UNIVERSITY OF CALIFORNIA, SAN DIEGO

Dynamics of the Plant-Pathogen Interaction: Strategies for Bacterial Virulence and Coordinating the Plant Defense Response

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Biomedical Sciences

by

Robert Houston Dowen III

Committee in charge:

Professor Jack E. Dixon, Chair
Professor Joseph Ecker
Professor Victor Nizet
Professor Bing Ren
Professor Palmer Taylor

2009

The Dissertation of Robert Houston Dowen III is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2009

Table of Contents

List of Figures

Chapter 2

Chapter 3

Appendix A

Appendix B

List of Tables

Chapter 3

Appendix A

Appendix B

Acknowledgements

First and foremost, I would like to thank my thesis advisor, Jack Dixon. He has provided a supportive laboratory environment where I was free to pursue the biological questions that I found the most interesting, regardless of how they meshed with other projects in the lab. Jack has taught me how to recognize good science and tell an interesting and compelling story. I will never forget the raw enthusiasm he has for cutting edge science and I will take this with me as I move forward as an independent researcher.

I was lucky enough to a have a co-mentor, Joe Ecker, during my time in graduate school. Joe has allowed me full access to the resources of his lab and has made all of the work presented in this dissertation possible, and for that I am very grateful. From Joe I have learned how to identify and attack the most challenging biological questions within one of the most competitive fields in science. He has a relentless passion for science that I will never forget.

I would also like to thank all the members of the Dixon and Ecker labs. Their kindness and intelligence made the lab an exciting and enjoyable place to work. Special thanks to James Engel, Ryan Lister, Hong Qiao, Carolyn Worby, Matt Rardin, Doug Mitchell, Cheri Lazar, Fred Robinson, Matthew Gentry, Dave Pagliarini, Cynthia Wong, and Melissa Lowe for helping me along the way.

I would also like to thank the additional members of my thesis committee for their insightful suggestions and for keeping the wheels moving along the way: Dr. Victor Nizet, Dr. Palmer Taylor, Dr. Bing Ren, and Dr. Scott Emr.

I also must thank my wife, Jill Dowen, who has faithfully stood by my side through all trying times of graduate school. She has always been supportive and

understanding of the demands of the lab and has contributed to my growth both personally and scientifically.

The text of Chapter 2 is a reprint of the material as it appears in the *Journal of Biological Chemistry,* 2009, Vol. 284, No. 23, Robert H. Dowen, James L. Engel, Feng Shao, Joseph R. Ecker, and Jack E. Dixon. The dissertation author was the primary researcher and the co-authors listed in the publication assisted and/or supervised the research that forms the basis of this chapter.

The text of Chapter 3, in part, is currently being prepared for submission for publication by Robert H. Dowen, Ryan Lister, Mattia Pelizzola, Joseph R. Nery, Jack E. Dixon, and Joseph R. Ecker. The dissertation author was the primary researcher and the co-authors listed either assisted and/or supervised the research that forms the basis of this chapter.

The text of Appendix A is a reprint of the material as it appears in *Journal of Cell Biology*, 2007, Vol. 178, No. 3, 477-488. Matthew S. Gentry, Robert H. Dowen III, Carolyn A. Worby, Seema Mattoo, Joseph R. Ecker, and Jack E. Dixon. The dissertation author was a major contributing researcher and second author of this paper.

The text of Appendix B, in full, has been submitted for publication by Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A.Thomson, Bing Ren, and Joseph R. Ecker. The dissertation author was a major contributing researcher and second author of this paper.

Vita

Education

2002-2009 **University of California, San Diego**
Ph.D., Biomedical Sciences
Thesis Advisors: Drs. Jack E. Dixon and Joseph R. Ecker
Thesis Title: Dissection of the Molecular Localization Strategies Utilized by Bacterial Type III Effector Proteins.

1998-2002 **University of North Carolina at Chapel Hill**
B.S., Chemistry (Highest Honors)
Honors Advisor: Dr. T. Kendall Harden
Honors Thesis Title: Identification of Regulatory Domains in the Human P2Y$_2$ Receptor.

Professional Experience

2002-2009 **University of California, San Diego**
Biomedical Sciences Ph.D. Candidate

- Utilized various biochemical approaches, including purification of recombinant proteins, to examine self-proteolytic processing and subsequent fatty acylation of bacterial type III effector proteins

- Genetically manipulated *Arabidopsis* and *Pseudomonas syringae* strains to investigate type III effector function *in vivo*

- Generated whole genome maps of *Arabidopsis* DNA methylation, mRNA transcripts, and smRNA levels in response to *Pseudomonas syringae* infection using Illumina GA sequencing

- Utilized computational approaches to probe the *Arabidopsis* and Human DNA methylomes (UNIX, MySQL, PHP)

1998-2002 **University of North Carolina at Chapel Hill**
Undergraduate Researcher

- Examined the mechanism of agonist-induced phosphorylation, desensitization, and internalization or the human P2Y$_2$ Receptor

- Performed a variety of cell culture-based radiochemical and cell biology experiments to track localization of the P2Y$_2$ Receptor

Publications

- **Dowen, R.H.**, Lister, R., Pelizzola, M., Dixon, J.E., and Ecker, J.R. Epigenetic regulation of the plant defence system against *Pseudomonas syringae*. In preparation.

- Lister, R., Pelizzola, M., **Dowen, R.H.**, Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. Human DNA methylomes at single-base resolution reveal widespread cell-specific epigenetic signatures. Submitted 2009.

- **Dowen, R.H.**, Engel, J.L., Shao, F., Ecker, J.R., and Dixon, J.E. A family of bacterial cysteine protease type III effectors utilize acylation-dependent and independent strategies to localize to plasma membranes. *J Biol Chem*. 2009 Jun 5;**284**(23):15867-79.

- Gentry, M.S., **Dowen, R.H.**, Worby, C.A., Mattoo, S., Ecker, J.R., and Dixon, J.E. The phosphatase laforin crosses evolutionary boundaries and links carbohydrate metabolism to neuronal disease. *J Cell Biol*. 2007 Jul 23;**178**(3):477-88.

- Digicaylioglu, M., Kaul, M., Fletcher, L., **Dowen, R.**, and Lipton, S.A. Erythropoietin protects cerebrocortical neurons from HIV-1/gp120-induced damage. *Neuroreport* 2004 Apr 9;**15**(5):761-3.

Seminar Presentations

- **Pharmacology Seminar Series**, 2008
  University of California, San Diego
  “Hijacking Plant Signaling Pathways by *Pseudomonas syringae*”
- **Biomedical Sciences Departmental Annual Retreat**, 2006
  La Jolla Shores Hotel and Conference Center
  “Hijacking Host Signal Transduction Pathways by Bacterial Effector Proteins”

- **Pharmacology Seminar Series**, 2005
  University of California, San Diego
  "Hijacking Host Signal Transduction Pathways by Bacterial Effector Proteins"
- **Pharmacology Departmental Annual Retreat**, 2004
  University of California, San Diego
  "Hijacking Host Signal Transduction Pathways by Bacterial Effector Proteins"

Teaching Experience

- **Salk Institute Mobile Science Lab**, 2004.
  Taught an outreach program that brings hands-on laboratory science to the classrooms of underprivelaged students in San Diego County.

References

- **Jack E. Dixon, Ph.D.**
  Vice President and Chief Scientific Officer Howard Hughes Medical Institute
  Professor of Pharmacology, Cellular & Molecular Medicine, and Chemistry & Biochemistry
  University of California, San Diego
  9500 Gilman Drive, Leichtag Bldg., Room 284
  La Jolla, CA 92093-0721
  Telephone: 858-822-0491 (UCSD), 301-215-8803 (HHMI)
  Fax: 858-822-5888 (UCSD), 301-215-8828 (HHMI)
  Email: jedixon@ucsd.edu (UCSD), dixonj@hhmi.org (HHMI)

- **Joseph R. Ecker, Ph.D.**
  Professor of Plant Molecular and Cellular Biology
  The Salk Institute for Biological Studies
  La Jolla, CA 92037
  Telephone: 858-453-4100 ext. 1795
  Fax: 858-558-6379
  Email: ecker@salk.edu

- **Victor Nizet, M.D.**
  Professor of Pediatrics & Pharmacy
  Chief, Division of Pediatric Pharmacology & Drug Discovery
  University of California, San Diego School of Medicine
  Skaggs School of Pharmacy & Pharmaceutical Sciences
  9500 Gilman Drive, CMM-East, Room 1066
  La Jolla, CA 92093-0687
  Telephone: 858-534-7408
  Fax: 858-534-5611
  Email: vnizet@ucsd.edu

ABSTRACT OF THE DISSERTATION

Dynamics of the Plant-Pathogen Interaction: Strategies for Bacterial Virulence and Coordinating the Plant Defense Response

by

Robert Houston Dowen III

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2009

Professor Jack E. Dixon, Chair

Using a myriad of genetic, biochemical, and cell biology based approaches, the interactions between pathogens and their hosts have been intensely examined since microbes were first described by Hooke and van Leeuwenhoek in the 17th century. Despite these efforts, the molecular mechanisms that underlie the basis of pathogenesis and host defense are still only partially unraveled. Our work has focused on the interaction between bacterial phytopathogens, specifically *Pseudomonas syringae*, and the reference plant *Arabidopsis thaliana*.

Here, we examined the sub-cellular localization strategies utilized by a subset of *P. syringae* virulence factors, the AvrPphB-like family of type III effector proteins.

Remarkably, some members of the AvrPphB family of effector proteins utilize their own cysteine protease activities to direct their localization within the plant cell. We have demonstrated that, following delivery through the type III secretion system, these effectors undergo self-proteolytic processing to reveal a novel amino terminus containing consensus sites for eukaryotic fatty acylation. We show that these effectors hijack the eukaryotic acylation machinery to ensure lipid modification, and are consequently delivered to the host plasma membrane. Additionally, we found that acylation of AvrPphB by the host lipidation machinery is absolutely required for successful cleavage of its *in planta* substrate. Finally, we have demonstrated that additional AvrPphB family members, surprisingly, employ acylation-independent localization strategies. Nonetheless, these effectors also localize to the host plasma membrane, underscoring the plant plasma membrane as a critical site for type III effector function.

In addition, we have also investigated specific aspects of the plant defense system. Although the signaling pathways that encode innate immunity against invading microbes have been well studied, epigenetic regulation of plant defenses through modification of heritable DNA methylation patterns represents an unexplored regulatory mechanism of plant defense. Here, we show that *Arabidopsis* mutants deficient in cytosine methyltranserase activity are markedly more resistant to *P. syringae*, a phenotype that strongly correlates with up-regulation of known plant defense genes. Remarkably, we have also demonstrated that wild-type plants utilize transgenerational memory of *P. syringae* infection to encode enhanced pathogen resistance in their progeny, a mechanism that requires the activity of the DRM1, DRM2, and CMT3 DNA methyltransferases. Furthermore, using high-throughput sequencing technology, we mapped methylcytosines across the genome of

Arabidopsis plants infected with *P. syringae*, and found multiple examples of pathogen-induced transient DNA methylation changes that correlate with transcriptional changes of proximal genes. Together, our results have revealed that cytosine methylation contributes to regulation of plant defense genes during infection, and that transient alterations in DNA methylation patterns may encode resistance to pathogens in subsequent generations.

CHAPTER 1

Introduction

Plants are constantly under attack from a wide spectrum of pest and pathogen species, including a variety of insects, viruses, oomycetes, fungi, bacteria, and nematodes. Disease or wounding that arises from infection can occur in a wide range of plant tissues, often leading to extraordinary agricultural losses each year. The interaction between model plants (*Arabidopsis thaliana* or tobacco, for example), as well as crop plants (maize, rice, tomato), and these pathogen species has recently become an intense arena for basic scientific research. The molecular interactions between the plant host and the pathogen have been difficult to dissect, however, several advancements in the phytopathogen field have underscored the complex and highly sophisticated mechanisms utilized the by the pathogen to propagate within the host and, in contrast, the elaborate plant defense response that is responsible for controlling these pests (1-3).

One particular subdivision of phytopathogen research has focused on the interaction between Gram-negative bacterial pathogens and the plant host. The *Pseudomonad* family of Gammaproteobacteria includes a large number of plant pathogens, as well as symbiotic plant growth-promoting bacteria, that are often found in the surrounding soil. Bacterial phytopathogens including *Pseudomonas syringae*, *Xanthomonas campestris*, *Erwinia amylovora*, and *Ralstonia solanacearum* enter the plant through wounds or open pores (stomata and hydathodes) in the leaf tissue and propagate within the intracellular space (apoplast), but are not believed to invade and replicate within the plant cells, a phenomena that has been observed with a variety of mammalian bacterial pathogens (4,5).

