice were used to amplify the 5’ end of \textit{SHANK3} mRNA with the Gene Racer kit (Invitrogen) based on the protocol supplied by the manufacturer. The mRNA was ligated to the Gene Racer oligo, reverse-transcribed, and amplified using \textit{SHANK3}-specific reverse primers R1 or R2 (Table 4; Appendix) with PfuUltra high-fidelity DNA polymerase (Stratagene) under the following 3-step ‘touch-down’ cycling parameters: (1) 5 cycles of 94°C for 30 sec, 72°C for 1 min, (2) 5 cycles of 94°C for 30 sec, 70°C for 1 min, (3) 30 cycles of 94°C for 30 sec, 62°C for 30 sec, and 72°C for 1 min, followed by 72°C for 10 min. The amplification products were gel purified, cloned into pCR4-TOPO (Invitrogen), and inserts were sequenced. The sequence data for the novel \textit{SHANK3} transcripts, 22t and 32t, have been deposited into the dbEST database and correspond to accession numbers GD253656 and GD253657, respectively. The unique first exon sequences of 22t and 32t correspond to chr15:89,354,730-89,355,012 and
chr15:89,363,250-89,363,804, respectively (Mouse July 2007 assembly; http://genome.ucsc.edu). Another transcript with a transcription start site downstream of 32t and lacking the full-length SHANK3 exon 18 was also identified by 5’-RACE (accession number: GD253658).

3.3.6 Reverse transcription, standard and real-time reverse transcription-PCR

Reverse transcription reactions were performed essentially as previously described\textsuperscript{231}. From mouse samples, we measured the expression of full-length SHANK3 and an internal control GusB with probe/primer assays Mm00498775_m1 and Mm00446953_m1 (Applied Biosystems), respectively, by real-time RT-PCR using the Opticon2 Continuous Fluorescence Detector (MJ Research). Expression levels of 22t and 32t were measured by RT-PCR using 18S and β-actin as internal controls for mouse and human samples, respectively. Primers and their corresponding PCR conditions are listed in Table 4 (Appendix).

3.3.7 Integration of promoter-associated features

For the SHANK3 locus (chr15:89,328,288-89,388,754; Mouse July 2007 assembly), we combined three distinct ‘features’ that are generally, yet not necessarily, associated with promoters. These features include ECRs, CAGE tags, and bivalent domains. Using the parameters described in the text, we identified ECRs throughout SHANK3 using ‘ECR Browser’\textsuperscript{232}: http://ecrbrowser.dcode.org. We obtained CAGE tag sequences along SHANK3 from a publicly available dataset\textsuperscript{222}: http://fantom31p.gsc.riken.jp/cage_analysis. ChIP-Seq data of H3K4me3 and H3K27me3 marks across SHANK3 in ES, NP,
and MEF cells were kindly provided by B. Bernstein from a dataset that is publicly accessible:\(^2\text{16}\): http://www.broad.mit.edu/seq_platform/chip. Because all of these features are sequence-based, we were able to precisely align them in relationship to the corresponding \textit{SHANK3} genomic sequence. The same ECR criteria and databases were used to analyze all three features within \(\sim\)500-bp of sequence flanking the \textit{NotI} site of each of the differentially methylated loci identified by RLGS\(^6\) that were confirmed.

### 3.3.8 Cloning of \textit{SHANK3} ECRs, transfection, and promoter-reporter assays

From mouse or human genomic DNA, selected ECR sequences were PCR amplified with PfuUltra high-fidelity DNA polymerase (Stratagene) using primers designed to contain specific restriction sites (Table 4; Appendix). We subcloned each PCR product into the TOPO-TA cloning vector, selected and sequenced positive colonies, and isolated plasmid DNA containing correct insert sequences. We digested the plasmids, gel purified the inserts, and re-ligated them into a similarly digested pGL3-Basic vector (Promega). We screened for and confirmed positive colonies by restriction digestion and sequencing, respectively, and isolated plasmid DNA. Using the FuGENE6 reagent (Roche) and according to the manufacturer’s instructions, 1 \(\mu\)g of each construct and 10 ng of an internal control vector (pRL-hTK; Promega) were co-transfected into HEK-293 cells that were cultured in 6-well plates containing DMEM media with 10% serum. The pGL3-Basic vector without insert and the pGL3 vector containing an SV40 promoter served as negative and positive controls, respectively. Firefly luciferase and Renilla luciferase activities were each measured 48 hours after transfection.
by the Dual-Luciferase Reporter Assay System (Promega); relative luciferase activities were calculated and normalized versus Renilla luciferase activity. Sequences containing promoter activity within ECR5, ECR22, and ECR32 have been deposited into the GenBank database and correspond to accession numbers FJ215690, FJ215689, FJ215688, respectively.

3.3.9 In vitro DNA methylation assay

Each pGL3-ECR promoter construct was treated with 2 mM S-adenosylmethionine (New England Biolabs) in the presence (methylated) or absence (‘mock’-methylated) of 6 units of M.SssI (CpG) methylase per µg of DNA for 4 hours at 37°C. As described in the text, aliquots of purified constructs were digested with HpaII to confirm the methylation status. Equal amounts of these constructs were transfected into cells for promoter-reporter assays as describe above.

3.4 Results

Genome-wide evaluation of DNA methylation across normal human tissues with microarrays revealed multiple CpG islands with tissue-specific DNA methylation, including a differentially methylated CpG island within the gene body of SHANK3. In order to determine whether this intragenic methylation has a function, we compared the DNA methylation patterns in conserved orthologous sequences in mouse and human SHANK3 in multiple tissue types. This comparative epigenetic approach revealed conserved patterns of DNA methylation in one SHANK3 CpG island (CpGi3) identified by our genome-wide study. By expanding the analysis of methylation to the three other CpG islands
throughout \textit{SHANK3}, we observed remarkable conservation of the levels of tissue-specific DNA methylation between mice and humans (Fig. 11). The 5’-most CpG island of \textit{SHANK3}, CpGi1, was unmethylated in all tissues analyzed in both mice and humans (Fig. 11B), consistent with the majority of CpG islands in the genome\textsuperscript{10}. In contrast, for three intragenic CpG islands (CpGi2-4), we observed dense methylation in PBL and intermediate or low levels of DNA methylation in brain or skin from mice, all of which were very similar to the analogous human tissues (Fig. 11C-E). In our hands, difficulty with reliably amplifying the orthologous human CpGi2 sequence from bisulfite-treated DNA precluded comparative epigenetic evaluation. However, Beri \textit{et al} successfully analyzed DNA methylation of CpGi2 in normal human PBL and brain tissue\textsuperscript{161}, the patterns of which were similar to those we observed in analogous mouse tissues. Also, the DNA methylation patterns of other CpG islands within \textit{SHANK3} described by Beri \textit{et al} are concordant with our results. The apparent persistence of the tissue-specific intragenic DNA methylation of \textit{SHANK3} over approximately 75 million years of evolution, supported by these independent observations, strongly implicates a functional, tissue-specific role for intragenic DNA methylation.
Figure 11: The tissue-specific DNA methylation patterns of SHANK3 CpG islands are evolutionarily conserved.

(A) Diagram of the ~60 kb SHANK3 gene composed of 22 exons (black bars) and four conserved CpG islands (grey bars) as defined by a length ≥200bp, an ObsCpG/ExpCpG ≥ 0.65, and a %GC ≥ 50%. Three of the CpG islands are intragenic, while one overlaps with the 5′-most translational start site (arrow). Exons encoding the known SHANK3 protein domains are indicated. Diagram is drawn to scale. (B-E) Comparison of the DNA methylation patterns of the SHANK3 CpG islands between normal mouse and human tissues. DNA methylation was determined by bisulfite treatment of DNA from the indicated tissues, PCR amplification of specific fragments within or encompassing each CpG island from the bisulfite-converted DNA, cloning of the PCR products, and sequencing of the indicated number of PCR clones (n). DNA methylation patterns and levels were determined only from highly (>95%) converted sequences. Each circle represents the overall methylation status of an individual CpG cytosine within the analyzed sequence, the color of which corresponds to the extent of methylation in all clones (white: 0-36% methylation; grey: 37-66% methylation; black: 67-100% methylation). The level of DNA methylation from each region is presented as a percentage calculated from the observed number of methylated CpG cytosines divided by the total number of CpG cytosines within the sequence analyzed from all clones. Sequence similarity was determined by
BLAT alignment (UCSC Genome Browser) of the mouse and human genomes to identify orthologous CpG island regions for direct comparison of DNA methylation patterns. In general, mouse and human CpG island sequences are >80% identical. The presentation of the data is based on 5’-3’ orientation from left to right, just as in A.

To determine the extent of tissue-specific DNA methylation throughout SHANK3, we measured DNA methylation levels of all four CpG islands in a larger panel of diverse mouse tissue/cell types using bisulfite-sequencing methods. This panel includes at least two distinct tissue/cell types arising from each of the three germinal layers. For comparison to these differentiated tissues/cells, we also included embryonic stem (ES) cells representing the earliest undifferentiated cell state in this panel. Analysis of the DNA methylation levels revealed inter- and intra- tissue and cell type-specific differences depending on the CpG island (Fig. 12A). Consistent with our initial results, CpGi1 was unmethylated in all tissue/cell types analyzed, while all intragenic CpG islands (CpGi2-4) displayed levels of DNA methylation that varied depending on the tissue/cell type (Fig. 12A). All somatic tissues harbored a measurable amount of DNA methylation, the levels of which differed significantly among the intragenic CpG islands even within the same tissue. For instance, as shown in Figure 12A, the overall percent methylation levels of CpGi2 and CpGi3 in brain were significantly different (9% and 41.7%, respectively; p<0.05, Student’s t-test), while in lung their methylation levels were nearly identical (32.1% and 37.4%, respectively). Together, these results demonstrate that the intragenic SHANK3 methylation is tissue-, cell type-, and gene region-specific, heterogeneous within
tissues, and not restricted to tissues or cell types arising from any one germinal lineage.

In the same panel of diverse tissue/cell types, we also evaluated the levels of SHANK3 expression using real-time quantitative PCR (qPCR). The highest level of expression was observed in brain tissue, with intermediate levels detected in lung and heart tissues, while low levels of expression were detected in the remaining tissue/cell types (Fig. 12A). These results are consistent with northern hybridization analyses of SHANK3 expression in analogous rat and human tissues18,139,160.

Figure 12: Characterization of SHANK3 DNA methylation and expression in diverse mouse tissue/cell types and during development.

(A) DNA methylation levels of all 4 SHANK3 CpG islands in normal mouse tissues/cell types and their corresponding expression levels. DNA methylation of the SHANK3 CpG islands was determined by bisulfite-sequencing of at least 8 clones, as described in Figure 11. For each tissue/cell type, SHANK3 expression was measured in triplicate by quantitative real-time PCR (qPCR) and calculated as a percentage relative to GusB levels. (B) SHANK3 expression during brain and lung development. The expression of SHANK3 was analyzed from brain and lung tissues of normal mice at the different stages of development (E13.5-P45, or ‘Adult’), and measured by qPCR as in A. (C-D) DNA methylation of SHANK3 CpG islands 2 (C) and 3 (D) during brain and lung development. The level of DNA methylation was measured and is presented as a percentage as described in Figure 11. Note that all expression and methylation data described here were
derived from tissues of at least two individual mice at each stage of development, except for E13.5 where whole brain tissues of 6 litter-matched embryos were pooled. For lung, E13.5 and P0 tissues were not collected for SHANK3 expression or methylation analyses. All SHANK3 expression data are representative of at least three independent qPCR experiments.

Compared to the relatively high intragenic DNA methylation levels of somatic tissues, the low levels observed in ES cells suggest that the methylation is acquired, rather than removed, during development. To investigate the dynamics of the establishment of intragenic DNA methylation patterns, we determined the methylation levels of CpGi2 and CpGi3 in brain and lung tissues from mice at different stages of development. The dynamics of methylation in these two CpG islands were studied because they exhibited inter- and intra-tissue specific differences in adult brain and lung tissue, which express SHANK3 at significantly different, yet detectable, levels (Fig. 12A). In addition to the intragenic DNA methylation, we evaluated the expression of SHANK3 in brain and lung tissues during development. Consistent with in situ hybridization results of SHANK3 mRNA in rat brain\(^{157}\), SHANK3 expression increased during the post-natal development of mouse brain (Fig. 12B). Notably, SHANK3 expression in brain increased two-fold between P7 and P14, concurrent with synaptogenesis. In contrast, SHANK3 expression in lung remained relatively low throughout post-natal development (Fig. 12B). Similar to the dynamics of SHANK3 expression, the methylation of CpGi2 and CpGi3 also varied during development depending on the tissue type. In CpGi2, the intermediate level of DNA methylation observed in adult lung tissue was established as early as P7 and then maintained through later development, whereas low levels of DNA methylation were
observed in all analyzed stages of brain development (Fig. 12C). In CpGi3, the dynamics of methylation during lung development were similar to that observed in CpGi2, however there were low levels of methylation in E13.5 brain that increased steadily throughout post-natal development to reach adult levels of methylation by P14, as late as a week after that of lung (Fig. 12D). We do not interpret this data to reflect a direct correlation between methylation and expression during development, as this would ignore the context-specificity of gene regulation and where along the \textit{SHANK3} locus expression was measured.