In contrast to the mammalian innate immune system that utilizes specialized cells to defend against infection, all plant cells appear to be equally equipped to initiate an innate immune response, as well as generate systemic signals to warn neighboring cells or tissues of invading microbes. Despite a seemingly primitive defense system, most invading bacteria are rapidly detected by plant cell surface receptors that act as an initial non-specific surveillance system. Activation of these membrane-localized pattern-recognition receptors (PRRs) results in initiation of a basal defense response that efficiently restricts the growth of most invading microbes (6).

Extracellular identification of invading bacteria by PRRs occurs through recognition of a variety of evolutionarily conserved nonspecific elicitors termed microbe-associated molecular patterns (MAMPs) that include bacterial cold-shock proteins, lipopolysacharides (LPS), and the bacterial flagellin structure (6). Interestingly, the potent mammalian pathogen *Escherichia coli* O157:H7 is unable to multiply or generate disease symptoms in the model plant *Arabidopsis thaliana*, and rapidly stimulates a characteristic set of MAMP-induced transcriptional changes, indicating a heavily conserved component of the bacterium is detected by the plant cell (7).

Virulent bacterial pathogens, however, have evolved complex strategies to thwart the plant basal defense response. These phytopathogens, which include a wide variety of *Pseudomonas syringae* strains, utilize a type III secretion system (TTSS) to inject an arsenal of virulence factors (or effector proteins) into the cytoplasm of host cells (1,8). Effector proteins act in a coordinated, and likely redundant, fashion to suppress basal defense pathways, thereby rendering the plant susceptible to disease. One of the most intensely studied plant pathogens, *Pseudomonas syringae* pv. *tomato* DC3000, secretes approximately 30 effector proteins into the plant cell and its host specificity is in part defined by this collection of effectors (9). Surprisingly, plant symbiots including a variety of *Rhizobia* and *Bradyrhizobia* species, as well as some *Pseudomonas* species (for example, *Pseudomonas fluorescens*), also employ a TTSS to deliver effectors into root cells during colonization, presumably to evade initial MAMP-induced defenses and escape clearance from the host (10). Although the molecular mechanisms underling effector action in the host cell are largely unknown, it has become clear that they are indispensable in promoting disease, as *Pseudomonas syringae* pv. *tomato* DC3000 mutants lacking a functional TTSS lose virulence on a variety of hosts species (11).

In contrast to susceptible plants, some plants are remarkably resistant to known pathogen species. These plants specifically recognize a pathogen strain and induce a sustained and potent defense response that includes induction of the Hypersensitive Response (HR, Martin *et al.* (2)). Bacterially-induced HR results in a localized programmed cell death response surrounding sites of infection and occurs coincidentally with pathogen growth restriction (12). Interestingly, as is the case for some fungal parasites, it appears that HR-induced cell death restricts nutrient availability from the feeding pathogen, thereby preventing spread of the organism to neighboring plant tissues (13). However, it currently remains unclear if HR plays an analogous role in restricting bacterial growth or whether this response is simply a byproduct of the anti-microbial conditions generated to combat infection.

Interestingly, this specialized defense response, characterized by HR induction, results in activation of similar signaling pathways and transcriptional changes as observed during MAMP-induced defenses; however, the initiation mechanism, amplitude, and timing of this response differs dramatically from that of basal defenses (14). Although the complex network of signaling molecules involved in HR are only beginning to be unraveled, it is clear that HR progression requires activation of plant disease resistance proteins (R proteins, encoded by *R* genes) that specifically recognize pathogen-derived elicitors (3). Bacterial effector proteins represent one such class of elicitors that can be monitored by the host *R* gene repertoire. The molecular mechanisms that result in activation of *R* gene-mediated defense have been an intense area of research and will be discussed further below. Remarkably, co-evolution of the host *R* genes and the bacterial effectors is an extremely dynamic process, representing an arms race that is dictated simply by the rate of evolution of each gene (14). Not surprisingly, some bacterial pathogens have evolved effector proteins that target downstream of R protein activation to inhibit HR and promote disease, illustrating the dynamic relationship that exists between the plant host and bacterial pathogen (15).

The Bacterial Type III Secretion System

Although many *P. syringae* species utilize multiple secretion systems (*P. syringae* pv. *tomato* DC3000 carries Type I-VI secretion systems), the type III secretion system is the only one that is essential for pathogenicity (11,16). The TTSS is encoded by a collection of *hrp* and *hrc* genes (~20 genes) that are almost always clustered with type III effector genes within specific genomic features termed pathogenicity islands, and not surprisingly, both classes of genes are transcriptionally regulated by common components (17,18). Interestingly, many of the conserved core components of the TTSS share striking sequence similarity to the flagella structure of Gram-negative bacteria and it is likely that the both systems share a common ancestral system (18,19).

The TTSS of most bacterial pathogens is composed of three major structural components, an anchor, a needle-like or pili structure, and a translocation complex, each of which are comprised of multiple types of proteins (18). The needle-like structure of *Pseudomonas syringae*, called a Hrp pilus, is extended, and likely decorated with cell wall-degrading harpins, in order to penetrate the cell wall of the plant cell. Amazingly, the protein that comprises the majority of the Hrp pilus structure, HrpA, shares very little sequence similarity to other HrpA-like proteins, even when compared to very closely related *Psedomonas syringae* strains. Evolutionary diversification of this widely used extracellular structure is likely an attempt to avoid a generalized recognition by the plant defense surveillance system (18).

Physical interaction between the pathogen and host cell initiates signals, via a poorly understood mechanism, within the bacterium to induce transcription of structural and regulatory components of the TTSS, as well as several type III effectors. Although the mechanism by which phytopathogen type III effectors are delivered through the TTSS into the eukaryotic cell has been quite controversial, it is likely that these proteins remain in an unfolded or partially unfolded state through the action of effector-specific chaperones. Consistent with this hypothesis, the *P. syringae* pv. *syringae* 61 strain utilizes the ShcA protein chaperone to efficiently deliver the HopPsyA effector through the TTSS (20), and it is likely that additional effectors use similar strategies. Additionally, it has become apparent that effector proteins also utilize an embedded *N*-terminal signal sequence, which generally possess predictable biochemical patterns, to ensure recognition and subsequent delivery through the TTSS (21,22). A number of recent studies have probed the mechanism by which type III effector proteins are regulated in the bacteria and subsequently presented to the TTSS for secretion, and examination of this component of molecular pathogenesis is essential for fully understanding the life cycle of these bacteria.

Thwarting MAMP-Induced Defense Responses

Perception of bacteria occurs at the plant cell surface by a collection of widely distributed membrane-associated receptors (PRRs) that monitor the surrounding environment for microbe-specific, evolutionary conserved molecules, or MAMPs (23). The best-characterized elicitor of MAMP defenses is the bacterial flagellin protein, which is rapidly sensed by the *Arabidopsis* leucine-rich repeat (LRR) receptor kinase FLS2 (24). Activation of FLS2, which can be stimulated in the laboratory using a synthetic peptide containing the flagellin epitope (flg22), rapidly induces transcription of a wide array of *Arabidopsis* defense genes (approximately 1000) that are likely key regulators of basal defense (24). Furthermore, flg22 treatments are capable of priming the plant against infection. For example, flg22 pre-treated *Arabidopsis* plants, which are then subsequently infected with virulent *P. syringae* pv. *tomato* DC3000 pathogen, are markedly more resistant to the bacteria (24), indicating that these early MAMP defenses are fully capable, when temporally activated, of encoding pathogen resistance. Another well-characterized PRR, the EFR LRR-kinase, non-specifically recognizes the bacterial elongation factor Tu (EF-Tu), and, upon activation, stimulates well-conserved stress pathways to induce a transcriptional response that is almost identical to that of the flg22 response (25).

PRRs, therefore, act in a coordinated and redundant fashion to induce basal defense pathways, ultimately resulting in transcriptional reprogramming; however, this unique activity makes them ideal targets of type III effector proteins. Indeed, a recent study has shown both EFR and FLS2 are targeted at the plasma membrane by the *Pseudomonas syringae* effector protein AvrPto (26). AvrPto recognizes and binds a conserved patch within the intracellular kinase domain of each receptor to suppress auto-phosphorylation activity and largely inhibit MAMP-induced defenses (26). Additionally, the *Arabidopsis* BAK1 protein, a receptor kinase that likely interacts with several PRRs to coordinate defense signaling, is targeted and inhibited by two effectors, AvrPto and AvrPtoB (26). PRRs serve as the initial line of defense in the plant and, in most cases, are sufficient for defense against invading microbes; however, a number of virulent pathogens have evolved sophisticated effector-dependent mechanisms to suppress PRR signaling at the plasma membrane.

Cell surface receptors are not the only targets of type III effectors. Downstream signaling molecules, including several components of the MAPK, salicylic acid (SA), and oxidative burst pathways, represent ideal targets for suppression of basal defense responses (3). Interestingly, the *P. syringae* effector HopAI1 directly inhibits MAPK signaling by irreversibly dephosphorylating the *Arabidopsis* MPK3 and MPK6 proteins using a unique phosphothreonine lyase activity (27). This remarkable activity suppresses flg22-induced transcriptional changes, as well as oxidative burst signaling, and directly enhances the pathogenicity of the bacteria (27). It is likely that additional effector proteins redundantly target these basal defense pathways to promote disease, however, the molecular mechanisms underling these suppressive activities are only beginning to be uncovered.

R Gene Mediated Defense

The most simple model for R protein activation requires a direct protein-protein interaction between the bacterial Avr protein (encoded by an avirulence gene) and the plant host R protein (encoded by a resistance gene) to initiate defense pathways, however, only a few of these “gene-for-gene” interactions have been discovered (28-30). A number of recent studies have supported a model for indirect recognition where an effector biochemically alters an intracellular host protein, which in turn is sensed by a single downstream R protein (14). Therefore, *R* genes are responsible for “guarding” against manipulation of a host protein by an effector Avr protein. It is likely that plants have diversified their individual collections of *R* genes throughout evolution in order to monitor as many host proteins or pathways as possible. Therefore, it is not surprising that *Arabidopsis* carries approximately 125 predicted *R* genes, while rice, which has a much larger genome, possesses approximately 600 individual *R* genes (3). As expected, genetic deletion of the bacterial avirulence gene or the corresponding plant *R* gene results in loss of the recognition event and leads to bacterial virulence.

The archetypal example of the “guard hypothesis” centers on two genetically distinct R proteins, RPM1 and RPS2, which are both responsible for monitoring effector-specific biochemical modification of the *Arabidopsis* protein RIN4. *RPS2* and *RPM1* belong to a large class of *R* genes that encode nucleotide-binding leucine-rich repeat (NB-LRR) proteins. Proteolytic elimination of RIN4 by the *P. syringae* effector AvrRpt2 leads to initiation of RPS2 defenses (31,32), while phosphorylation of RIN4 is induced by the effectors AvrRpm1 or AvrB, and results in activation of RPM1 defense pathways (33). Although a biochemical activity has yet to be assigned to RIN4, recent data have demonstrated that RIN4 is localized to intracellular plasma membranes of the plant cell through a carboxy-terminal acylation (34). Interestingly, both RPS2 and RPM1 proteins co-localize to plasma membranes (31,35) and physically interact with RIN4 (32,33), presumably generating a “primed” protein complex. Remarkably, RIN4 is targeted by three individual effector proteins from different *P. syringae* strains at the plasma membrane, implicating RIN4 in regulation of basal defenses. The R protein-mediated defense system, which is often initiated at the host plasma membrane, efficiently monitors RIN4 perturbations and stimulates a potent resistance response that culminates in induction of the Hypersensitive Response and cessation of pathogen growth.

Conclusions

Pseudomonas syringae strains carry large repertoires of type III effector proteins that, after secretion, redundantly target the plant's initial defense pathways to promote disease. Our understanding of the plant-pathogen interaction is limited by incomplete lists of the bacterial virulence factors and their respective molecular targets, including interactions between R proteins and their Avr protein complements. Approaches like large-scale, protein-protein interaction mapping would provide unprecedented insight into the molecular interactions between the plant and bacteria. Finally, investigation of the temporal regulation of type III effectors by the bacteria would likely provide key information about molecular communication that occurs between the microbe and host during the course of an infection.