Our more limited interpretation is that intragenic DNA methylation is established at different stages of development in a tissue- and gene region-specific manner. This specificity further suggests a functional role for the intragenic methylation.

Our observation that intragenic DNA methylation levels vary along the \textit{SHANK3} locus depending on the tissue/cell type and developmental stage suggests that the methylation may function in a gene region-specific manner. Interestingly, at least in primary cultured cells, the intragenic DNA methylation of \textit{SHANK3} influences expression of the purported ‘full-length’ \textit{SHANK3} transcript\textsuperscript{161}, although the mechanism is unknown.

We tested the hypothesis that the gene region-specific intragenic DNA methylation regulates the activity of alternate promoters embedded within the gene. We first searched for \textit{in vivo} evidence of promoter activity in the approximately 60-kb \textit{SHANK3} gene by integrating data from diverse genome-wide datasets of ‘features’ known to be associated with promoters. These
features included sequence conservation between mice and humans, evidence of transcription initiation, and ‘bivalent’ chromatin domains.

Cross-species comparisons of DNA sequence is an effective method used to identify functional elements in mammalian genomes, many of which have been verified experimentally\textsuperscript{233-236}. For \textit{SHANK3}, we determined evolutionarily conserved regions (ECRs) by comparing human and mouse genomes whose sequences were at least 100-bp in length and greater than 70\% identical using ‘ECR Browser’\textsuperscript{232}. Comparative genomic analyses using these criteria previously identified functional genomic elements in \textit{MeCP2}\textsuperscript{237}. We identified 66 ECRs throughout \textit{SHANK3}, nearly half of which included exons and UTRs, and covering nearly one-third of the gene (19.8-kb/60-kb). To further narrow our search for functional promoter sequences that might be influenced by differential DNA methylation, we aligned the 66 ECRs to mapped sequence tags from genome-wide cap analysis of gene expression (CAGE) studies\textsuperscript{222,238}. CAGE tags are derived from mRNA sequenced in the proximity of the 5’-cap site and those that map onto unique genomic regions correspond to a potential transcriptional start site\textsuperscript{239}. This ‘cap-trapping’ approach has identified multiple novel sites of transcription initiation in protein-coding genes\textsuperscript{222-225}, with estimates that each gene in the mammalian genome on average harbors approximately 3 potential alternate promoters\textsuperscript{225}. From 145 mouse cDNA expression libraries\textsuperscript{222}, a total of 129 CAGE tags were observed throughout \textit{SHANK3}, a surprising majority (96\%) of which mapped within the gene body. Of the 129 CAGE tags, 101 (78\%) mapped to \textit{SHANK3} ECRs, and 26/66 (39.4\%) ECRs contained at least 1 CAGE
Clusters of 4 or more CAGE tags on the same DNA strand that also overlapped with ECRs are presented in Figure 13A. Interestingly, half (4/8) of these clusters map to the CpG islands of \textit{SHANK3}.

Another criteria we used to corroborate whether these regions may harbor promoter activity is the co-occupancy of histone H3K4- and H3K27-trimethylation marks derived from genome-wide ChIP-Seq analyses in ES cells\textsuperscript{240}. The prevalence of these ‘bivalent’ chromatin domains are correlated with CpG-rich promoter regions and are thought to maintain the associated gene, or gene region, in a state poised for transcriptional activation or inactivation\textsuperscript{216,240}. Using a ChIP-Seq dataset (NCBI-GEO: GSE12241), we evaluated the distribution of bivalent domains in ES cells throughout \textit{SHANK3}, and integrated this data with the \textit{SHANK3} ECR and CAGE tag clusters. Six prominent bivalent domains were observed within the body of \textit{SHANK3} in addition to one at the 5’-end (Fig. 13A). All 7 of these bivalent domains mapped precisely to ECRs, 5 of which also contained the CAGE tag clusters (Fig. 13A). ECR5 and ECR6, overlapping with CpGi1, were the only sequences that harbored bivalent domains but lacked CAGE tags (Fig. 13A). Although transcription initiation, as defined by the presence of a CAGE tag, has been observed near CpGi1, the exact location of the promoter at the 5’-end of \textit{SHANK3} is unknown. Using an mRNA ‘transcript-walking’ by RT-PCR approach with brain cDNA, we observed \textit{SHANK3} transcription initiating within ECR5 (data not shown). Altogether, our integrated and cross-species approach narrowed from over the approximately 60-kb genomic sequence of \textit{SHANK3} to just 6 ECRs (representing 4.9-kb of total
sequence) having significant potential to harbor promoter activity, 5 of which where located within the gene body.

To confirm the presence of transcription initiating intragenically, we performed 5’-RACE using a 5’-linker primer and 3’-primers (R2 or R1; Table 4, Appendix) specific to known SHANK3 exons located downstream of the CAGE tag cluster sites (Fig. 13B). Using cDNA from brain and lung tissues, we observed consistent amplification of a single PCR product of the estimated size with the R2 primer and two PCR products of differing lengths with the R1 primer. Each of these amplicons was detected in brain, but not lung, under identical optimal PCR conditions (data not shown). As illustrated in Figure 13B, representative nucleotide sequences of the cloned PCR products were aligned onto the mouse genome. All of the sequences mapped to SHANK3 and five distinct 5’-ends were identified, each being sequenced multiple times (Fig. 13B). For the R2-amplified PCR product, two 5’-ends terminated 349-bp or 504-bp downstream of the 3’-end of ECR22 within the full-length SHANK3 exon 11, potentially resulting from pre-mature RACE termination, and intron 11, respectively (Fig. 13D). For the R1-amplified PCR products, three distinct sequences were observed, all with their 5’-ends terminating precisely within the ‘non-coding’ ECR32, part of SHANK3 intron 16 (Fig. 13D). One of the three sequences lacked exon 18, consistent with observations of the full-length SHANK3 where exon 18 is spliced out of all non-neuronal adult tissues19. The largest PCR products with their 5’-ends mapping intronically were labeled 22t (602-bp; GD253656) or 32t (772-bp; GD253657) to denote transcriptional
initiation originating from the nearest corresponding non-coding ECR. Both 22t and 32t are comprised of unique first exons and downstream sequences that correspond to the known exons of the full-length \textit{SHANK3} (Fig. 13B), suggesting that these transcripts may encode variant \textit{SHANK3} proteins. In further support of this notion, we identified conserved potential translational start sites (ATG) in each of the two transcripts in-frame with the full-length SHANK3 protein (Fig. 13D). Translation of 22t from its ATG, located within exon 12, would generate an amino acid sequence identical to that of the full-length SHANK3 from the methionine residue at position 346 of the 1082 amino acid sequence of SHANK3 (NM_021423). If 22t consists of all \textit{SHANK3} exons downstream from its unique first exon, its predicted amino acid sequence would yield a smaller SHANK3 protein composed of 738 residues and would lack the conserved ankryin repeats and SH3 domain that interact with cytoskeletal proteins$^{241}$ and the glutamate receptor interacting protein, GRIP$^{156}$, respectively. Relative to the transcriptional start site and first exon sequence of 22t, those of 32t are more highly conserved (87.5% homology) between mouse and human genomes (Fig. 13D). Translation of 32t from its ATG, located within intron 16, would generate 57 novel residues followed by an amino acid sequence identical to that of the full-length SHANK3 from the alanine residue at position 451 of SHANK3. If 32t consists of all \textit{SHANK3} exons downstream from its unique first exon, its predicted amino acid sequence would yield a smaller SHANK3 protein composed of 670 residues with a novel N-terminal sequence. In addition to the ankryin repeats and SH3 domain, this putative protein would lack the PDZ domain necessary for localizing
SHANK3 to the post-synaptic densities of neurons\textsuperscript{242}. A Kozak sequence\textsuperscript{243} (TCACGCC, mouse) immediately precedes the in-frame ATG of 32t, both of which are strongly conserved across several diverse species from humans to fugu. Notably, the Exoniphy program (UCSC Genome Browser) made a strong prediction of a coding exon in the ECR32 sequence based on identifying 6 insertions and deletions, all of which are multiples of 3 bases, in the orthologous ECR32 sequences of tetrapods (data not shown), which would preserve the reading frame of 32t with that of the full-length \textit{SHANK3}. Intriguingly, an immunoblot of synaptosomal proteins probed with a specific anti-mouse SHANK3 antibody yielded three bands of distinct molecular weights, the largest of which corresponded to the full-length SHANK3 protein, while the smaller bands may correspond to unknown isoforms of SHANK3\textsuperscript{143}. Analogous to the known isoforms of SHANK2\textsuperscript{142}, a closely related family member, our data suggest that the novel \textit{SHANK3} transcripts may produce SHANK3 isoforms that lack the protein interacting domains found along the N-terminus of the full-length SHANK3, but share a common C-terminus containing other known domains. However, the functional importance of the two novel \textit{SHANK3} transcripts and whether they are in fact protein-coding remains to be determined.
Figure 13: Novel tissue-specific transcripts of SHANK3 initiate from differentially methylated internal promoters.

(A) Diagram depicting the co-incidence of promoter-associated features throughout SHANK3. Integrated genomic and epigenomic analyses of the SHANK3 locus revealed ECRs that overlap with CAGE tags clusters and ‘bivalent’ chromatin domains, many of which also occur within or near CpG islands, and are features that have been associated with known promoters in vivo. Note that the CpG islands represented here are based on the criteria described except for a smaller CpG dense region within ECR32 (light-grey bar). The lengths of all bars in the diagram are drawn to scale. (B) Sequence alignment of 5'-RACE PCR products with the SHANK3 mouse genomic sequence, demonstrating novel transcriptional start sites. As illustrated, representative sequences of several clones were aligned against the mouse genome using the UCSC Genome Browser along with the indicated ECR sequences. The transcriptional start sites identified in the sequences are proximal to or overlap with CAGE tags occurring within ECRs 22 and 32, respectively. These ECRs also harbor bivalent domains and CpG rich sequence (shown in A). SHANK3 exons are numbered according to the full-length transcript. (C) ECRs with in vivo promoter-associated features harbor promoter activity in vitro. Promoter activity was determined by dual luciferase promoter-reporter assays using selected ECR fragments (as indicated in D) cloned into the pGL3 vector and transfected into HEK-293 cells. Promoter activity, displayed as ‘luciferase activity’, was calculated from the intensity of light produced as a consequence of beetle luciferin oxidation by Firefly luciferase expressed from each ECR construct relative to that of the promoter-less pGL3-basic vector after normalizing for
transfection efficiency as measured by the intensity of light produced as a consequence of coelenterazine oxidation by Renilla luciferase expressed from a co-transfected plasmid. The numeric values above each bar are the normalized ratio of luciferase activity observed for each construct. Transfection of each ECR-construct was performed in duplicate parallel experiments and all luciferase assays were performed in triplicate. The results shown here are representative of three independent experiments. (D) Diagram of ECR22 (left panel) and ECR32 (right panel) genomic sequences displaying the regions used for promoter activity, genomic conservation between mouse, rat, and human genomes, full-length SHANK3 exons (grey boxes), the transcriptional start sites of 22t and 32t, and the evolutionarily conserved potential translational start sites (ATG).

The precise convergence of the disparate yet related promoter-associated features and the confirmed sites of transcription initiation led us to three conserved sequences most likely to harbor promoter activity: One located at the 5’-most end of SHANK3, ECR5, and two located within the gene, ECR22 and ECR32. To determine whether these ECRs function as promoters that regulate the expression of their corresponding transcripts, we amplified and cloned each ECR sequence into a pGL3 luciferase expression vector, and transfected them into HEK-293 cells for in vitro promoter-reporter assays. The central core sequence within each of the three ECRs (illustrated in Fig. 13D; FJ215690, FJ215689, and FJ215688) demonstrated substantial and consistent promoter activity (Fig. 13C), while the entire non-coding portion of ECR22 and ECR32 exhibited reduced promoter activity (approximately 2-fold or lower relative to the promoter-less pGL3-basic vector; data not shown), suggesting that the ECRs may also harbor negative regulatory elements. The human ECR5 sequence had levels of promoter activity similar to that observed from the corresponding mouse construct, and there was reduced or no promoter activity from mouse or human pGL3 constructs containing sequences up to approximately 1500-bp
upstream or 500-bp downstream of ECR5 (data not shown). Thus, as predicted from our integrated analysis depicted in Figure 13A, sequences within all three of these ECRs demonstrated promoter activity in vitro. Three other ECRs that exhibit evidence of promoters, indicated in Figure 3A as ECR-47, -59, and -64, were not tested. Therefore, the possibility remains that other intragenic regions of SHANK3, in addition to those we have identified, may function as promoters or regulatory elements.

Thus far, we have identified conserved sequences within SHANK3 that exhibit in vitro promoter activity and from which novel SHANK3 transcripts initiate. We therefore consider these sequences to be alternate promoters of SHANK3. To determine whether the intragenic DNA methylation may influence the activity of these promoters, we first investigated the methylation status within these CpG-rich sequences in vivo. Based on sequence analysis, the SHANK3 CpGi2 5’-boundary annotated by UCSC Genome Browser should extend upstream of exon 11 to include most of the minimal sequence with promoter activity in ECR22 (Fig. 13D). Because attempts to amplify this region from mouse DNA using unbiased (no CpG) PCR primers failed, we used methylation-sensitive PCR (MSP) with primers flanking the ECR22 promoter sequence, and observed very similar tissue- and cell type-specific patterns of DNA methylation as was observed by bisulfite-sequencing of the downstream UCSC Genome Browser-defined ‘CpGi2’ region in the same panel of tissue/cell types (data not shown). Due to this similarity and since bisulfite-sequencing is a more quantitative and thorough method than MSP, we evaluated the methylation of the
CpGi2 region by bisulfite-sequencing as a surrogate and extension of the ECR22 promoter methylation assessed by MSP. Consistent with previous data of CpGi2, we observed tissue- and cell type-specific methylation of the sequence proximal to the ECR22 promoter (Fig. 14A). Additionally, the CpG-rich sequence of ECR32 was evaluated for methylation by bisulfite-sequencing using primers precisely flanking the minimal sequence with promoter activity. As shown in Figure 14A, the levels of DNA methylation varied among diverse tissue/cell types. Similar to that of the intragenic SHANK3 CpG islands, the overall levels of DNA methylation proximal to the ECR22 or within the ECR32 promoter sequences were low in ES cells, high in PBL, and intermediate in lung tissue (Fig. 14A). In brain tissue and primary astrocytes derived from cerebral cortex, we observed gene region-specific differences in the level of DNA methylation: low levels in ECR22 and intermediate to high levels in ECR32 (Fig. 14A).
Figure 14: Internal *SHANK3* promoter regions are differentially methylated and their respective transcripts are differentially expressed.

(A) Overall levels of DNA methylation of the ECR22 (*top*) and ECR32 (*bottom*) promoter regions in tissues or cells. The DNA methylation levels were measured as described in Figure 11. For individual clone methylation data, see Figure 15. 

(B) The expression patterns of the novel *SHANK3* transcripts are tissue-specific. The expression levels of 22t and 32t were determined by conventional reverse-transcription (RT)-PCR using mRNA from mouse tissues/cell types. Forward primers were designed to specifically recognize and amplify the internal transcripts based on the 5’-RACE sequences, thus amplification of the transcripts confirms the 5’-RACE results. 18S expression levels were examined in each tissue/cell type as an internal control. Water was used as a negative control for the PCR. Astro., astrocytes from cerebral cortex; ES, embryonic stem cells.

In many of the different tissue/cell types, the PCR-allelic DNA methylation patterns in the ECR22 and ECR32 promoters appeared bimodal (Fig. 15). That is, for individual bisulfite-sequenced clones each representing a single PCR allele, the methylation status of most of the CpG cytosines within the
sequence were uniformly methylated or uniformly unmethylated. In skin tissue, for example, the majority (>68%) of the CpG cytosines within the ECR32 promoter were methylated in 5/15 sequenced alleles, whereas nearly all of the CpG cytosines within the ten remaining alleles were unmethylated (Fig. 15). The frequencies of the bimodal patterns of allelic DNA methylation varied among different tissue/cell types and between the two internal ECR promoters.

**Figure 15:** The internal *SHANK3* promoters are differentially methylated. DNA methylation levels for each of the individual clones sequenced for ECR22 (A) and ECR32 (B) from normal mouse tissue/cell types. The percent methylation was calculated for each bisulfite-PCR clone amplified and sequenced across the ECR22 and ECR32 regions. The percent methylation for each PCR clone is represented by a circle, the filled color of which corresponds to a particular tissue/cell type. The overall level of DNA methylation for each sample is displayed in Figure 14A.

As stated earlier, cross-tissue comparisons of intragenic DNA methylation and expression have been problematic because of the *a priori* assumption that the methylation influences transcripts originating from the 5’-most promoter of the gene (*e.g.* an effect on elongation efficiency), and because the past comparisons do not adequately consider other tissue and cell type differences in gene regulatory factors. However, our results so far demonstrate
that differential intragenic DNA methylation occurs *within* embedded promoters. Thus, it may now be appropriate to make cross-tissue comparisons of the methylation status of the intragenic promoter sequence with that of the expression of transcripts, but only those transcripts originating from the specific intragenic site.

Given this perspective, we examined the expression of 22t and 32t using conventional RT-PCR with primers specific to the unique 5’ sequence of each transcript in the same panel of tissue/cell types studied for methylation. Consistent with the 5’-RACE results, both 22t and 32t transcripts were significantly expressed in brain tissue (Fig. 14B). Interestingly, cortical astrocytes significantly expressed 22t but not 32t, which correlated inversely with low or high levels of DNA methylation in the corresponding promoter sequence, respectively (compare Fig. 14B with Fig. 14A and Fig. 15). In all other tissue/cell types analyzed, we observed little, if any, expression of 22t and 32t (Fig. 14B). Except for the ES cells, these tissues harbored alleles that exhibited high levels (>50%) of DNA methylation in the ECR22 and ECR32 promoter regions, albeit at varying frequencies (Fig. 15). As in the mouse ECR32 promoter, the same tissue- and cell type-specific levels of DNA methylation were observed in the orthologous human ECR32 sequence in a similar panel of normal human tissue/cell types, again demonstrating evolutionarily conserved intragenic methylation (Fig. 16A-B). Additionally, RT-PCR analysis using human primers orthologous to the mouse 22t and 32t sequences revealed the presence of these novel *SHANK3* transcripts in human tissue/cell types (Fig. 16C). Similar to the
full-length *SHANK3* transcript, the tissue- and cell type-specific expression patterns of 22t and 32t in humans and mice appear conserved, underscoring the potential functional importance of these transcripts in mammalian cells. Taken together, results from mice and humans indicate that the low levels of DNA methylation along the embedded promoter sequences may be permissive for the expression of the associated transcript, whereas high levels may be repressive.

Figure 16: Evolutionary conservation of ECR32 methylation and expression of the novel *SHANK3* transcripts in humans.

(A) Overall levels of DNA methylation of the ECR32 promoter sequence in human tissue/cell types. The DNA methylation levels are displayed as a percentage as described in Figure 11. Primers used for bisulfite-sequencing were designed based on the conserved orthologous human sequence to that of the mouse ECR32 region analyzed (Table 4; Appendix). (B) The tissue-specific DNA methylation patterns of ECR32 are conserved between humans and mice. The levels of DNA methylation in humans were plotted against that of the
analogous tissue/cell types in mice. A linear regression line with its R-squared value is shown to demonstrate that the tissue-specific ECR32 methylation patterns between both species are significantly correlated (R=0.93, p=0.018; ANOVA regression analysis). (C) Conservation of the tissue-specific expression patterns of the novel SHANK3 transcripts in humans. RT-PCR expression of the indicated transcripts in a subset of normal human tissues analogous to that studied in mice confirms the conservation of expression of the two internal SHANK3 transcripts. Primers were designed based on the conserved orthologous human sequence to that of the mouse transcripts (Table 4; Appendix). β-actin was used as an internal control. Kera., keratinocytes.

The heterogeneous mixture of cell types in tissue, however, precludes the absolute determination of the contribution of the allelic DNA methylation pattern observed for each of the internal promoters to the expression state of the associated transcript. To determine whether cell type-specific differences could account for the bimodal distribution of allelic DNA methylation patterns observed in tissue, we examined the methylation of the ECR32 promoter sequence in purified cell types from brain and different brain regions. As observed in whole brain, the cerebellum and hippocampus (two regions known to express SHANK3) displayed a bimodal distribution pattern of ECR32 methylation (Fig. 17C). As shown in Figure 17C, all but one of the PCR alleles sequenced in neural progenitors and their in vitro differentiated cerebellar neurons exhibited low levels of DNA methylation (<30%), while all but one of the alleles sequenced in cortical astrocytes exhibited high levels (>50%). Thus, the low levels of DNA methylation of individual alleles in brain may be derived from neurons and their progenitors, whereas the alleles with high levels of methylation observed in brain may be derived from astrocytes. Additionally, cortical astrocytes expressed significantly lower levels of 32t compared to cerebellar neurons (Fig. 18D). These results show that the methylation status of
the ECR32 promoter sequence is inversely associated with the expression state of $32t$. Consistent with this relationship, astrocytes and neurons both expressed $22t$, where the associated ECR22 promoter methylation was low in all PCR alleles sequenced (Fig. 14 and data not shown). Taken together, these results indicate that the bimodal distribution of methylation of individual PCR alleles from tissues can be attributed to cell type- or brain region-specific differences in the methylation state of the internal promoter, which in turn contributes to the different expression states of the associated transcripts.

**Figure 17: Dynamics of $32t$ expression and ECR32 promoter methylation during development.**

(4) Increased $32t$ expression and differential transcript splicing during brain, but not lung, development. RT-PCR analysis of $32t$ transcript levels was performed
as described in Figure 14B using mRNA derived from normal mouse brain and lung tissue during different developmental time points. Two PCR products of differing sizes were amplified predominately during the development of brain tissue. This size difference (24-bp) corresponds to \textit{SHANK3} exon 18, which is spliced out of the full-length \textit{SHANK3} transcript in non-neuronal adult tissues\textsuperscript{19}. A gel image representative of three experiments is shown. Exon 18 is spliced out of \textit{32t} until P7-14, after which it is retained. This developmental period coincides with significant increases in \textit{SHANK3} and \textit{32t} expression and synaptogenesis. Furthermore, differential exon 18 splicing occurred between cerebellar granule neuronal progenitors and their differentiated neurons (Fig. 18D). Altogether, these findings implicate that exon 18, which is highly conserved across multiple vertebrate species, may be important for \textit{SHANK3}’s function in neuronal maturation during synaptogenesis. (B) ECR32 methylation is established early and varies during brain development. DNA methylation levels of the ECR32 promoter region were evaluated from the indicated time points during mouse brain or lung development. DNA methylation was measured as described in Figure 11. All brain and lung tissues were collected as described in Figure 12. (C) Comparison of DNA methylation levels of individual alleles in brain tissue and purified cell types. DNA methylation levels are presented for each of the individual clones sequenced across the ECR32 promoter region in mouse tissue/cell types (as in Fig. 18C and Fig. 15). Note that data from cortical astrocytes are displayed here. Cb., cerebellar tissue; Hi., hippocampal tissue.

Our \textit{in vivo} characterization of differential methylation of the internal promoter sequences and tissue/cell type-specific expression patterns of the associated \textit{SHANK3} transcripts are consistent with the known function of DNA methylation in gene silencing\textsuperscript{17,80,210,212,244}. To determine whether DNA methylation of these internal promoters influences their function, we conducted \textit{in vitro} and \textit{in vivo} experiments to alter DNA methylation and observed how this might affect promoter activity and transcript levels.

\textit{In vitro} studies of promoter methylation have consistently demonstrated that DNA methylation abrogates promoter activity\textsuperscript{245-251}. To determine if the internal \textit{SHANK3} promoters are similarly inhibited by methylation, the ECR22 and ECR32 pGL3 promoter constructs were methylated using the CpG methyltransferase enzyme \textit{M.SssI}. This enzyme recognizes and methylates all
CpG cytosines in a given sequence. Based on the dense CpG methylation across the ECR22 and ECR32 promoter sequences observed on single PCR alleles derived from tissues (Fig. 15), *M. SssI* should recapitulate *in vivo* CpG methylation levels. Efficient *M. SssI*-mediated CpG methylation was confirmed by digestion of unmethylated and *M. SssI*-methylated plasmid DNA with the methylation-sensitive enzyme *HpaII* (Fig. 18A). ‘Mock’-methylated and untreated control ECR22 and ECR32 promoter constructs exhibited substantial levels of luciferase activity, consistent with results described in Figure 13C, however the *M. SssI*-treated constructs did not produce significant luciferase activity (Fig. 18B). Similar results were obtained from plasmids generated by re-ligating the unmethylated luciferase cDNA sequence into a reporter-less ECR22 or ECR32 pGL3 promoter vector treated or untreated with *M. SssI* (data not shown), although the absolute luciferase activity produced from the untreated re-ligated control plasmids were reduced relative to that from the plasmids without a re-ligation step. These results indicate that CpG methylation *in vitro* abolishes the activity of the intragenic promoter sequences.
Figure 18: DNA methylation of the internal SHANK3 promoters suppresses their activity.