REFERENCES

1. Alfano, J. R., and Collmer, A. (2004) *Annu Rev Phytopathol* **42**, 385-414

2. Martin, G. B., Bogdanove, A. J., and Sessa, G. (2003) *Annu Rev Plant Biol* **54**, 23-61

3. Nimchuk, Z., Eulgem, T., Holt, B. F., 3rd, and Dangl, J. L. (2003) *Annu Rev Genet* **37**, 579-609

4. Hoefle, C., and Huckelhoven, R. (2008) *Cell Microbiol* **10**(12), 2400-2407

5. Kumar, Y., and Valdivia, R. H. (2009) *Cell Host Microbe* **5**(6), 593-601

6. Nurnberger, T., Brunner, F., Kemmerling, B., and Piater, L. (2004) *Immunol Rev* **198**, 249-266

7. Thilmony, R., Underwood, W., and He, S. Y. (2006) *Plant J* **46**(1), 34-53

8. Cornelis, G. R., and Van Gijsegem, F. (2000) *Annu Rev Microbiol* **54**, 735-774

9. Chang, J. H., Urbach, J. M., Law, T. F., Arnold, L. W., Hu, A., Gombar, S., Grant, S. R., Ausubel, F. M., and Dangl, J. L. (2005) *Proc Natl Acad Sci U S A* **102**(7), 2549-2554

10. Suss, C., Hempel, J., Zehner, S., Krause, A., Patschkowski, T., and Gottfert, M. (2006) *J Biotechnol* **126**(1), 69-77

11. Roine, E., Wei, W., Yuan, J., Nurmiaho-Lassila, E. L., Kalkkinen, N., Romantschuk, M., and He, S. Y. (1997) *Proc Natl Acad Sci U S A* **94**(7), 3459-3464

12. Scheel, D. (1998) *Curr Opin Plant Biol* **1**(4), 305-310

13. Morel, J. B., and Dangl, J. L. (1997) *Cell Death Differ* **4**(8), 671-683

14. Jones, J. D., and Dangl, J. L. (2006) *Nature* **444**(7117), 323-329

15. Jamir, Y., Guo, M., Oh, H. S., Petnicki-Ocwieja, T., Chen, S., Tang, X., Dickman, M. B., Collmer, A., and Alfano, J. R. (2004) *Plant J* **37**(4), 554-565

16. Cunnac, S., Lindeberg, M., and Collmer, A. (2009) *Curr Opin Microbiol* **12**(1), 53-60

17. Hacker, J., and Kaper, J. B. (2000) *Annu Rev Microbiol* **54**, 641-679

18. He, S. Y., Nomura, K., and Whittam, T. S. (2004) *Biochim Biophys Acta* **1694**(1-3), 181-206

19. Blocker, A., Komoriya, K., and Aizawa, S. (2003) *Proc Natl Acad Sci U S A* **100**(6), 3027-3030

20. van Dijk, K., Tam, V. C., Records, A. R., Petnicki-Ocwieja, T., and Alfano, J. R. (2002) *Mol Microbiol* **44**(6), 1469-1481

21. Guttman, D. S., Vinatzer, B. A., Sarkar, S. F., Ranall, M. V., Kettler, G., and Greenberg, J. T. (2002) *Science* **295**(5560), 1722-1726

22. Petnicki-Ocwieja, T., Schneider, D. J., Tam, V. C., Chancey, S. T., Shan, L., Jamir, Y., Schechter, L. M., Janes, M. D., Buell, C. R., Tang, X., Collmer, A., and Alfano, J. R. (2002) *Proc Natl Acad Sci U S A* **99**(11), 7652-7657

23. Zipfel, C. (2009) *Curr Opin Plant Biol* **12**(4), 414-420

24. Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E. J., Jones, J. D., Felix, G., and Boller, T. (2004) *Nature* **428**(6984), 764-767

25. Zipfel, C., Kunze, G., Chinchilla, D., Caniard, A., Jones, J. D., Boller, T., and Felix, G. (2006) *Cell* **125**(4), 749-760

26. Xiang, T., Zong, N., Zou, Y., Wu, Y., Zhang, J., Xing, W., Li, Y., Tang, X., Zhu, L., Chai, J., and Zhou, J. M. (2008) *Curr Biol* **18**(1), 74-80

27. Zhang, J., Shao, F., Li, Y., Cui, H., Chen, L., Li, H., Zou, Y., Long, C., Lan, L., Chai, J., Chen, S., Tang, X., and Zhou, J. M. (2007) *Cell Host Microbe* **1**(3), 175-185

28. Deslandes, L., Olivier, J., Peeters, N., Feng, D. X., Khounlotham, M., Boucher, C., Somssich, I., Genin, S., and Marco, Y. (2003) *Proc Natl Acad Sci U S A* **100**(13), 8024-8029

29. Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Teh, T., Wang, C. I., Ayliffe, M. A., Kobe, B., and Ellis, J. G. (2006) *Proc Natl Acad Sci U S A* **103**(23), 8888-8893

30. Jia, Y., McAdams, S. A., Bryan, G. T., Hershey, H. P., and Valent, B. (2000) *Embo J* **19**(15), 4004-4014

31. Axtell, M. J., and Staskawicz, B. J. (2003) *Cell* **112**(3), 369-377

32. Mackey, D., Belkhadir, Y., Alonso, J. M., Ecker, J. R., and Dangl, J. L. (2003) *Cell* **112**(3), 379-389

33. Mackey, D., Holt, B. F., 3rd, Wiig, A., and Dangl, J. L. (2002) *Cell* **108**(6), 743-754

34. Kim, H. S., Desveaux, D., Singer, A. U., Patel, P., Sondek, J., and Dangl, J. L. (2005) *Proc Natl Acad Sci U S A* **102**(18), 6496-6501

35. Boyes, D. C., Nam, J., and Dangl, J. L. (1998) *Proc Natl Acad Sci U S A* **95**(26), 15849-15854

CHAPTER 2

A family of bacterial cysteine protease type III effectors utilize acylation-dependent and independent strategies to localize to plasma membranes

ABSTRACT

Bacterial phytopathogens employ a type III secretion system to deliver effector proteins into the plant cell to suppress defense pathways; however, the molecular mechanisms and sub-cellular localization strategies that drive effector function largely remain a mystery. Here, we demonstrate that the plant plasma membrane is the primary site for sub-cellular localization of the *Pseudomonas syringae* effector AvrPphB and five additional cysteine protease family members. AvrPphB and two AvrPphB-like effectors, ORF4 and NopT, auto-proteolytically process following delivery into the plant cell to expose embedded sites for fatty acylation. Host-dependent lipidation of these three effectors directs plasma membrane localization and is required for the avirulence activity of AvrPphB. Surprisingly, the AvrPphB-like effectors RipT, HopC1, and HopN1 utilize an acylation-independent mechanism to localize to the cellular plasma membrane. While some AvrPphB-like effectors employ acylation-independent localization strategies, others hijack the eukaryotic lipidation machinery to ensure plasma membrane localization, illustrating the diverse tactics employed by type III effectors to target specific sub-cellular compartments.

INTRODUCTION

Plants have evolved sophisticated mechanisms to recognize invading bacterial pathogens and, upon infection, can coordinate an extremely efficient defense response. Detection of microbes occurs rapidly through recognition of a variety of nonspecific elicitors, or microbe-associated molecular patterns (MAMPs), that trigger a basal non-specific resistance response that is often sufficient in controlling most invading bacteria (1). However, phytopathogens including *Pseudomonas syringae*, *Xanthomonas campestris*, *Erwinia amylovora*, and *Ralstonia solanacearum* employ a type III secretion system (TTSS) to deliver an arsenal of virulence proteins (effectors) into host cells that suppress MAMP-induced defenses and render the plant susceptible to disease (2). *P. syringae* pv. *tomato* DC3000, for example, secretes approximately 30 effectors into the plant cell that are responsible for defining host specificity and, as a collection, are indispensable for disease progression (3,4).

However, resistant plants have developed mechanisms to defend against effector function and specifically recognize a given pathogen. These plants initiate a potent defense response that is often characterized by a localized programmed cell death reaction, or hypersensitive response (HR), at the site of infection, which often occurs concomitantly with cessation of pathogen growth (5). Although the underlying signaling molecules involved in HR induction are only beginning to be uncovered, it is clear that HR progression is dependent on plant disease resistance (*R*) gene products that specifically recognize bacterial effector avirulence (Avr) proteins (6). The simplest model for initiation of R protein defenses requires a direct protein-protein interaction between the bacterial Avr protein and the host R protein; however, only a handful of these “gene-for-gene” interactions have been uncovered (7-9).

Recent studies also support a model for an indirect recognition event whereby an effector biochemically alters a host protein, which in turn is sensed by a single downstream R protein or multi-protein complex (2). In this case, R proteins are responsible for “guarding” against manipulation of a host protein by an Avr effector protein.

The molecular mechanisms that dictate R protein activation in the large part remain a mystery; however, biochemical and functional data for a few Avr-R protein relationships have underscored the importance of sub-cellular localization in transducing plant defenses. The two archetypal examples of effectors whose functions are defined by specific sub-cellular localizations are the *P. syringae* Avr proteins AvrB and AvrRpm1, which both localize to host plasma membranes where they initiate R protein defenses (10). Interestingly, mislocalization of AvrB or AvrRpm1 abolishes their avirulence activities (10). These data emphasize the importance of proper effector localization in the host cell and allude to the fact that other type III effectors may employ similar strategies to promote their function.

We have previously shown that the *P. syringae* pv. *phaseolicola* effector AvrPphB is a member of the YopT family of cysteine proteases and specifically cleaves the *Arabidopsis* protein kinase PBS1 to initiate *RPS5*-dependent HR (11,12). Consistent with the guard model, PBS1 forms a protein complex with plasma membrane-localized RPS5 in “anticipation” of proteolytic cleavage by AvrPphB (13,14). Upon delivery into the host cell, AvrPphB auto-proteolytically processes to reveal a novel amino-terminus containing putative sites for both *N*-myristoylation and *S*-palmitoylation (10,15). While AvrPphB appears to interact with membranes through the putative myristoylation site (10), there has been no biochemical evidence supporting fatty acylation of AvrPphB, and to date it remains unclear if lipidation of AvrPphB is necessary for cleavage of PBS1 and subsequent HR induction *in planta*.

In this study, we have identified additional AvrPphB family members, utilized by evolutionary diverse phytopathogens, which remarkably possess auto-processing activity. Cleavage, in turn, reveals embedded sites for fatty acylation that are post-transcriptionally modified by the eukaryotic machinery *in vivo*. Consequently, host lipidation of these AvrPphB-like effectors ensures plasma membrane localization. We demonstrate that acylation of AvrPphB is absolutely required for cleavage of PBS1 and induction of *RPS5*-dependent defenses at the plant plasma membrane. Surprisingly, some AvrPphB family members do not auto-process and, in turn, are not acylated. Nonetheless, these effectors localize to plasma membranes using acylation-independent strategies. Together, these studies illustrate the complex tactics employed by type III effectors to localize within specific sub-cellular compartments, thereby enhancing their effective concentrations and likely promoting their biological function.

EXPERIMENTAL PROCEDURES

Plasmids and Pathogen Strains

All PCR based cloning was performed using standard procedures and all point mutations were generated using the QuickChange Site Directed Mutagenesis kit (Stratagene) using manufacturer's instructions. Effector cDNAs were cloned from genomic DNA: *P.s.* pv. *phaseolicola* (ATCC 11355D), *P.s.* pv. *tomato* DC3000 (isolated using standard procedures), *R.s.* GMI1000 (gift from Timothy Denny), and *Rhizobium* sp. NGR234 cosmid pXB740 (16). cDNAs were cloned into the mammalian expression vector pcDNA3.1 (Invitrogen) for *in vitro* transcription/translation experiments, into the yeast expression vector pRS425-GAL (gift from Richard Kolodner) for *S. cerevisiae* experiments, and into the plant 35S CMV expression vector pCHF3-YFP for localization studies (17). *avrPphB* and *orf4* alleles were expressed behind their native promoters in *P. syringae* pv. *tomato* DC3000 (gift from Brian Staskawicz) or *P. fluorescens* pLN1965 (gift from James Alfano) using the broad-host-range plasmid pVSP61 (18). The *P.f.* pLN1965 strain is identical to *P.f.* pLN18 (19) but allows for use the pVSP61 plasmid encoding kanamycin resistance. The pVSP61::*avrPphB* plasmid has been described previously (20). The *orf4* gene (200 bp upstream and 400 bp downstream) was cloned into pVSP61. All pVSP61 plasmids were introduced into *Pseudomonas* strains (grown on KB media at 28°C) via triparental mating using a DH5α helper strain carrying the plasmid pRK2013. The *AtPBS1* genomic clone was isolated from Col-0 genomic DNA as described previously (11), tagged with a 3×HA epitope directly upstream of the stop codon, and cloned into the binary vector pJHA212B. The resulting vector was transformed into *pbs1-1* (21) plants via *Agrobacterium*-mediated floral dipping and T_1 *pbs1-1:PBS1-HA* plants were isolated by Basta selection in soil.