(A) Validation of in vitro DNA methylation of constructs by methylation-sensitive restriction enzyme digestion. The internal promoter constructs described in Figure 3C were used for M.SssI-mediated CpG methylation. Plasmid DNA corresponding to M.SssI methylated (Me) or mock methylated (Mock) ECR-promoter constructs were digested with the methylation-sensitive enzyme HpaII, separated through an ethidium-bromide containing agarose gel along with undigested plasmid DNA (Un) and a 1-kb ladder (M), and visualized. (B) In vitro DNA methylation of the internal SHANK3 promoters abolished their activities. The promoter activities of Me-, Mock-, or Un- ECR22- or ECR32-promoter constructs were determined using dual promoter-reporter assays as described in Figure 3C. Representative data of three independent experiments are shown. (C) ECR32 DNA methylation differs significantly between morphologically similar astrocytes derived from distinct brain regions. The percent methylation of each bisulfite-PCR clone amplified and sequenced across the CpG-dense promoter region of ECR32, as described in Figure 4A, is plotted from untreated primary cultures of astrocytes derived from normal P7 mouse hippocampal (Hi.) or cortical (Ctx.) brain regions. The overall DNA methylation levels (Hi., 42.6%; Ctx., 60.4% methylation; solid lines) between these two groups are significantly different (p<0.001, Student’s t-test). (D) Increased 32t transcription in cortical, but not hippocampal, astrocytes after treatment with the DNA demethylating agent 5-aza-2’-deoxycytidine (5azadC). The Hi. and Ctx. astrocytes were treated with 5 µM 5azadC, separately or in combination with TSA, and 32t transcript levels were determined by RT-PCR analysis as described in Figure 4B. RNA from untreated primary cultures of cerebellar granule neural progenitor cells (CGNPs), their in vitro differentiated neurons (CG neurons), and whole brain tissue were used as positive controls. The size difference (24-bp) in amplicons from CGNPs and CG neurons corresponds to splicing of the SHANK3 exon 18 in the 32t transcript. 18S levels were measured as internal controls for the PCR. The fold-change displayed was calculated from the intensity of the 32t transcript levels from treated cells relative to that of the untreated controls after normalizing to 18S. This gel image is representative of three independent experiments. (E) Increased expression of full-length SHANK3 in astrocytes treated with the histone deacetylase inhibitor, TSA. SHANK3 expression was measured by qPCR using primers that specifically amplify a 5’ region of the full-length transcript, as used for experiments described in Figure 2. SHANK3 expression was measured in triplicate from mRNA derived from treated or untreated astrocytes as described in B and is shown as a percentage relative to GusB levels. Data displayed here are representative of three independent qPCRs. The increased expression of 32t in Ctx. astrocytes in response to 5azadC or in combination with TSA and the increased expression of full-length SHANK3 in Ctx. and Hi. astrocytes after treatment with TSA alone or in combination with 5azadC are all statistically significant (p<0.05, Student’s t-test).
These in vitro methylation data predict that if DNA methylation is a dominant mechanism suppressing promoter activity in vivo, then demethylation should re-activate it. Demethylation can be achieved using the DNA methyltransferase inhibitor 5-aza-2'-deoxycytidine (5azadC), which covalently traps the maintenance methyltransferase enzyme Dnmt1 upon incorporating into DNA during DNA synthesis and is generally associated with a release of transcriptional silencing of genes with hypermethylated promoters. To determine whether demethylation of the SHANK3 promoter activates the expression of the associated transcript in vivo, we used early-passage cultures of primary cortical astrocytes that have high levels of ECR32 promoter methylation and a low level of 32t expression (Fig. 14 and Fig. 15B). From the brain tissue of P7 mice, we isolated cortical and, for comparison, hippocampal astrocytes that were grown for 2 cell passages under identical culturing conditions. We evaluated the methylation status of the ECR32 promoter sequence in astrocytes derived from the two different brain regions. The overall levels of DNA methylation as well as the frequency of densely methylated alleles were significantly higher in cortical astrocytes compared to that of hippocampal astrocytes (Fig. 18C). Additionally, a low level of 32t expression was observed in untreated cortical astrocytes relative to a higher level in hippocampal astrocytes (Fig. 18D). To test whether this inverse relationship of ECR32 promoter methylation and 32t expression is functionally relevant, we treated both populations of astrocytes with 5azadC. As shown in Figure 18D, 32t expression was increased in the highly methylated cortical astrocytes treated with 5azadC.
compared to that of the untreated control, whereas 32t expression remained unchanged in the less methylated hippocampal astrocytes after 5azadC treatment. These results suggest that DNA methylation of the ECR32 promoter sequence prevents 32t expression \textit{in vivo}. The levels and allelic patterns of ECR32 DNA methylation as well as 32t expression were reproducible in the two independently isolated cortical astrocytes derived from the same age and strain of mice (compare Fig. 14 and Fig. 15B with Fig. 18C-D).

DNA methylation is often, though not always, associated with transcriptionally repressive heterochromatin, in part composed of hypoacetylated histones, which work cooperatively to maintain genes in a transcriptionally silent state\textsuperscript{257}. In specific cases, 5azadC-mediated demethylation and promotion of histone acetylation by the HDAC inhibitor trichostatin A (TSA) interact synergistically to increase the expression of epigenetically silenced genes\textsuperscript{66,258,259}. To determine whether histone acetylation status might influence 32t expression, we also treated the cortical and hippocampal astrocytes with TSA alone or in combination with 5azadC. In both astrocyte populations, there were no significant differences in 32t expression levels in cells treated with TSA alone nor were there any further changes to 32t expression in cells treated with a combination of TSA and 5azadC compared to 5azadC treatment alone (Fig. 18D). The observation that demethylation treatment alone was sufficient to increase the expression of 32t in cells where the associated promoter is normally highly methylated argues that the \textit{in vivo} function of the intragenic DNA
methylation is to mediate the silencing of the intragenically originating \textit{SHANK3} transcripts by inactivating their respective promoter.

Although we have discovered a functional promoter sequence for the ‘full-length’ \textit{SHANK3} transcript initiating within ECR5 (Fig. 13C and data not shown), how its activity might be regulated is unknown, particularly since it is unmethylated in all tissues/cell types analyzed (Fig. 12A). Because astrocytes express low basal levels of the full-length \textit{SHANK3}, we tested whether treatment of these cells with 5azadC and/or TSA could increase its expression. ECR5, encompassing the 5’-most \textit{SHANK3} promoter, is unmethylated in astrocytes (data not shown), and 5azadC treatment did not affect expression of \textit{SHANK3} from this 5’-most promoter (Fig. 18E). In contrast, treatment of either hippocampal or cortical astrocytes with TSA alone increased the level of full-length \textit{SHANK3} expression (3-fold or 9-fold relative to untreated controls, respectively) with no further increase when TSA was combined with 5azadC (Fig. 18E), suggesting that the histones within the 5’-most promoter region may be hypoacetylated. These results indicate that, in contrast to the intragenic promoters, the activity of the normally unmethylated 5’-most \textit{SHANK3} promoter may be influenced by the local histone acetylation status, an epigenetic modification known to be associated with gene expression states\textsuperscript{37}. Concordant with the expression level of 32t, hippocampal astrocytes also exhibited significantly higher (~5-fold) basal levels of full-length \textit{SHANK3} expression relative to cortical astrocytes (Fig. 5E; p<0.05, Student’s t-test). Further studies are required to determine the functional significance of these cellular context-
specific differences in DNA methylation and SHANK3 expression (i.e. differences between the morphologically similar astrocytes derived from distinct brain regions).

Based on the 5azadC and TSA experiments in astrocytes, the full-length SHANK3 and 32t appear to be regulated by distinct, but interrelated epigenetic mechanisms. However, both transcripts are equivalently expressed within brain and neurons (Fig. 18D and data not shown), and we found by in situ hybridization that they are expressed in the same subsets of neurons in the brain of adult mice (data not shown). Additionally, similar to the expression pattern of the full-length SHANK3 transcript (Fig. 12B), 32t transcript levels increased coordinately during brain development, in contrast to the low levels of expression in lung that remained largely unchanged (Fig. 17A). Taken together, these results suggest that although the expression of the full-length SHANK3 and 32t transcripts may be regulated by distinct epigenetic mechanisms, they appear to be coordinately expressed.

To further substantiate the conclusion that intragenic DNA methylation represses the activity of alternate SHANK3 promoters in vivo, we examined tissue from mice with genetically-induced demethylation. Two hypomorphic alleles of Dnmt1 (Dnmt1N and Dnmt1R) introduced into the mouse germline generate Dnmt1N/R animals, referred to hereafter as ‘NR’, with reduced levels of Dnmt1 expression thereby causing a reduction in global levels of DNA methylation227, though at any given single copy locus in a specific tissue, the reduction in the methylation level may be stochastic, in part. Thus, we first screened for SHANK3
expression differences in two different tissue types from 8 Dnmt1 NR and 8 litter- and/or age-matched Dnmt1 wild-type (WT) adult animals, then examined DNA methylation in a subset of samples with significant expression differences. Based on the results from astrocytes, we focused our analysis of expression on 32t and subsequent methylation analysis of the ECR32 promoter sequence. By RT-PCR, we observed a significant difference in the level of 32t expression in only one hypomorphic animal, NR-2, compared to that of WT or other NR mice in one tissue type. Figure 19A shows significantly increased expression of 32t in lung, but not cerebellar tissue from the NR-2 animal compared to that of a genetically identical litter-matched control, NR-1, whose 32t expression levels in both tissue types were similar to that observed in WT animals. Analysis of the methylation status in the ECR32 promoter sequence revealed significantly reduced levels of DNA methylation in both cerebellum and lung of the NR-2 animal compared to that observed in the NR-1 and a representative WT animal (Fig. 19B-C). In contrast, we observed little difference in the methylation of other intragenic sequences of SHANK3 in the brain and lung of these NR-1 and NR-2 animals (Fig. 20). Although mice with hypomorphic alleles of Dnmt1 have globally lower levels of DNA methylation compared to their wild-type counterparts, our results from a limited number of animals altogether suggest that demethylation of CpG cytosines at this specific locus is infrequent, varies between NR animals and tissue types, and may be stochastic. These caveats notwithstanding, we identified one NR animal with significant demethylation of the ECR32 promoter sequence. The lower level of ECR32 promoter methylation
in the NR-2 lung compared to that of a genetically identical littermate, NR-1, is associated with increased 32t transcript levels. In contrast, this association was not observed in cerebellum, which is primarily unmethylated (Fig. 19C).

Figure 19: Genetic demethylation of the ECR32 promoter region in a lung sample correlates with increased expression levels of the associated 32t transcript.

(A) Increased 32t transcription in the NR-2 lung tissue with decreased DNA methylation levels of the associated ECR32 promoter region. Quantitation of 32t transcript levels was performed from RT-PCR analyses of the indicated tissues from two Dnmt1 hypomorphic (NR) mice, NR-1 and NR-2. The level of expression was calculated from the intensity of the transcript amplicon relative to that of the housekeeping gene, 18S. Error bars represent the variability between independent RT-PCRs performed. *-denotes significant differences in 32t expression between NR-1 and NR-2 tissues (p<0.05, Student’s t-test). A representative RT-PCR gel image is displayed (inset). (B) Decreased levels of DNA methylation in the ECR32 promoter region in tissues from a mouse with hypomorphic alleles of Dnmt1. The levels of ECR32 DNA methylation in cerebellum and lung tissues from the NR-2 mouse (grey bar) were significantly lower than that observed in a wild-type (dashed line) or a litter-matched, NR-1 (black bar), control. The overall levels of DNA methylation were calculated as described in Figure 11. *-denotes significant differences in ECR32 methylation between NR-1 and NR-2 tissues (p<0.05, Student’s t-test). (C) DNA methylation of single alleles of cells in NR-2 tissue is significantly decreased compared to that of NR-1. The percent methylation was calculated for each bisulfite-PCR clone amplified and sequenced across the CpG-dense promoter region of ECR32, as described in Figure 14A, in cerebella or lung tissues from the two NR mice. These values are represented by colored circles corresponding to data derived
from NR-1 (black) or NR-2 (grey) mice. (D) Demethylation of CpG cytosines within conserved transcription factor binding sites in the ECR32 promoter sequence. The decrease in DNA methylation of each CpG cytosine analyzed in the ECR32 promoter region between the NR-1 and NR-2 lung tissues was calculated from the difference between the percent of methylated CpG cytosines observed in the NR-1 lung and that observed in the NR-2 lung. This percent difference is displayed graphically where each CpG is aligned to its position on a comparative genomic alignment of the ECR32 promoter sequence. This diagram also shows the CpGs (grey circles), the transcriptional start sites identified by 5’-RACE numbered relative to the nucleotide distance from the putative ATG at +1, and conserved transcription factor binding motifs (horizontal bars).

We also compared the differences in the percent methylation at each of the 19 individual CpG cytosine nucleotides within the ECR32 promoter region over all alleles sequenced between the NR-1 and NR-2 lung (Fig. 19D). Conserved transcription factor binding motifs containing CpG dinucleotides were identified in the ECR32 promoter sequence. Figure 19D illustrates these motifs aligned to the differential CpG methylation of each cytosine across the ECR32 promoter sequence. Interestingly, the conserved CpGs within all 3 of the E2F1 binding sites in the promoter exhibited significant demethylation in the NR-2 lung, where increased $32\text{t}$ expression was also observed. All 3 E2F1 sites are near the transcriptional start sites in ECR32 (Fig. 19D). Notably, E2F1 DNA binding is methylation-sensitive$^{260,261}$. E2F binding sites are one of two evolutionarily conserved sequence motifs shared between all three $SHANK3$ promoters. These preliminary data suggest that E2F1 could be involved in the regulation of the ECR32 and other $SHANK3$ promoters, though this remains to be substantiated.
Figure 20: Region-specific differences of \textit{SHANK3} DNA methylation in tissues of mice with hypomorphic alleles of \textit{Dnmt1}.

DNA methylation levels were determined by bisulfite-sequencing as described in Figure 11 and displayed as a percentage for each CpG island throughout \textit{SHANK3} and a non-CpG island, but CpG-rich, region (nCGi; within \textit{SHANK3} intron 12) of brain (top) and lung (bottom) tissues from the NR-1 and NR-2 mice. Other than ECR32, the only other region with a difference in methylation levels between NR-1 and NR-2 brain and lung tissue was in CpGi3, though they were not significantly different compared to that of WT (data not shown).

To determine whether the role of intragenic DNA methylation in \textit{SHANK3} applies to other genes, we evaluated 35 validated differentially methylated DNA sequences\textsuperscript{6} from the mouse genome for features of promoters which, similar to the integrated approach utilized identifying the internal \textit{SHANK3} promoters, included sequence conservation, CAGE tags from mouse tissues, and bivalent chromatin domains in mouse ES cells. Consistent with genome-wide studies of DNA methylation\textsuperscript{10,262}, many (48.9%; 22/35) of the differentially methylated loci were intragenic, a minority (11.4%; 4/35) were at the known 5’-promoter regions of genes, and the remainder (25.7%; 9/35) were
classified as intergenic (Table 1). As expected, four of four of the differentially methylated 5’-promoter sequences exhibited promoter-associated features, whereas all (9/9) intergenic loci lacked at least two of the three features (Table 1). Importantly, the majority (68.2%; 15/22) of differentially methylated intragenic loci harbored either two or three of the three promoter-associated features. These results implicate a more generalizable role for differential intragenic methylation in the regulation of alternate promoters embedded within genes and transcripts initiating intragenically (see Table 3; Appendix).

Table 1: Regions of tissue-specific intragenic DNA methylation commonly harbor promoter-associated features.

<table>
<thead>
<tr>
<th>Location of loci</th>
<th>Confirmed DMRs</th>
<th>CpG islands</th>
<th>3 of 3 PA Features</th>
<th>2 of 3 PA Features</th>
<th>0-1 of 3 PA Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’-end</td>
<td>4/35 (11.4%)</td>
<td>3/4 (75%)</td>
<td>4/4 (100%)</td>
<td>0/4 (0%)</td>
<td>0/4 (0%)</td>
</tr>
<tr>
<td>Intragenic</td>
<td>22/35 (48.9%)</td>
<td>8/22 (36.4%)</td>
<td>12/22 (54.5%)</td>
<td>3/22 (13.6%)</td>
<td>7/22 (31.8%)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>9/35 (25.7%)</td>
<td>0/9 (0%)</td>
<td>0/9 (0%)</td>
<td>0/9 (0%)</td>
<td>9/9 (100%)</td>
</tr>
<tr>
<td>Total:</td>
<td>35/35 (100%)</td>
<td>11/35 (31.4%)</td>
<td>16/35 (45.7%)</td>
<td>3/35 (8.6%)</td>
<td>16/35 (45.7%)</td>
</tr>
</tbody>
</table>

Confirmed DMRs, differentially methylated regions identified by RLGS and confirmed by independent methods; CpG islands are defined as in \(^{110}\); PA, promoter-associated features are defined in Figure 13A. For detailed information on these loci, refer to Table 3 (Appendix).

3.5 Discussion & Conclusion

Although a substantial proportion of intragenic DNA methylation is tissue-specific, its biological significance has been controversial. Some studies suggest that intragenic DNA methylation may not be functionally important\(^{219,221}\), while others have suggested that intragenic methylation reduces gene expression\(^{220}\) by regulating efficiency of transcriptional elongation\(^{135}\), or that it can determine alternative polyA site choice\(^{263}\). In addition to a constitutively unmethylated 5’-promoter associated CpG island, \(SHANK3\) harbors three evolutionarily conserved intragenic CpG islands and a smaller CpG-rich sequence which exhibit tissue- and
cell type-specific DNA methylation. Consistent with our initial report on one of these CpG islands\textsuperscript{7} and a subsequent report\textsuperscript{161}, we show that these tissue- and cell type-specific DNA methylation patterns are evolutionarily conserved between species that have diverged over 75 million years ago, suggesting that the intragenic methylation is functional. Rather than a uniform pattern across the gene body, however, we found that the intragenic DNA methylation of \textit{SHANK3} is gene region-specific within individual tissues or cell types. Brain region-specific differences were also observed within a single cell type (astrocytes). Further, the establishment of these methylation patterns occurred at multiple distinct stages of embryonic and postnatal development, depending on the tissue type. These observations suggest the existence of an unexpectedly specific control over and potentially multiple functions for intragenic DNA methylation within a single gene and tissue/cell type. Results from our integrated comparative genomics and epigenomics approach suggest an important role for tissue/cell type-specific intragenic DNA methylation in regulating the expression of novel \textit{SHANK3} transcripts from alternate promoters. Consistent with correlative studies of intragenic DNA methylation and sites of transcription initiation\textsuperscript{264}, our analysis of 22 additional genes exhibiting tissue-specific intragenic DNA methylation suggests the findings for \textit{SHANK3} may be widely applicable.

Intragenic DNA methylation of \textit{SHANK3} appears to contribute to alternate promoter usage. As summarized in Figure 21, our results from correlative and functional \textit{in vivo} and \textit{in vitro} DNA methylation experiments demonstrate that the tissue/cell type-specific intragenic DNA methylation interferes with the expression
of novel SHANK3 transcripts by inhibiting promoter activity, consistent with the known role of DNA methylation at 5’-promoters\textsuperscript{11,265,266}. In contrast to the intragenic promoters (ECR22 and ECR32), the 5’-most promoter (ECR5) of SHANK3 remained unmethylated even in tissue/cell types that did not express the associated full-length gene, indicating that the activity of the promoter is regulated via a DNA methylation-independent mechanism. The specific and significant TSA-induced increase of the full-length SHANK3 expression in two astrocyte populations isolated from different brain regions suggest that the chromatin in the 5’-promoter region is composed of hypoacetylated histones which could account for transcriptional silencing of the full-length SHANK3. These results suggest that the 5’-promoter and intragenic promoters of SHANK3 are under the control of distinct epigenetic mechanisms. In addition to providing important insight into how SHANK3 is transcriptionally regulated, which is critical for normal brain development, these results help define an important role for intragenic DNA methylation in transcription.
Figure 21: Model of epigenetic regulation of SHANK3 alternate, internal promoter usage.

(A) Schematic representation of SHANK3 transcripts and promoters. SHANK3 exons are shown in blue. All functional promoters described in this study are shown in orange and labeled according to their ECR. The transcriptional start sites (TSS) corresponding to each of the SHANK3 transcripts are indicated by arrows. (B) Summary of four in vivo methylation and expression states are shown. The style and weight of each TSS arrow indicates the relative levels of expression observed in each state: high expression, bold; intermediate expression, solid but not bold; low or undetectable expression, dashed; silenced expression, crossed-out arrows. Ovals indicate regions that are differentially methylated, the filled pattern of which denotes the density of the methylation. States I-IV corresponds to data derived from neurons, astrocytes, lung, and PBL, respectively. In general, low levels of DNA methylation create a permissive state for the expression of all three SHANK3 transcripts described in this study (State I). Methylation of intragenic promoters is associated with low or no expression of the associated transcript, whereas low levels of methylation are associated with either expression or lack of expression (i.e. permissive) of the associated transcript depending on the tissue/cell type. Demethylation of the ECR32 promoter is associated with increased 32t transcription as observed in astrocytes and lung. This process is indicated by partial CpG methylation ovals between State II and State III. At ECR5, no methylation is observed in any normal tissue/cell type analyzed in this study, indicating other mechanisms are involved in regulating the expression of the full-length SHANK3. Induced histone acetylation, presumably over the ECR5 promoter, is associated with increased expression of the full-length SHANK3 in astrocytes. This process is indicated by red ovals marked by ‘Ac’ in State II.
The conflicting views of whether intragenic DNA methylation influences transcription arose in part because endogenous genes typically contain heterogeneous levels of methylation in tissues, making it impossible to assign a single methylation state to the expressing state of a particular cell. In order to effectively determine how DNA methylation is associated with expression, analysis of single cell types are required. However, after many passages in culture, cells generally do not faithfully maintain their in vivo DNA methylation states\textsuperscript{16,267,268}. Thus, studying both tissues and cultured, but early-passage, primary cells are complementary and necessary to determine whether intragenic DNA methylation influences transcription in vivo. In addition to several tissue types, we have analyzed DNA methylation in nearly homogenous populations of early passaged neural progenitors, differentiated neurons, and cortical and hippocampal astrocytes. Comparing the methylation patterns of these different brain cell types with that of the tissues from which they were derived revealed that the tissue-specific intragenic DNA methylation of \textit{SHANK3} is also cell type-specific. Analysis of DNA methylation within the ECR32 promoter sequence in cortical and hippocampal astrocytes led to the remarkable observation that not only is the methylation tissue-specific and cell-type specific, it also varies substantially between the same cell type derived from different brain regions. These differences are not due to strain, age, or gender since the astrocytes were collected from the same set of litter-matched individuals. Our results are consistent with and expand upon the observation that different normal brain regions exhibit differential DNA methylation at specific loci\textsuperscript{269} and suggests that these brain region-specific
differences may in part be attributed to varying DNA methylation levels of morphologically similar cell types from functionally distinct brain regions. Additionally, in brain and astrocytes, we observed that the intragenic DNA methylation levels varied between specific loci along SHANK3. It is important to note that although the function of the full-length SHANK3 is known in neurons, it is expressed in other tissue/cell types, albeit at lower levels, such as we observed in lung and to a lesser extent hippocampal astrocytes. SHANK3 may have a different function in non-neuronal cells, such as its role in signal transduction in epithelial cells\textsuperscript{158}. Additionally, although the expression levels of the SHANK3 transcript variants are conserved between mouse and human tissue/cell types, are highly expressed in neurons, and differentially expressed in astrocytes derived from distinct brain regions, the functional significance of these transcripts remains unknown. Because of SHANK3’s role in neurological diseases, further studies are warranted to determine the function(s) of the transcript variants. Altogether, our results highlight the complexity and specificity of intragenic DNA methylation patterns which may contribute to brain region–specific gene expression differences\textsuperscript{270,271} and/or alternate promoter usage of genes relevant to brain function in addition to SHANK3\textsuperscript{206}.

Alternate promoter usage, which our results suggest is in part regulated by DNA methylation, gives rise to specific SHANK3 transcript variants predominantly expressed in brain. Localized to the post-synaptic densities (PSD) of excitatory neurons, SHANK3, like other SHANK proteins, is important in neural plasticity and certain cognitive functions including memory and
learning\textsuperscript{141,156}. A recent study characterizing \textit{SHANK1}-deficient mice suggested that the expression level of \textit{SHANK1} is important for normal brain development and function\textsuperscript{159}. Indeed, by inference from clinical studies of 22q13 deletion syndrome and autism, the expression level of \textit{SHANK3} may also be critical for the appropriate function of the PSD and thus must be regulated precisely during development\textsuperscript{18,19,159}. Our results indicate that epigenetic mechanisms are involved in the regulation \textit{SHANK3} transcripts during development. However, further studies, including a much larger cohort of \textit{Dnmt1} hypomorphic animals, are required to determine how \textit{trans}-acting factors, such as E2F1, might be involved in regulating the activity of the \textit{SHANK3} promoters, how their activity is affected by DNA methylation at this locus, and to determine the extent to which these \textit{cis}- and \textit{trans}-acting mechanisms may contribute to alternate promoter usage of other genes.

A majority of protein-coding genes in the mammalian genome have more than one promoter\textsuperscript{225} which produce either different proteins necessary for the function of specific cell types\textsuperscript{222,272}, or may produce untranslated RNAs. However, how alternate promoters embedded in gene bodies are regulated by epigenetic mechanisms in a tissue- and/or context-specific manner is less clear. Our model of DNA methylation-dependent alternate promoter usage is compatible with observations of other genes for which intragenic DNA methylation has been studied thus far\textsuperscript{206,264}, and may be a mechanism that can explain the regulation of alternate promoters in a context-specific manner\textsuperscript{273}. 