Phylogenetic Analysis of the YopT Family

The YopT family members were identified by PSI-BLAST using the YopT and AvrPphB amino acid sequences as the queries (22). Generation of the YopT phylogenetic tree was performed using the ClustalW server (http://align.genome.jp) using the default parameters and the tree was produced by selecting the "Dendrogram" option.

In vitro Auto-processing Assays

Effectors in pcDNA3.1 were *in vitro* transcribed and translated in wheat germ extract (Promega) in the presence of [^{35}S]methionine (Amersham) according to manufacturer's instructions. At the specified time points, 5 μl of the total reaction volume (50 μl) were removed, quenched in 2× SDS sample buffer, subjected to SDS-PAGE, and analyzed by autoradiography.

In planta and *in vitro* TnT Expression of HopC1 and HopN1

Full-length *hopC1*-2xFlag and *hopN1*-2xFlag effectors were cloned into the dexamethasone-inducible pTA7002 vector (23). The resulting vectors were transformed into *pbs1-1 Arabidopsis* plants via *Agrobacterium*-mediated floral dipping (24) and T_1 plants were isolated by hygromycin selection. Four to five-week-old T_2 plants were sprayed with 20 μM DEX and leaf tissue was harvested 8 hr later. Plant tissue was ground in liquid nitrogen and homogenized in lysis buffer: 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 5 mM DTT, 1% Triton X-100, and 2× plant protease inhibitor cocktail (Sigma). Lysate was cleared by centrifugation at 10,000×g and protein concentrations were determined with Bradford Reagent (Bio-Rad). Total

Protein (50μg) was subjected to SDS-PAGE and western blot analysis using the mouse anti-Flag M2 monoclonal antibody (Sigma, F1804). Additionally, the full-length HopC1-2xFlag and HopN1-2xFlag effector proteins were *in vitro* transcribed and translated (TnT) in wheat germ extract (Promega) according to manufacturer's instructions and samples were analyzed by SDS-PAGE and western blot analysis.

Protein Purification for Edman Degradation

Full-length proteins (MBP-ORF4-His$_6$, NopT-His$_6$, and RipT-Flag) were expressed in *E. coli* (DE3) RIL cells transformed with the appropriate pET vector. All strains were grown in LB media to an OD$_{600}$ of 0.7, then induced with 0.4 mM isopropyl-β-D-thiogalactopyranoside for 12 hr before harvesting the cells at 4°C. Maltose binding protein N-terminally fused to ORF4 (pET28-MBP) was purified from 2 L of culture as previously described (25) except protease inhibitors were removed from all buffers and the bacterial lysate was batch bound to 4 ml amylose resin (NEB) for 2.5 hr at 4°C. After extensively washing the resin, the ORF4-bound beads were moved to room temperature and rotated overnight in the final wash buffer to allow auto-processing of ORF4 from the MBP tag. Eluted protein was passed over a Ni$^+$-agarose (Qiagen) affinity column (1 ml bed volume) as described previously (26). The NopT-His$_6$ protein (pET21a) was purified from 4 L of culture using a Ni$^+$-agarose column (2 mL bed volume) as described above. C-terminally Flag-tagged RipT (pET21a-2xFLAG) was purified from 4 L of bacterial culture. The bacterial pellet was resuspended in lysis buffer (50 mM Tris-HCl pH 8, 250 mM NaCl, 5 mM DTT, 10% glycerol, 1mM EDTA) and cells were lysed by a French pressure cell. RipT-Flag protein was batch purified from lysate using 0.5 mL of anti-Flag M2 Agarose resin (Sigma). After 6 hr of binding to the resin at 4°C, the anti-Flag M2 Agarose was washed 5 times with lysis buffer before elution with 100 μg/ml Flag peptide (Sigma) dissolved in lysis buffer. All proteins were concentrated with Amicon Ultra concentrators (Millipore), subjected to SDS-PAGE, and transferred to a PVDF membrane. The membrane was stained with 0.1% Coomassie Blue R (Sigma) in 50% methanol, destained with 50% methanol/10% acetic acid (2-3 washes), washed twice with water, and air-dried. Edman sequencing was performed by the University of California at San Diego Protein Sequencing Facility.

In vitro Myristoylation Assays

For *in vitro* *N*-myristoylation labeling experiments, proteins were generated by TnT in the presence of either [^{35}S]methionine to demonstrate protein synthesis or [^{3}H]myristic acid to demonstrate myristoylation. Briefly, [^{3}H]myristic acid (NEN) in ethanol was dried down via vacuum and resuspended at 1 μCi/μL in wheat germ extract before addition to the TnT reactions according to manufacturer's instructions. After a 2 hr incubation at 30°C, samples were quenched in 2× SDS sample buffer, subjected to SDS-PAGE (10% volume of ^{35}S-labeled samples, 20% volume of ^{3}H-labeled samples), and analyzed by autoradiography.

Yeast *In vivo* Labeling

In vivo labeling of *S. cerevisiae* was performed as previously described with modifications (27). Briefly, C-terminally 2×Flag tagged effectors were expressed under a galactose-inducible promoter (pRS425-GAL) in the protease-deficient RDKY1293 strain (MATα, *ura3-52*, *trp1Δ63*, *leu2Δ1*, *his3Δ200*, *pep4*::*HIS3*, *prb1Δ1.6R*, *can1*, *GAL*; gift from R. Kolodner). Yeast strains grown to mid-log phase in YPAD were diluted (5:50 ml) into complete minimal media containing 3% raffinose and grown to stationary phase. Approximately 1.7×10^{9} cells were then resuspended in 100 ml of rich media (1% Bacto Yeast Extract, 2% Bacto Peptone, and 4% galactose) and grown for 4 hr at 30°C. Cells (~1×10^{9}) were resuspened in 25 ml of rich media containing 3% galactose, 3 μg/ml cerulenin (Sigma), and 30 μCi/ml [^{3}H]myristic acid or 50 μCi/ml [^{3}H]palmitic acid (NEN). Yeast were labelled for 4 hr at 30°C before harvesting. Cell lysis and protein immunoprecipitation has been described previously (27). For this study, pre-equilibrated anti-Flag M2 Agarose resin (100 μl) was added to the Protein A Agarose (Invitrogen) precleared lysate and rotated overnight at 4°C. The immunoprecipitations were washed extensively with lysis buffer and eluted with 2× SDS sample buffer. Samples were subjected to SDS-PAGE and the gel was treated with Amplify (Amersham) to enhance the tritium signal before drying. The gel was analyzed by autoradiography ([^{3}H]myristic acid, 1 month exposure; [^{3}H]palmitic acid, 10 month exposure). To determine protein expression, identical samples were analyzed by western blotting with an anti-Flag M2-Peroxidase conjugate antibody (Sigma).

Total Yeast Membrane Fractionation and Chemical Treatments

Strains were grown to saturation in YPAD before dilution (1:10 ml) into complete minimal media containing 2% raffinose and 0.25%-2% galactose (concentration was varied based on protein expression levels). Cells were grown for 8 hr and ~1.7×10^{8} cells were harvested. Cells were resuspended in 300 mM sorbitol, 100 mM NaCl, 5 mM $MgCl_2$, 10 mM Tris-HCl, pH 7.4, and protease inhibitor tablets (Roche) before addition of acid-washed glass beads (Sigma). Cell disruption was carried out by vigorous vortexing and cell debris was removed by centrifugation at 500×g for 10 min at 4°C, giving the “total” fraction. The soluble fraction was separated from the insoluble membrane fraction by ultracentrifugation at 100,000×g for 1 hour. The membrane pellet was resuspended in lysis buffer containing 1% Triton X-100 to completely solubilize the membranes. Equal volumes (80 μl) of each fraction were added to 20 μl of 5× SDS sample buffer, subjected to SDS-PAGE, and analyzed by western blotting (anti-Flag M2, Sigma F1804). A duplicate blot was probed with a monoclonal anti-v-H-Ras antibody (Oncogene Research Products) to detect the yeast plasma membrane marker Ras1p.

Total membranes for chemical treatment experiments were isolated as described above except total lysate was split in 5 equal volumes before ultracentrifugation. Membrane samples were resuspended in either lysis buffer, 1 M NaCl/10 mM Tris-HCl pH 7.4, 2 M urea/10 mM Tris-HCl pH 7.4, 0.1 M Na_2CO_3 pH 11.5, or 1% Triton X-100/10 mM Tris-HCl pH 7.4 and incubated for 1 hr on ice. Samples were re-fractionated by ultracentrifugation and equal volumes of soluble and insoluble fractions were subjected to SDS-PAGE and western blot analysis.

Total Plant Membrane Fractionation

Full-length wild-type effectors (*ripT*-2xFlag*, hopC1*-2xFlag*,* and *hopN1*-2xFlag) were cloned into the DEX-inducible pTA7002 vector (23). The RipT-2xFlag and HopC1-2xFlag proteins were transiently expressed in five-week-old Col-0 plants by *Agrobacterium*-mediated delivery as previously described (10). DEX-induced (20 μM) leaf tissue was harvested 8 hr after treatment. The T_2 DEX::*hopN1*-2xFlag transgenic plants (*pbs1-1* background) were sprayed with 20 μM DEX and leaf tissue was collected after 8 hours. Fractionation of *Arabidopsis* total membranes was performed as previously described (10). Equal volumes of the total, soluble, and insoluble fractions were subjected to SDS-PAGE, blotted, and probed with either an anti-Flag

M2 monoclonal antibody (Sigma) or an anti-H+-ATPase antibody (plasma membrane marker, gift from Maarten Chrispeels).

Transformation of Chinese Cabbage and Microscopy

Chinese cabbage (*Brassica campestris* subsp. *napus* var. *pekinensis*) leaf slices were transformed by particle bombardment using a Biolistic PDS-1,000/He particle delivery system (Bio-Rad Laboratories). Gold particles (1.0 μm) were coated with the individual 35S::*effector*-YFP (pCHF3-YFP) or 35S::*AtPIP2A*-CFP (pCHF1-CFP) plasmids according to manufacturer's instructions. The tissue was bombarded twice using 1,100 psi rupture discs under a vacuum of 26 in Hg. After a 9 hr incubation at 22°C, epidermal peels were imaged using a Zeiss Axiovert microscope (Carl Zeiss Microimaging, Inc.) equipped with a MicroMax digital camera (Roper-Princeton Instruments) controlled by MataFluor software (Universal Imaging, Corp.).

Stable Expression of the AvrPphB Acylation-Deficient Mutant in *Arabidopsis* and HR Assays

The *avrPphB(Δ62)GC/AS*-2xFlag allele was cloned into the DEX-inducible pTA7002 vector (23). This vector, as well as the empty pTA7002 vector, were transformed into Col-0 *Arabidopsis* plants by *Agrobacterium*-mediated floral dipping (24), and T_1 plants were selected for using hygromycin as described above. Four-week-old T_2 plants were sprayed with 20 μM DEX and the plants were photographed 16 hr later. Two independent transgenic plant lines were assayed and produced similar results.

Plant HR and Bacterial Growth Assays

Arabidopsis thaliana plants were grown in a Promix-HP:vermiculite (2:1) soil mix under a 9 hr photoperiod at 22°C. HR assays were performed in 4-6 week old plants by syringe infiltration of bacteria (~3.75×10^{7} cfu ml^{-1}) as previously described (11). Plants were photographed and scored for HR 16-20 hpi for *P.s.* pv. *tomato* DC3000 treated plants or 45 hpi for *P. fluorescens* treated plants. *P. syringae* growth assays were performed by dipping two-week-old seedlings in bacteria (~$2.\ 5\times10^{7}$ cfu ml^{-1}) exactly as previously described (28). Data is represented as the mean and standard error of the decimal logarithm (log[cfu mg^{-1} fresh weight]) of four replicates.