97
We conclude that differential intragenic DNA methylation may be a general mechanism involved in regulating the use of alternate promoters embedded within gene bodies in a tissue type-, cell type- or cellular context-specific manner. Additionally, our findings for SHANK3 provide crucial insight into how the gene is transcriptionally regulated, which has implications for clinical investigations directed at determining the role of SHANK3 and its expression level in neurological disorders. Our study underscores the complexity of DNA methylation patterns in vivo, and the potentially more significant role for intragenic methylation in gene regulation, compared to its relatively minor role in regulating 5’-CpG island promoters.

**NOTE:** A version of this Chapter has been submitted for publication in 2008.
4. Chapter IV: Conclusions

4.1 Summary of Findings

4.1.1 Genome-wide CpG island methylation

In order to determine the extent of differential DNA methylation of CpG islands genome-wide, we have developed and validated a novel approach that takes advantage of using methylation-sensitive restriction endonucleases, whose sites are enriched within CpG islands, and previously well-established BAC microarray technology. Using this approach to survey differential DNA methylation in normal human tissues/cell types, we observed that an unexpectedly large fraction of CpG islands were methylated. Taking into account the estimated number of CpG islands in the human genome (~30,000)\textsuperscript{274}, the frequency of the methylation sensitive enzyme recognition sequences within CpG islands (74% and 94% for \textit{BssHII} and \textit{NotI}, respectively)\textsuperscript{191} and the ratio of differentially methylated BACs we identified that contained only one \textit{BssHII} or \textit{NotI} site (0.02-0.03), we estimate that 1.5-3% of all CpG islands in the human genome are methylated in any given normal tissue. This estimate is conservative because, in this calculation, we did not include BACs that harbor more than one restriction site that we know exhibit differential DNA methylation (\textit{e.g.} the BAC covering the \textit{SHANK3} locus). These results directly challenge the long-held view that, at least in normal cells, CpG islands are constitutively unmethylated\textsuperscript{109}, thus bringing to question the role of tissue-specific DNA methylation within CpG islands.

Our analysis of this tissue-specific methylation and the expression of associated genes altogether highlight the importance of where along the gene the
differentially methylated CpG island occurs (i.e. 5’-promoter or intragenic). Perhaps not surprisingly, we observed that methylation of CpG islands within 5’-promoter regions appear to preclude the expression of the associated gene (e.g. C2orf32), consistent with the known function of DNA methylation at CpG island-less gene promoters\textsuperscript{266}. In contrast, however, the association of differentially methylated intragenic CpG islands and the expression of the underlying gene is more complicated – it appears that the lack of methylation may be necessary, but not sufficient, for gene expression (e.g. SHANK3). Interestingly, more than half of the differentially methylated CpG islands identified and validated occur within gene bodies. This surprising observation enticed us to further characterize intragenic DNA methylation in evolutionarily divergent species.

Since our initial report of these data in 2005, many other approaches have been developed and utilized to interrogate, in some cases more in-depth than our initial method, the global distribution of DNA methylation in various organisms\textsuperscript{124}. In general, findings from these analyses reinforce our observations that converge on two central themes currently under investigation: (1) 2-8% of CpG islands genome-wide are differentially methylated, and (2) intragenic DNA methylation is more common than methylation at 5’-promoter regions of genes.

### 4.1.2 Evolutionary conservation of tissue-specific DNA methylation

In part from the studies described in the previous section, we now know that the occurrence of \textit{in vivo} tissue-specific DNA methylation at known 5’-promoter regions is significantly less common than that observed in gene
bodies\textsuperscript{8,10}. However, unlike 5′-promoter methylation, whether or not tissue-specific intragenic DNA methylation has a biological function is unclear. The interesting observation that many divergent species of plants, insects, and mammals exhibit extensive intragenic DNA methylation is very suggestive that the methylation has an important function\textsuperscript{124}. By examining the methylation patterns of several CpG islands and CpG-rich sequences throughout \textit{SHANK3} in several tissue/cell types from rats, mice, and humans, we conclude that the intragenic DNA methylation is tissue- and cell type-specific and evolutionarily conserved. We used this example of intragenic DNA methylation as a model to investigate its role in relationship to gene expression.

### 4.1.3 Function of tissue-specific intragenic DNA methylation

Our integrated comparative genomic and epigenomic approach of the \textit{SHANK3} locus revealed two intragenic regions that harbor promoter activity, are differentially methylated \textit{in vivo}, and facilitate the tissue/cell type-specific transcription of novel and potentially protein-coding \textit{SHANK3} transcripts. Further correlative and functional experiments have demonstrated that the activity of these alternate promoters is influenced by DNA methylation (\textit{i.e.} DNA methylation ablates promoter activity). To extend our findings, we analyzed loci with known differentially methylation in mice and found that 68.2\% (15/22) of the intragenic regions also display features of promoters, much like sequences at the 5′-end of genes. Although the methylation is intragenic, the observation that DNA methylation suppresses promoter activity is consistent with its known role
at the 5’-promoter regions of genes\textsuperscript{266}, suggesting that the underlying mechanism is similar.

\section*{4.2 Implications of Findings}

\subsection*{4.2.1 Comparative genomics and epigenomics}

Our novel approach provides a practical method to screen for differentially methylated CpG island sequences genome-wide. This approach is expandable, as new generation tiling arrays consisting of 32,000 BAC clones have been developed\textsuperscript{187}, which will allow for a more thorough coverage of the genome. Further, as we have shown, use of other methylation-sensitive enzymes will allow for methylation analysis of additional sites. Additionally, our lab has developed next-generation sequencing methods for mapping the methylome.

Comparative genomics is a powerful approach to identify functional DNA elements, assuming in each case that sequence conservation dictates function. However, highly conserved non-coding sequences may have no discernable function, and other poorly conserved sequences have been proven functional\textsuperscript{8,199,275-278}. Comparative epigenomics can be performed with any type of epigenetic modification data such as DNA methylation, as was the case for \textit{SHANK3} and other genes\textsuperscript{8}, or with histone tail modification data, such as acetylation that allowed the discovery of novel enhancer elements\textsuperscript{279}. Although we identified sites of differential DNA methylation corresponding to promoters, there are examples where differential methylation contributes to the regulation of enhancer elements\textsuperscript{207,280}. Thus, integrating comparative genomics with comparative epigenomic approaches should provide additional insights into
regulatory DNA elements throughout the genome\textsuperscript{207,279,280}. Emerging technologies that allow profiling of genomes and epigenomes will significantly enhance this integrated approach\textsuperscript{281}.

4.2.2 5'-promoter versus intragenic DNA methylation

The extensive tissue-specific intragenic DNA methylation we and others have observed from global analyses implies a larger role for methylation within gene bodies than at the 5'-end of genes. In normal tissues, DNA methylation at the 5'-end of genes, including some imprinted genes\textsuperscript{282}, genes on the inactive X-chromosome\textsuperscript{283}, and some tissue-specific genes\textsuperscript{9,204,206,208}, maintains transcriptional silencing of the associated gene. There are two mechanisms by which DNA methylation silences gene expression: (1) DNA methylation can directly block transcriptional regulatory factors from binding to their target sequences within promoters\textsuperscript{266}. For example, the gene that encodes the glial fibrillary acidic protein (GFAP) is activated during astrocyte differentiation by the demethylation of a CpG cytosine that lies in its 5'-promoter region within the STAT3-binding element\textsuperscript{284}. (2) DNA methylation can also indirectly repress gene expression through several methyl-CpG-binding proteins. For instance, MeCP2 forms a complex with HDACs and a co-repressor protein, Sin3a, to repress transcription in a methylation-dependent manner\textsuperscript{61,62}. Another example includes the gene \textit{BDNF} that harbors multiple promoters, one of which is dynamically regulated by the methylation of specific CpG cytosines near a CREB transcription factor-binding site. Observed in post-mitotic neurons, when the cytosines are methylated, the methyl-binding protein MeCP2 sterically
obstructs CREB binding, thereby preventing transcription of \textit{BNDF}. In response to KCl-induced membrane depolarization, however, active demethylation of the site occurs, through an as yet unknown mechanism, allowing CREB to bind and facilitate transcription of the gene\textsuperscript{206}. Additionally, other methyl-CpG-binding proteins, such as MBD2, can form a complex with NuRD, which contains an ATP-dependent chromatin-remodelling protein, Mi-2, and HDACs, that can repress methylated promoters and remodel methylated chromatin with high efficiency\textsuperscript{285,286}, thereby linking DNA hypermethylation and histone deacetylation in transcriptional repression. Although important for transcriptional silencing, methylation at the 5’-ends (\textit{i.e.} promoters) of genes constitutes a relatively small proportion of DNA methylation in the genome\textsuperscript{8,10,124}.

In contrast to the 5’-end of genes, DNA methylation is commonly observed within gene bodies\textsuperscript{8}. Interestingly, from our more detailed analysis of \textit{SHANK3} and 15 other genes with similar features, this intragenic DNA methylation may function similarly to 5’-promoter methylation – intragenic DNA methylation appears to maintain transcriptional silencing of alternate transcripts originating within gene bodies, also through promoter elements. This notion is compatible with observations that intragenic DNA methylation may contribute to silencing of spurious or cryptic sites of initiation in plants\textsuperscript{14}, although this function may only apply to methylation that is constitutive. Differential intragenic DNA methylation, however, may explain how alternate sites of transcription initiation are normally utilized in a tissue- and cell type-specific manner. Indeed, genome-wide mRNA ‘cap-trapping’ analyses of several diverse
normal tissue and cell types from humans and mice have observed that most protein-coding genes exhibit multiple alternate sites of transcription initiation in addition to those described at the 5’-ends of genes\textsuperscript{222-225}, with an estimate that every gene in the mammalian genome harbors at least 3 alternate promoters, on average\textsuperscript{225}. The overlapping distribution of these transcription initiation sites with sites of tissue-specific DNA methylation imply that DNA methylation plays a significant role in the differential usage of alternate promoters\textsuperscript{264,273}.

\subsection*{4.2.3 Transcriptional regulation of \textit{SHANK3}}

Different promoters of \textit{SHANK3} are under the influence of distinct epigenetic mechanisms. The 5’-promoter appears to be regulated by histone acetylation while the intragenic promoters are influenced by DNA methylation. However, the similar expression patterns of the full-length \textit{SHANK3} and 32t during development suggest that the transcripts can be coordinately regulated. Additionally, our \textit{in situ} hybridization showed that the full-length \textit{SHANK3} mRNA and 32t are present in the same subset of neurons in the cerebellum and the hippocampus. Taken together, these data imply that the promoters may share common \textit{trans}-acting factors whose binding or activity could be influenced similarly by different epigenetic modifications. Intriguingly, highly conserved CpG-containing E2F and SP1 binding sites, whose corresponding proteins are inhibited by methylation of their DNA binding motifs\textsuperscript{261,287}, are the only apparent sites shared between all three promoters (data not shown), potentially a mechanism for coordinated regulation of the expression of the \textit{SHANK3} transcripts. Consistent with this notion, we observed E2F1 binding sites with
CpG cytosines that were specifically and significantly demethylated in the ECR32 promoter sequence in a lung tissue sample with increased levels of 32t expression. Thus, the intragenic DNA methylation observed in SHANK3 in certain tissue/cell types may be involved in restricting access of regulatory proteins to their targeted promoter sequences, either directly if the regulatory proteins are methylation-sensitive (e.g. E2F1) or indirectly through recruiting methyl-binding proteins that preclude binding of the regulatory proteins. Further investigation is required to elucidate how changes in histone modifications and DNA methylation might be coordinated at the SHANK3 promoters to regulate the expression of the associated transcripts during development.

The distinct epigenetic control of the promoters may function to permit the expression of specific SHANK3 transcripts depending on the tissue/cell type. For instance, in lung the 5'-promoter sequence is unmethylated and presumably composed of acetylated histones, while both internal promoters exhibit significant levels of methylation. Together, these epigenetic mechanisms could permit expression of the full-length SHANK3, albeit at lower levels than in brain, while preventing expression of the internally initiated SHANK3 transcripts. In contrast to lung, astrocytes do not express the full-length SHANK3, presumably due to hypoacetylation of its promoter. Cortical astrocytes express the 22t transcript where the ECR22 promoter is unmethylated, whereas the 32t transcript is maintained in a silenced state by promoter methylation. Although DNA methylation of the internal SHANK3 promoters is associated with silencing of the
associated \textit{SHANK3} transcript, some tissue/cell types do not require methylation for silencing. In ES cells that do not express the \textit{SHANK3} transcripts, we observed low levels of DNA methylation in the \textit{SHANK3} promoters, consistent with the methylation status of other promoters in these cells\textsuperscript{9,288}. Interestingly, these ES cells harbor bivalent domains in all of the \textit{SHANK3} promoter sequences, perhaps as a mechanism to maintain silencing of expression of the associated transcripts and to concurrently prime them for activation in specific somatic lineages upon differentiation, as suggested for other genes\textsuperscript{240,289,290}. Bivalent domains are not unique to ES cells as they occur in differentiated cells\textsuperscript{216,288-290}, and may account for the lack of expression of the \textit{SHANK3} transcripts in the cell types that also harbor low levels of DNA methylation in the corresponding promoter regions (\textit{e.g.} keratinocytes and sperm). In differentiated tissue/cell types that express the \textit{SHANK3} transcripts, the bivalent domains observed in ES cells must resolve to create an ‘open’ chromatin state, presumably involving H3K4-methylation, histone acetylation, and/or little or no DNA methylation. These conditions would permit the necessary \textit{trans}-acting factors to bind the promoter regions and induce the expression of the \textit{SHANK3} transcripts. For example, the cerebellar granule neural progenitors and their differentiated neurons have low levels of DNA methylation within the \textit{SHANK3} promoters and express all three \textit{SHANK3} transcripts. Consistently, the bivalent marks over the \textit{SHANK3} promoters are resolved in ES cell-derived neural progenitor cells, that tend to differentiate into astrocytes\textsuperscript{216}: the 5’-promoter loses the repressive H3K27 methylation mark but retains the active H3K4 methylation
mark, while both modifications are significantly lost in the ECR22 and ECR32 promoter regions (NCBI-GEO: GSE12241; data not shown). Thus, we speculate that as the bivalent domains resolve upon ES cell differentiation, DNA methylation may be established to maintain silencing of selected \textit{SHANK3} transcripts in specific somatic lineages (e.g. lung and astrocytes) that may require the expression of a particular \textit{SHANK3} transcript, while other cell types that do not require the \textit{SHANK3} transcripts remain silenced via DNA methylation-independent mechanisms that include specific histone modifications of the corresponding promoter regions. Additional investigation is required to substantiate this notion and to determine how it may be relevant to alternate promoter usage of other genes.

4.2.4 Role of the novel \textit{SHANK3} transcript variants

Despite our enhanced understanding of how \textit{SHANK3} is transcriptionally regulated, the function of the novel \textit{SHANK3} transcript variants remains unknown. Initiated from distinct promoters within intronic regions of \textit{SHANK3}, these ‘alternate’ transcripts differed in their 5’ non-coding sequences but shared the same open reading frame of the full-length SHANK3, suggesting that SHANK3 proteins with different domain compositions may be produced. \textit{SHANK1} and \textit{SHANK2} give rise to SHANK proteins with different protein-interacting domain compositions\textsuperscript{139,142}. The different isoforms could allow for a spectrum of interactions with SHANK-interacting proteins, each of which could have a specialized role in neural transmission and signal transduction at the post-synaptic densities (PSD) in distinct neuronal cell types\textsuperscript{139,156}. Because SHANK
proteins are important in neural plasticity and certain cognitive functions including memory and learning\textsuperscript{141,156}, the level, timing, and specificity of SHANK expression during development, as well as that of its transcript variants, may be quite important for the proper function of the PSD in neurons\textsuperscript{18,19,159}. Alternate promoter usage, which is in part regulated by DNA methylation established during development, may contribute to the control of expression of the \textit{SHANK3} transcript variants, which are predominately expressed in brain. In addition to the full-length \textit{SHANK3}, these variant transcripts may have important implications for the neurological disorders involving the gene, particularly since the majority of mutations and copy number variations could affect the full length \textit{SHANK3} mRNA as well as 22t and 32t. This warrants further investigation to better understand the function of the novel \textit{SHANK3} transcripts in normal brain development and whether aberrant alternate promoter usage, likely involving intragenic DNA methylation, may contribute to disease.

4.3 Limitations

4.3.1 Genome-wide CpG island DNA methylation approach

Although we identified differentially methylated CpG islands, our array approach has limitations. For example, the approach assesses a relatively small number of methylation-sensitive restriction endonuclease sites which are specifically selected for their biased towards cutting in CpG islands (\textit{e.g.} 94\% of \textit{Not}I site are within a CpG island, but only 10\% of CpG islands have a \textit{Not}I site\textsuperscript{191}). The resolution is also limited by the contents of the BAC array itself. We used an array that covers up to 10\% of the human genome with BAC clones.
spread roughly 1-Mb apart. The next generation of BAC arrays are complete tiling arrays with 32,000 BAC clones\textsuperscript{187}. Additionally, interpreting differential methylation of BACs containing multiple methylation-sensitive restriction sites can be problematic since any site can potentially contribute to methylation differences. Each site would need to be validated individually by alternative methods, such as bisulfite-sequencing, to accurately identify differentially methylated sequences. We have performed such analyses for specific BACs, including one that contained the \textit{SHANK3} locus. However, this approach is not practical for every potentially differentially methylated BAC since bisulfite-sequencing is expensive and labor-intensive. For those loci we examined by bisulfite sequencing, the concordance with the array data was high. More recent methods developed in our lab employing next-generation sequencing overcome the array limitations.

### 4.3.2 Other potential regulatory regions within \textit{SHANK3}

Our data suggest that DNA methylation precludes the activity of alternate promoters embedded within \textit{SHANK3} and potentially those within other genes for which intragenic DNA methylation has been observed. However, because we did not directly test for promoter activity of other regions within \textit{SHANK3} and the other differentially methylated genes exhibiting \textit{in vivo} evidence of internal promoters (\textit{i.e.} CAGE tags, ECRs, and bivalent chromatin domains), whether we have discovered a more generalizable role for intragenic methylation is not yet proven. One previous study has suggested that DNA methylation plays a role in regulating alternate promoter usage of \textit{fpgs}, where the differentially methylated
promoter was at the 5’-end of the gene. Our data suggest a similar role for intragenic DNA methylation, but for alternate promoters embedded within gene bodies, which is further supported by correlative data of gene expression and intragenic methylation from other genes. Thus, with ongoing investigation of the genome-wide distribution of DNA methylation across multiple tissues and cell types (e.g. the AHEAD and the NIH Roadmap epigenome projects), we anticipate that there will be other examples for which intragenic DNA methylation is associated with alternate promoter usage. Additionally, we do not exclude the possibility that intragenic DNA methylation has other roles, including regulating RNA polymerase-mediated transcriptional elongation and alternate polyA site choice. Whether or not these roles additionally apply to the regulation of \textit{SHANK3} expression requires further investigation.

### 4.3.3 Demethylation experiments

In order to test whether the intragenic DNA methylation observed for \textit{SHANK3} maintains transcriptional silencing of the \textit{SHANK3} transcript variants \textit{in vivo}, we decreased DNA methylation chemically (i.e. 5azadC) and genetically (i.e. \textit{Dnmt1}-hypomorphic animals) and measured changes in expression levels. Although results from these experiments support the hypothesis that the methylation maintains silencing of the intragenically originating \textit{SHANK3} transcripts, both demethylation strategies affect methylation throughout the genome. Thus, demethylation-induced expression of \textit{trans}-acting activators of the ECR32 promoter could account for the increased expression of \textit{32t} we observed in cortical astrocytes and the NR-2 lung. The observation that, in the
NR-2 lung, there was demethylated CpGs in the ECR32 promoter concomitant with elevated 32t expression argues for a direct effect, though does not exclude an indirect effect as well. Additionally, 5azadC is known to have pleiotropic effects on cells\textsuperscript{255}. Thus, we cannot completely rule out that other factors may have been unexpectedly activated as a consequence of global demethylation. For this reason, we have complemented these \emph{in vivo} demethylation experiments with \emph{in vitro} methylation experiments, results from which suggest the DNA methylation directly suppresses promoter activity and that the \emph{in vivo} intragenic DNA methylation observed in the alternate \textit{SHANK3} promoters may serve to inactivate the associated \textit{SHANK3} transcript variants.

\textbf{4.3.4 Tissue heterogeneity}

As mentioned in \emph{Section 3.4}, DNA methylation levels and patterns are heterogeneous within tissues. Thus, interpreting whether DNA methylation and gene expression are related from tissue data alone is problematic. Our results from analyzing primary cultures of nearly homogenous cells (\textit{i.e.} astrocytes and neurons) have demonstrated that the heterogeneity of DNA methylation within tissues (\textit{i.e.} brain) may be attributed to cell type-specific differences. We also show that expression of the \textit{SHANK3} transcripts is negatively associated with the presence of methylation within their respective promoters among these cell types, further supporting our hypothesis that intragenic DNA methylation influences alternate promoter activity. However, the cell types we analyzed were derived from different brain regions (\textit{i.e.} astrocytes were derived from cortical tissue, whereas neurons were derived from cerebellar tissue). Thus, we cannot rule out
the possibility that the apparent cell type-specific differences in DNA methylation between these cells could be due to the region of the brain that the cells were derived from. Indeed, our data that the same cell type (i.e. astrocytes) derived from different brain regions (i.e. hippocampus and cortex) showing significant differences in DNA methylation of the ECR32 promoter provides precedence for brain region-specific DNA methylation in a single cell type. Analysis of DNA methylation in cell types derived from identical brain regions will be required to definitively determine whether the differential DNA methylation observed within *SHANK3* is due to brain region, cell type, or both.

### 4.4 Future Directions

The primary hypothesis that intragenic DNA methylation influences alternate promoter usage is supported by experimental evidence described in this manuscript. While addressing this hypothesis, we made additional observations about *SHANK3* that, because of its known role in neurological disorders, warrant further investigation.

We discovered novel *SHANK3* transcript variants that originate from within the gene, which we designated 22t and 32t. As described in *Chapter III*, these transcripts may encode SHANK3 protein isoforms that have a more specialized role in neurons than the full-length protein, analogous to isoforms of SHANK2. We provide evidence from comparative genomics analysis of conserved open-reading frames suggesting that these transcripts are translated *in vivo*. Alternatively, these transcripts could be functional, but non-coding RNAs. Thus, experiments that address whether 22t and/or 32t are translated
would be an immediate priority for future study. Such experiments could include *in vitro* translation assays, which will first require knowledge of the complete composition of the \(22t\) and \(32t\) transcripts *in vivo*. This can be accomplished by extending the 5’-RACE PCR further downstream of \(SHANK3\) from the original reverse primers used in our study, so that relevant cDNA constructs can be engineered. If proteins can be generated *in vitro*, then antibodies specific to unique amino acid residues translated from these transcripts (e.g. the 57 novel potential residues of \(32t\)) can be developed to analyze the presence of these potential proteins *in vivo* using conventional techniques such as Western blotting. Results from these experiments would provide an initial indication of the functional relevance of these alternate \(SHANK3\) transcripts, an important objective for further study.

Several approaches could help discern the functional importance of the \(SHANK3\) transcript variants, if initial data suggests they encode proteins. Because the \(SHANK3\) transcript variants share exons in common with the longer full-length \(SHANK3\), results from genetic ablation of these transcripts would be difficult to interpret. An alternative would be to genetically knock-out the promoters utilized to drive the transcription of the \(SHANK3\) variants (*i.e.* ECR22 and/or ECR32) using Cre-loxP technology. This approach would accomplish two objectives: (1) provide validation of promoter activity *in vivo*, and (2) reduce or eliminate the expression of the alternate transcripts specifically, without affecting the full-length \(SHANK3\). This could be performed in animals or in cell lines. Alternatively, two other complementary strategies could involve over-expressing the \(SHANK3\)
transcript variants in cells lacking their expression (e.g. cortical astrocytes) or reduce the expression of the variant transcripts in expressing cells (e.g. neurons) using siRNA directed at the unique 5’-end of each transcript. Because the full-length SHANK3 has a role in synapse formation, dendritic spine density \(^{293}\), and Erk/PI3K-signaling \(^{158}\), measurements of these functions following perturbation of the variant transcripts could be informative. Importantly, whether these transcripts are involved in disease could be further investigated in animal models by coupling analyses of the PSD structure and composition, neural transmission, or Erk/PI3K-signaling with behavioral testing (e.g. learning and/or memory).