In planta PBS1 Cleavage Assay

T_2 *pbs1-1:PBS1-HA* plants were inoculated with *P. syringae* (3.75×10^{7} cfu ml^{-1}) strains by syringe infiltration. Tissue from three independent plants was harvested 14 hpi and homogenized in lysis buffer containing 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 5 mM DTT, 1% Triton X-100, and 2× plant protease inhibitor cocktail (Sigma). Lysate was cleared by centrifugation at 10,000×g for 5 min and protein concentrations were determined with Bradford Reagent (Bio-Rad). Total Protein (10μg) was subjected to SDS-PAGE and western blot analysis using an anti-HA.11 monoclonal antibody (Covance).

In vitro Cleavage Assays

Recombinant proteins (pET21: AvrPphB(Δ62)-His_6 and pET21: ORF4(Δ111)-His_6) were purified in the absence of protease inhibitors from *E. coli* (DE3) RIL cells using Ni^{+}-agarose affinity chromatography as described above. *In vitro* cleavage assays of rabbit reticulocyte-generated (TnT) PBS1 were performed using recombinant AvrPphB(Δ62)-His_6 (0.2 μg/μL) or ORF4(Δ111)-His_6 (0.2 μg/μL) proteins as previously described (11).

Secretion of AvrPphB Effectors by *P. syringae*

P. syringae strains were grown in KB media to mid-log phase before resuspending the bacteria at an OD_{600} of 0.3 in Hrp-inducing minimal media (pH 6.0) containing 10 mM fructose as previously described (29). An overnight-induced (22°C) 40 mL culture was centrifuged (4,300×g, 15 min) and 20 mL of the supernatant was re-centrifuged for 40 min at 17,200×g. 10 mL of the resulting supernatant was removed and protein was precipitated with 11.5% TCA, washed with acetone, and resuspended in 2× SDS sample buffer. Cell-bound fractions (bacterial pellet) and secreted fractions (supernatant protein, 7.5x concentrated) were subjected to SDS-PAGE and probed with a polyclonal anti-AvrPphB antibody (whole serum, 1:10,000) or a polyclonal anti-Neomycin Phosphotranserase II antibody (Upstate). Anti-sera was generated against recombinant AvrPphB(Δ62)-His_6 protein (11) in rabbit (Cocalico Biologicals).

Induction of *PR1* Expression

Two-week-old plants were infected by dipping (~2. 5×10^7 cfu ml^{-1}) and aerial tissue from 4-5 plants (one biological replicate) was collected 24 hpi. RNA was isolated using the RNeasy plant mini kit (Qiagen) and cDNA was generated from total RNA (1μg) using the SuperScript III kit (Invitrogen) and oligo(dT) primers according to manufacturer's instructions. qPCR reactions were run on a MX4000 Multiplex QPCR machine (Stratagene) using Power SYBR Green PCR Mastermix kit (Applied Biosystems). Primer pairs for *PR1* (target) and *TUA3* (endogenous control) have been previously described (30). Ct values were generated using default parameters and relative expression values were calculated using the formula $2^{-((\mathrm{Ct}_{\mathit{PR1}}\mathrm{Treatment}\ -\ \mathrm{Ct}_{\mathit{TUA3}}\mathrm{Treatment})\ -\ ((\mathrm{Ct}_{\mathit{PR1}}\mathrm{Mock}\ -\ \mathrm{Ct}_{\mathit{TUA3}}\mathrm{Mock}))}$. Data presented is the mean and standard error of the fold change in *PR1* transcript compared to mock (no bacteria) of at least 5 biological replicates (comprised of three technical replicates) from two independent experiments.

RESULTS

Identification of an AvrPphB-Like Effector Subfamily within the YopT Family of Cysteine Proteases

Bioinformatic analyses of the YopT family suggest that more than 30 evolutionary diverse bacterial organisms utilize putative cysteine protease virulence factors to promote disease in animal, marine, or plant species (Figure 2.7.A; Shao *et al.* (12)). Using PSI-BLAST (22), we searched for novel AvrPphB family members from recently sequenced plant pathogens or symbiotes that were identical to AvrPphB at the C/H/D catalytic residues, as well as the invariant residues W105 and P228 (numbered from the AvrPphB sequence). Thirteen sequences from different bacterial strains were identified, and the full-length proteins were aligned using the BLOSUM matrix. Although residues surrounding the catalytic amino acids are heavily conserved, similarity outside of these regions is extremely limited (Figure 2.1). Surprisingly, all AvrPphB-like effector proteins are predicted to have a secondary structure (http://bioinf.cs.ucl.ac.uk/psipred/) similar to that of the known AvrPphB structure (31), suggesting that these proteins have been evolutionarily tailored by the bacteria to maintain cysteine protease activity, but likely target different host proteins. Surprisingly, we also identified a conserved patch of residues in the amino terminus of the AvrPphB family members that comprise an embedded consensus site for eukaryotic fatty acylation (Figure 2.2.A), implying that proteolytic processing of these effectors at specific residues may generate functional eukaryotic lipidation motifs.

AvrPphB Family Members Self-Process at Specific Residues within the *N*-Terminus

The *P.s.* pv. *phaseolicola* effector AvrPphB is a 35 kDa protein that self-proteolytically processes into a 28 kDa mature protein, requiring the C/H/D catalytic triad for this unique activity (Figure 2.2.A; Shao *et al.* (12), Puri *et al.* (15)). We examined five additional AvrPphB family members (Figure 2.2.B), from evolutionary diverse pathogens, for self-processing activity using an *in vitro* cleavage assay. Interestingly, three other AvrPphB-like effectors (ORF4, NopT, and RipT) self-cleaved into approximately 28 kDa proteins when they were expressed in wheat germ extract (Figure 2.3.A). Auto-proteolytic processing requires the catalytic cysteine, suggesting that the *orf4*, *nopT*, and *ripT* genes encode functional cysteine proteases. Additionally, we observed a more rapid processing of AvrPphB compared to the other effectors in the context of this assay, suggesting that the structure of AvrPphB or the chemical composition of the internal cleavage site is better suited for auto-processing activity compared to that of ORF4, NopT, and RipT. Surprisingly, the two effectors HopC1 and HopN1 do not auto-process *in vitro*. Furthermore, self-processing was not detected by western blot analysis when HopC1 or HopN1 were expressed in *S. cerevisiae* (Figure 2.8.A) or in *Arabidopsis thaliana* (Figure 2.4), eliminating the possibility that a eukaryotic “activator” is required for processing.

In order to determine the sites of auto-proteolytic cleavage, we expressed full-length, C-terminally epitope-tagged proteins in *E. coli*, purified the proteins from lysates, and determined their cleavage sites by Edman degradation. ORF4, NopT, and RipT were efficiently processed in bacteria and sequencing revealed that self-cleavage occurs prior to a glycine found in the P1′ position (Figure 2.3.B). Interestingly, self-cleavage of AvrPphB, as well as cleavage of its substrate, PBS1, occurs proximal to a GDK motif that is found in both sequences (Figure 2.3.C). Mutation of all three GDK residues in PBS1 completely inhibits cleavage (11), suggesting that the P1, P2, and P3 residues may be important for effector self-processing. We found that triple mutation of the P1-3 residues in AvrPphB, ORF4, NopT, and RipT prevents self-cleavage (Figure 2.3.D), indicating that these residues are required for recognition and subsequent auto-proteolysis of the amino-terminus.

Self-Proteolysis of AvrPphB-Like Effectors Exposes Post-Translational Lipid Modification Sites

It has been proposed that auto-proteolytic processing of AvrPphB generates sites for fatty acylation (10); however, there is no direct biochemical evidence supporting *N*-myristoylation or *S*-palmitoylation of AvrPphB or any additional family members. We aligned the amino termini of the processed AvrPphB-like effectors and found conserved glycine (P1′) and serine (P5′) residues in AvrPphB, ORF4, and NopT (Figure 2.5.A) that are consistent with the myristoylation consensus sequence (32). All three effectors also possess potential sites for cysteine palmitoylation. Interestingly, RipT, as well as the non-processed effectors HopC1 and HopN1, lack amino terminal acylation consensus sites and are therefore unlikely to be lipidated.

To determine if the AvrPphB family members are *N*-myristoylated, we *in vitro* transcribed and translated full-length effector proteins in the presence of ^{3}H-myristic acid. Radiolabeled myristate was efficiently incorporated into AvrPphB, ORF4, and NopT self-processed proteins (Figure 2.6). Myristoylation of these effectors require auto-processing activity to expose the embedded myristoylation site, as well as the P1′ glycine modification site, since the C/S and G/A mutants are not lipidated (Figure 2.6). To examine myristoylation of AvrPphB family members in a cellular system, we expressed the effectors in *S. cerevisiae* in the presence of ^{3}H-myristic acid. Consistent with the *in vitro* studies, AvrPphB, ORF4, and NopT are myristoylated in yeast and acylation is dependent on the P1′ glycine residue (Figure 2.5.B). RipT, HopC1, and HopN1 do not possess myristoylation consensus sites and were not modified. In eukaryotic systems *N*-myristoylation of proteins often occurs concomitantly with *S*-palmitoylation of nearby cysteines, a post-translational modification that enhances membrane association of lipidated proteins (33). To determine if AvrPphB family members can be palmitoylated by the eukaryotic machinery, we expressed the effectors in *S. cerevisiae* in the presence of ^{3}H-palmitic acid. AvrPphB, ORF4, and NopT, which each possess cysteines proximal to the myristoylation site, are palmitoylated in yeast (Figure 2.5.C). However, the myristoylation-deficient mutants (GC/AA) are likewise not palmitoylated. These mutants are likely not palmitoylated due to either loss of the myristoyl moiety that often initiates subsequent palmitoylation or mutation of the cysteine modification sites. Interestingly, the acylated effectors generally lack any additional putative *S*-palmitoylation sites outside of the *N*-terminal motif (AvrPphB, 3 additional cysteines; ORF4, 1 cysteine; NopT, 0 cysteines; Figure 2.1), and all additional cysteines are positioned in the catalytic core and are unlikely candidates for lipidation. Therefore, it is probable that palmitoylation occurs proximal to the myristoylation sites found in AvrPphB, ORF4 and NopT. These data represent the first direct biochemical evidence for dual acylation of AvrPphB family members and suggest that lipidation by the eukaryotic host machinery may play an important role in effector function.

Interestingly, the AvrPphB family members are not uniformly auto-processed or lipidated. To investigate if the acylated effectors are evolutionary distinct from the non-acylated effectors, we generated a YopT phylogenetic tree and searched for trends in the auto-processing or lipidation phenomena (Figure 2.7.A). The acylated proteins (AvrPphB, ORF4, and NopT) cluster into a common clade that is distinct from the non-acylated effectors (RipT, HopC1, and HopN1), suggesting that an evolutionary division from a common protease ancestor may have given rise to the lipidation feature. Alternatively, the auto-processing activity of HopC1 and HopN1 and the lipidation sites in RipT may have been lost in order to redirect localization of these effectors in their respective hosts. We scanned all the remaining untested AvrPphB family members for embedded myristoylation consensus sites, and found putative sites for both *N*-myristoylation and *S*-palmitoylation in the sequences of four additional effectors (Blr2058, Blr2140, HopAW1, and YP_272236 which is identical to ORF4; Figure 2.7.B). Although these putative lipidation sites lack experimental validation, it is notable that these effectors phylogenetically cluster with the known acylated effectors (Figure 2.7.A).

The Acylated and Non-Acylated AvrPphB-Like Effectors are Differentially Associated with the Plasma Membrane

Traditionally, dual acylation of eukaryotic proteins with myristoyl and palmitoyl moieties directs proteins to cellular membranes, often plasma membranes, where they are oriented into the cytoplasmic face of the lipid bilayer. To test if auto-processing and subsequent lipidation promotes membrane attachment of AvrPphB family members, we expressed the effectors in *S. cerevisiae* and performed biochemical sub-cellular fractionation experiments. The acylated effectors AvrPphB, ORF4, and NopT co-fractionate with yeast membranes and the farnesylated Ras1p plasma membrane marker (Figure 2.8.A). Furthermore, these associations require functional acylation sites since the GC/AA mutants localize exclusively to the soluble fraction. We also observed significantly higher expression levels of AvrPphB protein, but a smaller proportion of membrane-associated AvrPphB protein, compared to the other acylated effectors. These data, as well as the modest amounts of ^{3}H-myristic acid and ^{3}H-palmitic acid incorporated into AvrPphB (Figure 2.5.B, 2.5.C), indicate that lipidation of AvrPphB in *S. cerevisiae* occurs slowly compared to ORF4 and NopT. Differential lipidation rates between effectors are likely a consequence of the chemical context of the residues surrounding the acylation sites and the substrate selectivity of the eukaryotic acylation machinery; however, it is clear that AvrPphB, ORF4, and NopT are all capable of being acylated at the GC motif and are subsequently localized to cellular membranes.

Surprisingly, the non-acylated effectors RipT, HopC1, and HopN1 also fractionate to the insoluble membrane fraction. We observed localization of the RipT P1′ glycine mutant (G65A), as well as the HopC1 and HopN1 catalytically inactive mutants (C/S), exclusively in the membrane fractions, further substantiating that these effectors associate with membranes independent of lipidation. To eliminate the possibility of artifacts of the yeast expression system, we expressed RipT, HopC1, and HopN1 in *Arabidopsis* and performed similar membrane fractionation experiments. Identical results were observed *in planta* (Figure 2.9), indicating that the non-acylated AvrPphB-like effectors are likely *bona fide* membrane proteins.

The membrane fractionation experiments indicate that the acylated and non-acylated AvrPphB-like effectors employ different mechanisms for membrane association and possibly possess different membrane binding affinities. To evaluate the effector-membrane association, we treated yeast membranes with high salt, denaturing, alkaline, or detergent-containing buffers. As expected, the acylated effectors can only be extracted with detergent as exemplified by the farnesylated Ras1p protein (Figure 2.8.B), suggesting lipidation is the predominant component responsible for membrane association. In contrast, the non-acylated effectors are partially extracted with urea, alkaline buffer, and detergent. While the mechanisms of membrane attachment for the non-lipidated effectors remains unclear, these data suggest that the AvrPphB-like effectors utilize different strategies to localize to membranes and posses different membrane binding affinities.

An overwhelming majority of dual acylated eukaryotic proteins are preferentially localized to plasma membranes (PM) rather than endomembranes (33). To further investigate the cellular localization of the AvrPphB family members, we transiently expressed yellow fluorescent protein (YFP) tagged effectors in Chinese cabbage epidermal cells and examined localization by fluorescence microscopy. AvrPphB, ORF4, and NopT exhibit a clear plasma membrane localization that is indistinguishable from the *Arabidopsis* PM marker PIP2A (Figure 2.10). In contrast, expression of the acylation-deficient mutants generates an unmistakable cytoplasmic localization that is identical to the soluble YFP control staining. Consistent with the biochemical fractionation data, RipT is largely enriched in plasma membranes via an acylation-independent mechanism, since localization of the G65A mutant is identical to that of the wild-type protein. Additionally, the non-lipidated effectors HopC1 and HopN1 are enriched in the plasma membranes of Chinese cabbage cells (Figure 2.10). Interestingly, we observed a unique punctate staining of HopN1 in the plasma membranes of Chinese cabbage cells, as well as tobacco epidermal cells (data not shown), that was absent in the additional effectors and the PIP2A PM marker (Figure 2.10, inset panels), suggesting that HopN1 may target to a lipid microdomain. It is notable that we observed strong nuclear staining in a large proportion of cells expressing both the wild-type and mutant effectors; however, we also observed this phenomena in cells expressing the known plasma membrane protein PIP2A (Figure 2.10), suggesting that over-expression of proteins in this system likely results in nuclear localization artifacts. Although we cannot completely eliminate the possibility that some or all the AvrPphB-like effectors localize to endomembranes at low levels, additional sucrose gradient purification of plasma membranes from *S. cerevisiae* crude membrane fractions revealed that all the wild-type effectors are strongly enriched in the plasma membrane (data not shown). Collectively, these data implicate the plasma membrane as a crucial site for localization of AvrPphB family members and suggest that proper plasma membrane localization may be important for directing effectors to their respective substrates.

The Acylated AvrPphB-Like Effectors Possess Distinct Substrate Specificity

The lipidated AvrPphB-like effectors employ identical strategies to ensure plasma membrane localization; however, it is unknown if localization alone is sufficient to direct substrate specificity. AvrPphB proteolyticly cleaves the *Arabidopsis* PBS1 protein to initiate HR (11,13), and it is possible that additional AvrPphB-like effectors target PBS1. To investigate if the acylated effectors are functionally equivalent, we exogenously expressed ORF4, which has the highest similarity to AvrPphB among all the family members (processed proteins: 45% similar, 27% identical), under control of its native promoter in the plant pathogen *P. syringae* pv. tomato DC3000 (*Pst*) and verified its expression by RT-PCR and western blot analysis (data not shown). We inoculated resistant *Arabidopsis* plants with the avirulent *Pst*(*avrPphB*) strain or the *Pst*(*orf4*) strain at high bacterial densities in order to produce a visually scorable HR-associated tissue collapse. The virulent *Pst*(empty vector) pathogen produces no HR 20 hours post infection; however, *Pst*(*avrPphB*) induces a striking tissue collapse phenotype in 93% of the infected leaves (Figure 2.11.A). Interestingly, strains carrying *orf4* fail to generate a HR. To ensure that the endogenous repertoire of *Pst* effectors is not interfering with ORF4 function, we also performed HR assays using *P. fluorescens* (*Pf*) strains carrying the same effector alleles. Consistent with the *Pst* infections, the *Pf*(*avrPphB*) strain, but not *Pf*(*orf4*), generated a weak, but reliable HR (37 of 59 infected leaves, Figure 2.11.A). Furthermore, recombinant ORF4 protein has no activity against PBS1 in an *in vitro* cleavage assay (Figure 2.12). Together, these data demonstrate that plasma membrane targeting alone is not sufficient to cleave PBS1, and suggest that the acylated AvrPphB-like effectors possess different substrate specificity.

Dual Acylation of AvrPphB is Required for Cleavage of PBS1 and Initiation of Defenses in Resistant *Arabidopsis* Plants

Our sub-cellular localization studies provide strong evidence that AvrPphB is driven to the host plasma membrane by eukaryotic acylation. Interestingly, we have also observed lipidation of PBS1 and RPS5 (data not shown), and additionally, RPS5 associates with *Arabidopsis* membranes (14). Together, these data suggest that PBS1 likely co-localizes with RPS5 at the plasma membrane to guard against AvrPphB; however, it remains unclear if acylation of AvrPphB is required for cleavage of PBS1 and subsequent HR induction.

Cleavage of PBS1 by AvrPphB can be observed *in planta* when both components are over-expressed (11); however, over-expression of the soluble AvrPphB mutant protein in transgenic plants (DEX::*avrPphB GC/AS*) results in an acylation-independent HR that is likely due to overwhelming expression levels in the plant cell (Figure 2.13). To circumvent these over-expression artifacts and ensure proper cellular localization, we performed *in planta* PBS1 cleavage experiments at near endogenous expression levels using *Pst*-delivered AvrPphB proteins and transgenic plants carrying a *PBS1-HA* genomic clone. Partial cleavage of PBS1 occurs in plants inoculated with the *Pst*(*avrPphB*) strain; however, the PBS1 protein is unaffected when strains delivering the acylation-deficient effectors are used (G63A, C64S, and GC/AS, Figure 2.11.B). To ensure that the acylation-deficient mutant is only impaired in sub-cellular localization and not the intrinsic protease activity, we performed *in vitro* cleavage assays and found the AvrPphB GC/AS mutant to be equally efficient as the wild-type protein in cleaving PBS1 (Figure 2.12).

To determine if lipidation of AvrPphB is required for efficient HR induction, we inoculated resistant (Col-0) or susceptible (*pbs1-1*) plants with *Pst* strains carrying the wild-type or mutant alleles and performed HR assays. AvrPphB function is severely reduced in resistant plants by mutation of either the myristoylation site (G63A, 32% responding) or palmitoylation site (C64S, 27% responding) when compared to the wild-type protein (25 of 27 leaves or 93% responding, Figure 2.11.C). Mutation of both acylation sites thoroughly diminishes the avirulence activity (13% responding). The AvrPphB mutants were markedly deficient in their ability to generate HR despite the fact that they were all properly delivered through the TTSS as full-length proteins (Figure 2.11.D).

To further examine the role of acylation in promoting the avirulence function of AvrPphB, we tested the ability of the AvrPphB acylation-deficient mutants to suppress growth of the virulent *Pst* DC3000 strain in *Arabidopsis*. As expected, expression of AvrPphB in *Pst* ensures avirulence and limits bacterial growth in resistant (Col-0), but not susceptible (*pbs1-1*) plants (Figure 2.11.E). The acylation-deficient single and double mutants, however, are all defective in avirulence function since these strains grow in Col-0 to similar levels as the virulent *Pst*(empty vector) strain.

Local and systemic defense against the virulent *Pst* pathogen, as well as some avirulent pathogens, requires accumulation of salicylic acid (SA) and modulation of SA-responsive genes for a maximal resistance response (34). Therefore, we examined expression levels of the SA-inducible gene *PR1* (pathogenesis-related gene 1) in plants 24 hours after infection with *Pst* strains carrying the wild-type or acylation-deficient *avrPphB* alleles. Inoculation of resistant plants with the *Pst*(*avrPphB*) strain results in a *PBS1*-dependent five-fold increase in *PR1* gene expression relative to mock treated plants; however, delivery of the acylation-deficient double mutant by *Pst* results in *PR1* induction levels that are equivalent to those generated by the virulent *Pst*(empty vector) strain (Figure 2.11.F). These data indicate that acylation of AvrPphB is essential for up-regulation of *PR1* transcript. Using a variety of genetic and biochemical approaches to examine multiple aspects of the AvrPphB resistance response, we have unambiguously shown that host acylation of AvrPphB drives cleavage of PBS1 and subsequent HR induction in the plant cell.

DISCUSSION

Phytopathogens inject an arsenal of type III effectors into the host cell to thwart defenses and promote disease; however, the molecular strategies that are employed by effectors to target plant signaling components remain largely unknown. Here, we demonstrated that four effectors from the AvrPphB family of cysteine proteases possess a unique auto-proteolytic processing activity. Self-cleavage, in turn, reveals embedded consensus sites for eukaryotic acylation. We demonstrated that AvrPphB, ORF4, and NopT are indeed *N*-myristoylated, as well as *S*-palmitoylated by the eukayotic host machinery, consequently directing them to the plasma membrane (Figure 2.14). Furthermore, host-dependent acylation of AvrPphB is necessary for its avirulence activity, and it is likely that lipidation of ORF4 and NopT is indispensable for their function as well. We have also shown that RipT, HopC1, and HopN1 are not lipidated by the host machinery; however, they are nonetheless directed to the plasma membrane where they likely disrupt host defense signaling networks (Figure 2.14). Although the molecular targets of ORF4, NopT, and RipT are unknown, it is possible that substrate specificity can be partially inferred from the amino acid context of the auto-processing sites. For example, auto-processing of AvrPphB, as well as cleavage of PBS1, occurs proximal to a GDK motif, suggesting that it may be possible to define the substrate specificities, and putative molecular targets, for ORF4, NopT, and RipT based on the three residues that we identified as essential for auto-processing (Figure 2.14). However, bioinformatic approaches to identify specific *in planta* substrates for these effectors are restricted by the size and chemical makeup of the auto-processing motif and additional information about the auto-processing specificity will be required to generate an experimentally testable substrate pool.

We propose that auto-processing of AvrPphB, ORF4, NopT, and RipT occurs within the plant cell following delivery of the full-length proteins through the TTSS. Supporting this hypothesis, we identified key signatures of the type III secretion signal within the full-length effector sequences but not the auto-processed sequences (35,36). Additionally, we observed preferential secretion of the full-length AvrPphB protein by *Pst* grown in culture. Auto-processing, however, does not exclusively occur within the plant cell since we observed self-cleavage of these effectors in *E. coli*, indicating that a eukaryotic “activator” is not required for this activity. Therefore, these proteins comprise an effector protease family unique to phytopathogens that is mechanistically distinct from the only other biochemically validated cysteine protease effector AvrRpt2, which requires modification by the eukaryotic peptidyl-prolyl isomerase cyclophilin for activation and subsequent self-processing (37). Surprisingly, two effectors, HopC1 and HopN1, were not capable of self-proteolysis. Although protease activity has yet to be ascribed to HopC1, HopN1 possess *in vitro* protease activity, as well as HR suppression activity in tobacco, both of which require the catalytic triad (38). These two effectors therefore represent an evolutionary distinct non-processing, yet catalytically active, class of cysteine protease effectors within the AvrPphB family.