The data described in Section 3.4 indicate that the expression of the full-length \(SHANK3\) and the intragenically originating \(SHANK3\) transcript variants may be coordinately regulated by distinct epigenetic mechanisms (i.e. histone acetylation and DNA methylation, respectively), which likely interact with trans-acting factors (e.g. E2F1) that influence promoter activity. To validate and extend the chemical deacetylation and demethylation experiments that led to this inference, measurements of chromatin structure and composition would be important follow-up experiments. For example, chromatin-immunoprecipitation (ChIP) technology in relevant cell types (i.e. neurons and astrocytes) could be used to assess their role at the \(SHANK3\) promoters. Antibodies for ChIP should include those that specifically recognize histone tail modifications that are associated with gene expression states (e.g. H3K4-me3, H3K27-me3, and H3-Ac). Antibodies that recognize CpG methylation, methyl-binding proteins, RNA polymerase II, and candidate transcription factors like E2F1 could also be utilized to determine the
components of active and inactive \textit{SHANK3} promoters. Additionally, to identify and validate the \textit{trans}-acting factors involved in promoter-mediated regulation of the expression of the \textit{SHANK3} transcripts \textit{in vitro}, site directed mutagenesis assays of potential transcription factor binding sequences within the \textit{SHANK3} promoters could be performed, followed by functional assays of promoter activity. Data from all these experiments could be combined for a more in-depth and comprehensive understanding of the epigenetic regulation of the \textit{SHANK3} transcripts. These and similar experiments could also be performed to determine whether the additional intragenic regions identified by our integrated analysis exhibit promoter activity. The extent to which intragenic DNA methylation plays a role in regulating alternate promoters could be inferred from such analyses.

Finally, our data of the dynamics of DNA methylation of \textit{SHANK3} during normal development indicate that intragenic DNA methylation is established in a gene region-, cell type-, and developmental stage-specific manner. How is this accomplished? This question also applies to many other loci whose DNA methylation status is developmentally regulated\textsuperscript{294}. We have determined that \textit{Dnmt1} and \textit{Dnmt3b} are both involved in maintaining the intragenic DNA methylation of \textit{SHANK3}, albeit in cancer cell lines (data not shown). However, this data does not indicate which DNA methyltransferase was involved in establishing the intragenic methylation. Additionally, as described in \textit{Section 1.2.1}, other factors, including HATS, HDACs, MBDs, and ATP-dependent chromatin remodeling complexes, cooperate with the DNA methyltransferases and may interpret epigenetic marks (\textit{i.e.} histone tail modifications) during the initial steps of
methylation-driven silencing\textsuperscript{212}. The establishment of \textit{de novo} methylation may require the remodeling action of protein complexes such as SWI/SNF to make the DNA accessible to the particular Dnmt that performs the modification at the targeted promoter sequence\textsuperscript{295}. However, the precise temporal and functional connection between different epigenetic mechanisms that regulate specific loci remains unclear. Using ChIP technology with antibodies for these factors, for example, the intragenic DNA methylation of \textit{SHANK3} could provide a relevant model to determine how methylation is targeted during development, which could be applicable to the regulation of many other genes in the mammalian genome. Together, these future investigations of epigenetic control, of which intragenic DNA methylation may play a central role, will reveal insight into the maintenance of stable, cell type-specific gene expression domains critical for the identity and function of a cell, which will also have critical clinical implications.
References


## Appendix

### Table 2: Primer sequences and cycling conditions

#### A. Bisulfite PCR primer sequences

<table>
<thead>
<tr>
<th>Species</th>
<th>Target Locus</th>
<th>Name</th>
<th>Sequence</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>hSHANK3 CpG1</td>
<td>h799F10-F4</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>h799F10-R4</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 58</td>
</tr>
<tr>
<td></td>
<td>hARHGEF17</td>
<td>RP11-96D22-oF1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RP11-96D22-oR1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RP11-96D22-iF1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RP11-96D22-iR1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 58</td>
</tr>
<tr>
<td></td>
<td>mShank3 CpG1</td>
<td>m799F10-oF3</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>m799F10-oR3</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>m799F10-iF4</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>m799F10-iR4</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 55</td>
</tr>
<tr>
<td>Mouse</td>
<td>mShank3 CpG1</td>
<td>PR11-82D16F1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR11-82D16R1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR11-82D16F2</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PR11-82D16R2</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-in: 55</td>
</tr>
<tr>
<td>Rat</td>
<td>mShank3 CpG1</td>
<td>r799F10-oF1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r799F10-oR1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r799F10-iR1</td>
<td>GAT TGT TTT TGG GYG G</td>
<td>Nested-out: 48</td>
</tr>
</tbody>
</table>

(A) PCR reactions were incubated at 95°C for 2 min and then cycled as follows. For touchdown PCR (TD), 40 cycles of 95°C for 30 sec, 66°C for 30 sec (reducing the annealing temperature by 2°C every 2 cycles until 56°C was reached), 72°C for 60 sec, with a final extension of 72°C for 10 min. For Nested PCR, 30 cycles of 95°C for 60 sec, listed annealing temperature for 60 sec, 72°C for 60 sec, with a final extension of 72°C for 10 min. A 2 ul aliquot was used for the inner nested PCR using the same parameters as the outer, except that the cycle annealing and extension temperatures were held for 30 sec each. (B) Real-Time RT-PCR was conducted using the Applied Biosystems’ (ABI) Assays-on-Demand Assay or the ABI SYBR Green Master Mix using the sequences outlined. The cycling conditions were per manufacturer’s instructions.
<table>
<thead>
<tr>
<th>Confirmed by</th>
<th>ES cell</th>
<th>Liver</th>
<th>Kidney</th>
<th>Brain</th>
<th>Colon</th>
<th>Muscle</th>
<th>Testis</th>
<th>Promoter-Associated Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequenom, bis-seq</td>
<td>2 2</td>
<td>2 2</td>
<td>2 2</td>
<td>2 2</td>
<td>2 2</td>
<td>2 2</td>
<td>2 2</td>
<td>2 2</td>
</tr>
<tr>
<td>sequenom</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
</tr>
<tr>
<td>sequenom, MSP</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
<td>1 1</td>
</tr>
<tr>
<td>sequenom</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>sequenom</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>sequenom</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>sequenom</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>sequenom</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>sequenom</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
</tbody>
</table>

The chromosome position of each locus is shown (based on the Feb. 2006 assembly of the mouse genome; UCSC Genome Browser). CpG islands are defined as in 110. Promoter-associated features are defined and analyzed as described in Figure 13A. ‘–triMe’, trimethylation. Differentially methylated loci identified by RLGS were confirmed using Sequenom Mass Spectrometry-based methylation assays 296, MSP, and/or bisulfite-sequencing methods.
Table 4: List of primers, their applications, and reaction conditions.

<table>
<thead>
<tr>
<th>Species</th>
<th>Target Locus</th>
<th>Name</th>
<th>Application (template)</th>
<th>Primer Sequence</th>
<th>Conditions</th>
<th>Product Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Shank3 CpGi1</td>
<td>hSHK3-CpGi1-F</td>
<td>Touchdown PCR</td>
<td>GTT TYG TCG TAG AAG TGT TTG</td>
<td>TD: 67 -&gt; 57, 50 cyc</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>Shank3 CpGi1-R</td>
<td>hSHK3-CpGi1-R</td>
<td>(bisulfite-gDNA)</td>
<td>TCC CAG YAT TAT TAC CCG ACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 CpGi2</td>
<td>hSHK3-CpGi2-F</td>
<td>Touchdown PCR</td>
<td>GTT TYG TCG TAG AAG TGT TTG</td>
<td>TD: 65 -&gt; 55, 40 cyc</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Shank3 CpGi2-R</td>
<td>hSHK3-CpGi2-R</td>
<td>(bisulfite-gDNA)</td>
<td>CTA CTT CAC CAA AAA ACT CCT AC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 CpGi4</td>
<td>hSHK3-CpGi4-F</td>
<td>Touchdown PCR</td>
<td>TGG TYG GAC AGT ATT TAT TGT</td>
<td>TD: 54 -&gt; 54, 46 cyc</td>
<td>346</td>
</tr>
<tr>
<td></td>
<td>Shank3 CpGi4-R</td>
<td>hSHK3-CpGi4-R</td>
<td>(bisulfite-gDNA)</td>
<td>ATA AGG ACA CTA AAC AAT ATA CCA AA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32</td>
<td>hECR32-bis-oF</td>
<td>Outer PCR</td>
<td>ATG TTG AGT TTT GAG TTT TGT GGT</td>
<td>Tm: 51, 36 cyc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-R</td>
<td>hECR32-bis-oR</td>
<td>(gDNA)</td>
<td>TCA TCT CAT CTA ACC ACC CCA TTA T</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-iF</td>
<td>hECR32-bis-iF</td>
<td>Inner PCR</td>
<td>AAA ACC TGG TAA ACC CAA ACT ATA C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-iR</td>
<td>hECR32-bis-iR</td>
<td>(outer product)</td>
<td>AAA ACC TGG TAA ACC CAA ACT ATA C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-RT-F</td>
<td>hECR32-RT-F</td>
<td>RT-PCR</td>
<td>GCC CCA GCC CCG TAC ATG GCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-RT-R</td>
<td>hECR32-RT-R</td>
<td>(outer product)</td>
<td>GGC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>hECR32-RT-F</td>
<td>RT-PCR</td>
<td>GCC CCA GCC CCG TAC ATG GCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>hECR32-RT-R</td>
<td>(outer product)</td>
<td>GGC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-actin</td>
<td>hActinB-F</td>
<td>RT-PCR</td>
<td>GGA TCT TCA TGA GGT AGT CAG TC</td>
<td>Tm: 60, 25 cyc</td>
<td>626</td>
</tr>
<tr>
<td></td>
<td>Shank3 22t</td>
<td>hECR22-RT-F</td>
<td>RT-PCR</td>
<td>GCA TCT TAC GTA ACG TCA TCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>hECR32-RT-F</td>
<td>RT-PCR</td>
<td>GCC CCA GCC CCG TAC ATG GCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>hECR32-RT-R</td>
<td>(outer product)</td>
<td>GCC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>Shank3 ECR5</td>
<td>hECR5-F</td>
<td>PCR</td>
<td>(KpnI: GG GGTACC) DCG CCA GAG TGC AGT GGA AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR5-R</td>
<td>hECR5-R</td>
<td>(gDNA)</td>
<td>(KpnI: GG GGTACC) DCG CCA GAG TGC AGT GGA AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR22</td>
<td>hECR22-F</td>
<td>PCR</td>
<td>(KpnI: GG GGTACC) DCG CCA GAG TGC AGT GGA AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR22-R</td>
<td>hECR22-R</td>
<td>(gDNA)</td>
<td>(KpnI: GG GGTACC) DCG CCA GAG TGC AGT GGA AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32</td>
<td>hECR32-bis-oF</td>
<td>Outer PCR</td>
<td>TAT TTG AAG AGG GGA TAG TTG GTT</td>
<td>Tm: 52, 35 cyc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-bis-oR</td>
<td>hECR32-bis-oR</td>
<td>(bisulfite-gDNA)</td>
<td>ATT CTA CTC TAA AAA TCC ACC CTA C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-bis-iF</td>
<td>hECR32-bis-iF</td>
<td>Inner PCR</td>
<td>ATT CTA CTC TAA AAA TCC ACC CTA C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-bis-iR</td>
<td>hECR32-bis-iR</td>
<td>(outer product)</td>
<td>ATT CTA CTC TAA AAA TCC ACC CTA C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-RT-F</td>
<td>hECR32-RT-F</td>
<td>RT-PCR</td>
<td>GCC CCA GCC CCG TAC ATG GCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 ECR32-RT-R</td>
<td>hECR32-RT-R</td>
<td>(outer product)</td>
<td>GCC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>hECR32-RT-F</td>
<td>RT-PCR</td>
<td>GCC CCA GCC CCG TAC ATG GCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>hECR32-RT-R</td>
<td>(outer product)</td>
<td>GCC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18S</td>
<td>m18S-F</td>
<td>RT-PCR</td>
<td>AGT GCG GGT CAT AAG CTT GC</td>
<td>Tm: 60, 20 cyc</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>Shank3 22t</td>
<td>mECR22-RT-F</td>
<td>RT-PCR</td>
<td>CGC CTT GTG GGG TTC CTG TGT CA</td>
<td>Tm: 64, 36 cyc</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>mECR32-RT-F</td>
<td>RT-PCR</td>
<td>GCC CCA GCC CCG TAC ATG GCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 32t</td>
<td>mECR32-RT-R</td>
<td>(outer product)</td>
<td>GCC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shank3 3t</td>
<td>mECR32-RT-R</td>
<td>(outer product)</td>
<td>GCC CTA CTG GAC TGG TCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note: PCRs were incubated at 95°C for 2 min and then cycled as follows: For touchdown (TD) PCR, 2 cycles of 95°C for 30 sec, 1st Tm°C for 30 sec (reducing the annealing temperature by 2°C every 2 cycles until the last Tm°C listed was reached at which point the listed number of cycles were performed), 72°C for 45 sec, with a final extension of 72°C for 10 min. For Nested PCR, 30 cycles of 95°C for 60 sec, listed annealing temperature (Tm) for 60°C, 72°C for 60 sec, with final extension for 10 min. A 2 µl aliquot was used for the inner ‘nested’ PCR using the same parameters as the outer PCR, except that the cycle annealing and extension temperatures were held for 30 sec each. *-denotes that 1X Q-solution (Qiagen) was added to the PCR. Expected product sizes are shown.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
UCSF Library Release

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

[Signature]

Author Signature

[Date]

11-24-2008