We clearly demonstrated that auto-processing of AvrPphB, ORF4, and NopT results in fatty acylation of these effectors by the eukaryotic lipidation machinery. Interestingly, these three auto-processed effectors, but not RipT, are predicted to be myristoylated using a variety of eukaryotic prediction models (32,39,40), suggesting that these bacterially-generated acylation sites have been engineered to conform to the restraints of the plant acylation machinery. It is possible that myristoylation of additional AvrPphB family members can be predicted according to these parameters using the following generalized consensus motif: GX_2XXS, where X_2 is a non-acidic residue. Futhermore, genetic experiments suggest that additional effectors may be acylated by the host machinery including AvrPto (41), HopF2 (42), XopE1/XopE2/XopJ (43), and multiple HopZ alleles (44), while several more contain putative consensus sites for acylation (45). Prior to our study, however, *in vivo* biochemical evidence supporting host-dependent acylation of effector proteins was limited to *N*-myristoylation of only two *P. syringae* effectors, AvrRpm1 and AvrB (10). We have identified three additional effectors that are myristoylated *in vivo* and have provided the first direct biochemical evidence for modification of effectors by *S*-palmitoylation. Although an overwhelming majority of protein myristoylation is believed to occur co-translationaly at the ribosome where *N*-myristoyltransferases are enriched (46), there is compelling evidence for non-ribosomally associated myristoylation: the mammalian protein BID is myristoylated at an embedded acylation site following proteolytic cleavage by caspase 8 (47). Therefore, the AvrPphB-like effectors are likely myristoylated independently of the ribosome-associated *N*-myristoyltransferases, resulting in a weak plasma membrane association that can be fully stabilized through *S*-palmitoylation of the effectors by plasma membrane-localized palitoyltransferases (48,49). Additionally, palmitoylation is a reversible lipid modification that can be dynamically regulated in the eukaryotic cell. We suspect that active depalmitoylation of the acylated effectors by acyl-protein thioesterases *in planta* would likely disrupt the effector-membrane associations and attenuate function.

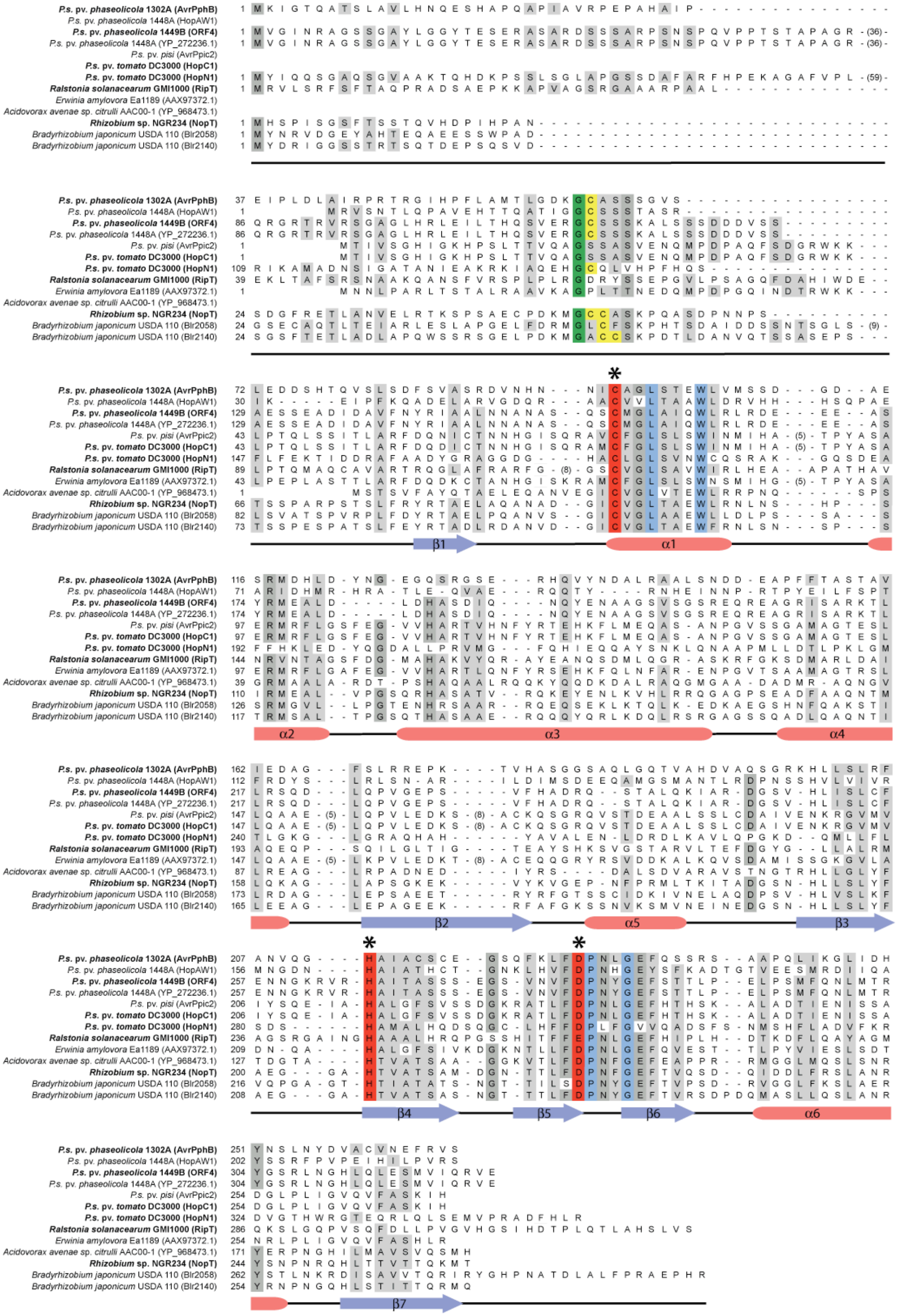
Previous studies examining the role myristoylation plays in promoting the avirulence function of AvrPphB have provided somewhat conflicting results (10,50). Nimchuk and colleagues demonstrated that the putative myristoylation site in

AvrPphB is required for maximal induction of HR when transiently expressed in *Arabidopsis*; however, an AvrPphB myristoylation-independent HR has also been observed in different plant species using *Agrobacterium* and viral over-expression systems (50). While it is possible that there are host-specific differences in R protein recognition of AvrPphB, it seems likely that over-expression of AvrPphB results in loss of the myristoylation dependence due to high protein concentrations in the plant cell. We have also observed this phenomenon in transgenic *Arabidopsis* plants over-expressing the AvrPphB acylation-deficient mutant. We delivered AvrPphB and the mutant proteins at near endogenous levels by exogenously expressing the alleles under control of their native promoters in *Pst*, and found an absolute requirement for host acylation of AvrPphB to promote the avirulence function in *Arabidopsis*. Consistent with our results, delivery of AvrPphB using the *P.s.* pv. *phaseolicola* R6 strain induces HR in bean pods in a myristoylation-dependent manner (50). Interestingly, a recent report demonstrated a NopT acylation-independent HR when the G50A myristoylation mutant was over-expressed in tobacco (51). Our data suggest that delivery of the NopT G50A mutant at endogenous levels may provide additional insight into the function of the G50 residue.

Type III effectors are likely secreted at low concentrations relative to host signaling molecules, and therefore require a potent, but specific biochemical activity that may be enhanced by increasing the effectors' local concentrations via sub-cellular localization. We have identified six additional bacterial effectors, including AvrPphB, which localize to host plasma membranes. Although some AvrPphB family members utilize the host lipidation machinery to direct their association with plasma membranes, we demonstrated that others employ acylation-independent plasma membrane localization mechanisms. There are myriad examples of membrane proteins that lack lipid modifications and localize via protein-protein interactions, phospholipid-protein (electrostatic) interactions, or hydrophobic (integral membrane) associations (33). Although the molecular targets of the AvrPphB-like effectors are completely unknown, localization of these effectors to the plasma membrane likely restricts the host substrate pool to co-localizing plasma membrane proteins. An overwhelming amount of evidence has implicated the plasma membrane as a crucial site for initiation of both basal and *R* gene-mediated defenses (52). Mediators of basal defense pathways are often plasma membrane-localized pattern-recognition receptors (PRRs) and include FLS2, EFR, and BAK1, which are all targeted by the plasma membrane localized effector AvrPto to promote virulence (53,54). Therefore, plasma membrane-associated R proteins and PRRs, as well as their associated signaling molecules, all serve as possible virulence targets for the AvrPphB family members. The mechanisms that drive type III effector function are only beginning to be unraveled; however, our findings provide critical insight into the diverse sub-cellular localization mechanisms employed by AvrPphB family members and illustrate the convoluted subversion strategies utilized by the bacterial pathogen.

Figure 2.1 Multiple amino acid sequence alignment of the AvrPphB family**.**

Members of the AvrPphB family were identified by PSI-BLAST using the AvrPphB sequence as the query. The catalytic residues (indicated by asterisk) are shown in red. Invariant and highly conserved residues are colored in blue and dark gray, respectively. Residues that share similar chemical properties are shown in light gray. The embedded myristoylation (in green) and palmitoylation (in yellow) sites are also shown. The known AvrPphB secondary structure is shown below the alignment (31). Proteins examined in this study are in bold. Additional accession numbers are as follows: AvrPphB, Q52430; HopAW1, AAX12112; ORF4, AAD47206; AvrPpic2, CAC16701; HopC1, AAO54131; HopN1, AAO54892; RipT, NP_521333; NopT, AAB91961; Blr2058, NP_768698; Blr2140, NP_768780.


P.s. pv. phaseolicola 1302A (AvrPphB)
P.s. pv. phaseolicola 1448A (HopAW1)
P.s. pv. phaseolicola 1449B (ORF4)
P.s. pv. phaseolicola 1448A (YP_272236.1)
P.s. pv. pisi (AvrPpic2)
P.s. pv. tomato DC3000 (HopC1)
P.s. pv. tomato DC3000 (HopN1)
Ralstonia solanacearum GMI1000 (RipT)
Erwinia amylovora Ea1189 (AAX97372.1)
Acidovorax avenae sp. citrulli AAC00-1 (YP_968473.1)
Rhizobium sp. NGR234 (NopT)
Bradyrhizobium japonicum USDA 110 (Blr2058)
Bradyrhizobium japonicum USDA 110 (Blr2140)
*
β1
α1
α2
α3
α4
β2
α5
β3
*
*
β4
β5
β6
α6
β7

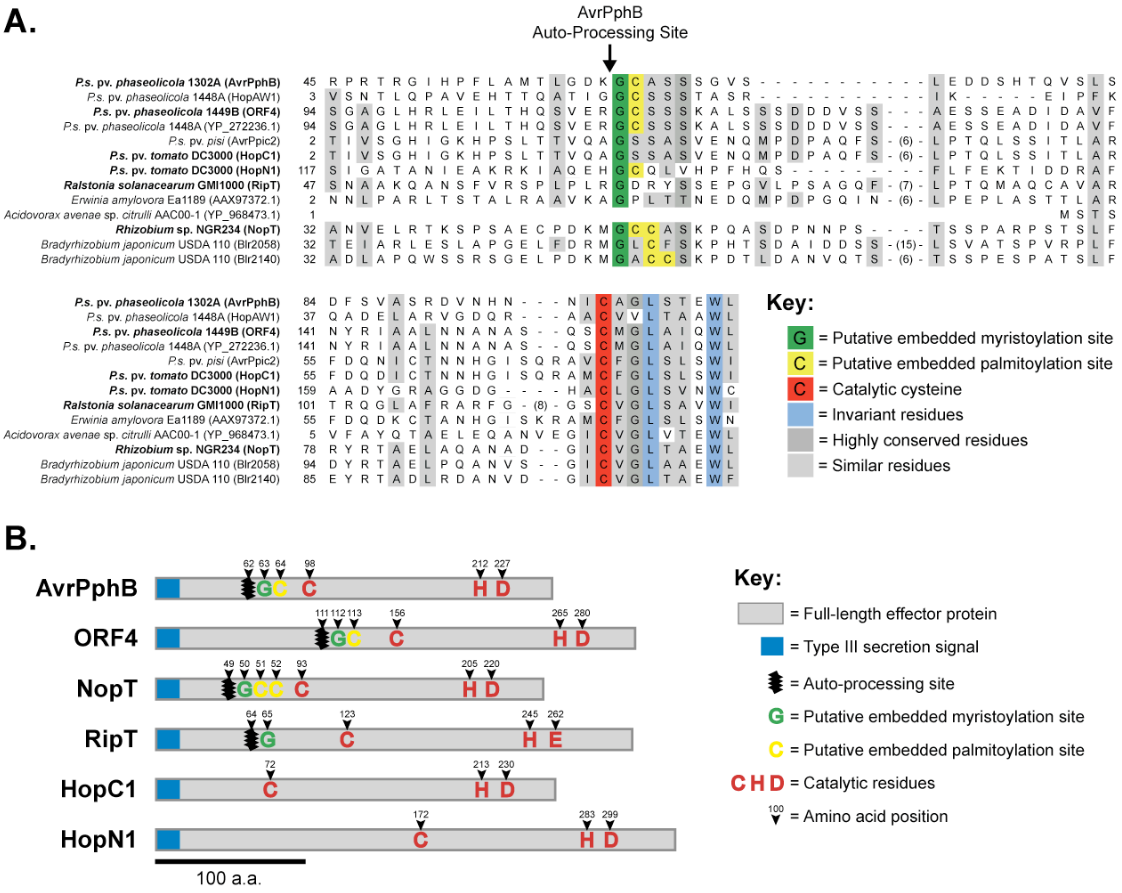

A.
AvrPphB
Auto-Processing Site
P.s. pv. phaseolicola 1302A (AvrPphB)
P.s. pv. phaseolicola 1448A (HopAW1)
P.s. pv. phaseolicola 1449B (ORF4)
P.s. pv. phaseolicola 1448A (YP_272236.1)
P.s. pv. pisi (AvrPpic2)
P.s. pv. tomato DC3000 (HopC1)
P.s. pv. tomato DC3000 (HopN1)
Ralstonia solanacearum GMI1000 (RipT)
Erwinia amylovora Ea1189 (AAX97372.1)
Acidovorax avenae sp. citrulli AAC00-1 (YP_968473.1)
Rhizobium sp. NGR234 (NopT)
Bradyrhizobium japonicum USDA 110 (Blr2058)
Bradyrhizobium japonicum USDA 110 (Blr2140)
Key:
G = Putative embedded myristoylation site
C = Putative embedded palmitoylation site
C = Catalytic cysteine
= Invariant residues
= Highly conserved residues
= Similar residues
B.
AvrPphB
ORF4
NopT
RipT
HopC1
HopN1
100 a.a.
Key:
= Full-length effector protein
= Type III secretion signal
= Auto-processing site
G = Putative embedded myristoylation site
C = Putative embedded palmitoylation site
C H D = Catalytic residues
100 = Amino acid position

Figure 2.2 Multiple amino acid sequence alignment of the AvrPphB family reveals conserved residues within the *N*-terminus.

(A) Members of the AvrPphB family were identified by PSI-BLAST using the AvrPphB sequence as the query and the amino termini were aligned. Residues that share homology, as well as the catalytic cysteine and the embedded acylation sites, are colored according to the key. The known auto-processing site in AvrPphB is also indicated (Shao *et al.* (12), Puri *et al.* (15)). Proteins examined in this study are in bold. Additional accession numbers are as follows: AvrPphB, Q52430; HopAW1, AAX12112; ORF4, AAD47206; AvrPpic2, CAC16701; HopC1, AAO54131; HopN1, AAO54892; RipT, NP_521333; NopT, AAB91961; Blr2058, NP_768698; Blr2140, NP_768780.
(B) A schematic of the full-length effector proteins examined in this study. The catalytic residues, auto-processing sites, and acylation sites are displayed according to the key.

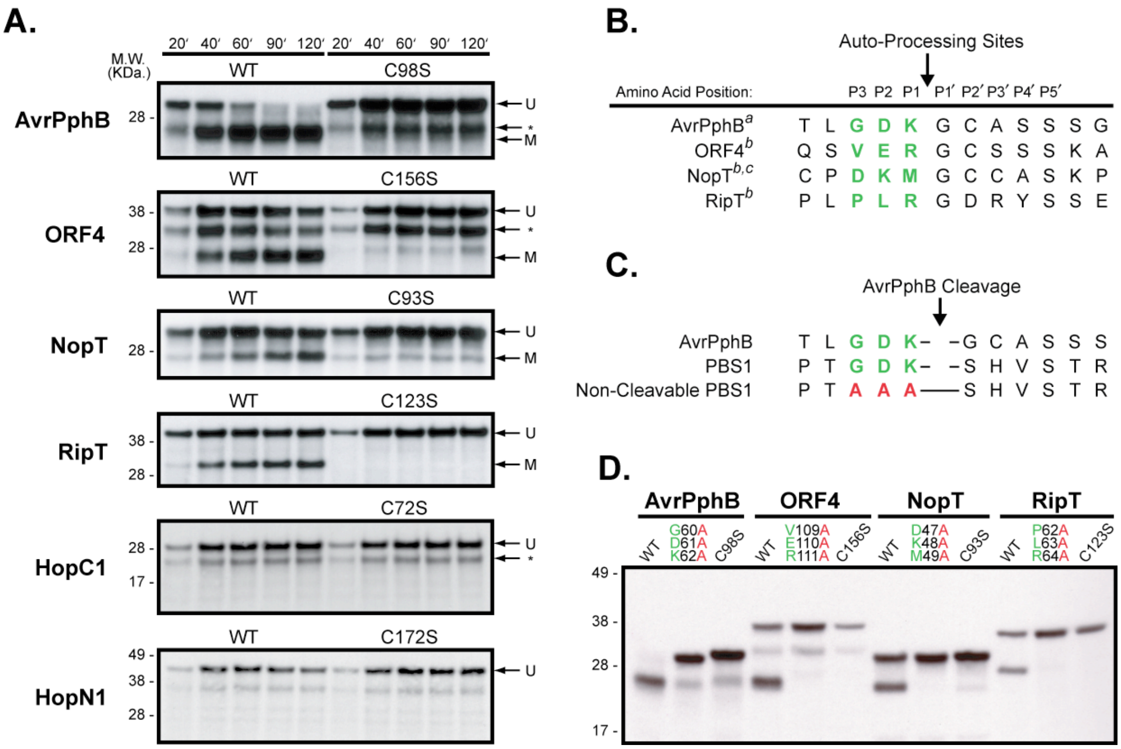


Figure 2.3 Additional AvrPphB family members undergo auto-proteolytic cleavage at specific residues.

(A) The indicated proteins were *in vitro* transcribed and translated in the presence of [^{35}S]methionine and aliquots were removed at the indicated time points. The samples were subjected to SDS-PAGE and proteins were visualized by autoradiography (Unprocessed proteases, U; mature proteases, M). Secondary start methionines produce additional protein species during translation (asterisks) that generate the following proteins: AvrPphB, M=22 KDa. and *(from Met57)=23 KDa.; ORF4, M=23 KDa. and *(from Met51)=30 KDa.; HopC1, U=29 KDa. and *(from Met29)=26 KDa. The experiment was performed three times with similar results.
(B) The auto-processing sites of the mature recombinant proteins were determined by Edman degradation ([a] data from Puri *et al.* (15); [b] data from this study; [c] data from Dai *et al.* (51)). The three residues that precede the cleavage site are shown in green.
(C) Sequence alignment of the known auto-proteolytic processing site in AvrPphB and the cleavage site in PBS1 with the three conserved amino acids that precede the cleavage sites shown in green. Mutation of these residues to alanine (red) in PBS1 inhibits cleavage by AvrPphB (Shao *et al.* (11)).
(D) The indicated P1/P2/P3 triple mutants were generated and analyzed as described in (A). Residues that allow auto-proteolytic processing are shown in green and mutant residues that prevent cleavage are colored in red. The catalytically inactive mutants (C/S) are deficient in auto-processing activity. Additional protein species generated from secondary start methionines (AvrPphB and ORF4) are as described in (A). Each experiment was repeated twice with similar results.

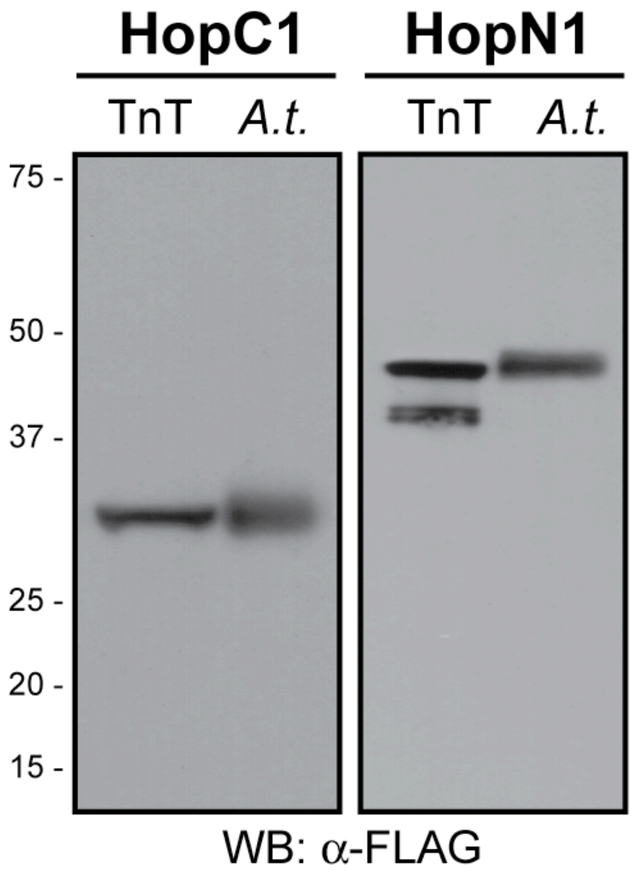


Figure 2.4 HopC1 and HopN1 do not auto-proteolytically process *in planta*.

T_2 transgenic plants carrying DEX::*hopC1*-2xFlag or DEX::*hopN1*-2xFlag were sprayed with 20 μM DEX and tissue was harvested 8 hr later. Plant protein extracts (*A.t.*) and *in vitro* transcribed/translated Flag-tagged proteins (TnT) were subjected to SDS-PAGE, blotted, and probed with an anti-Flag antibody. Secondary start methionines produce additional protein species during translation of HopN1 in the TnT reaction.

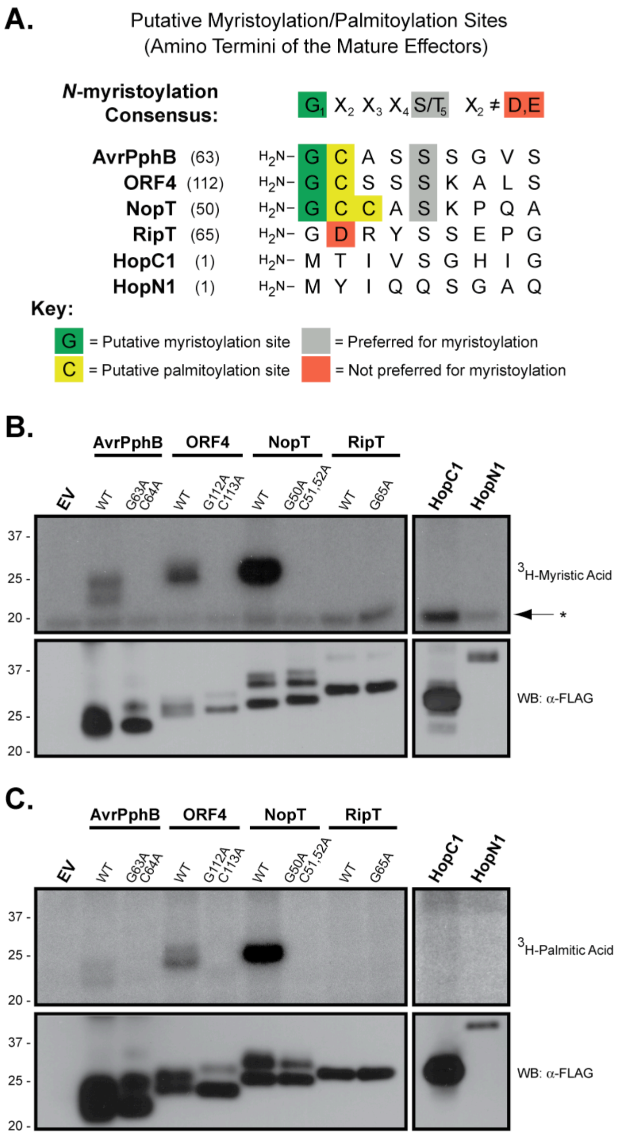


Figure 2.5 Auto-proteolytic processing of AvrPphB family members results in *N*-myristoylation and *S*-palmitoylation of the new amino terminus.

(A) The *N*-terminal sequences of the auto-processed, mature proteins were examined for eukaryotic acylation consensus sites. Important residues are colored according to the key. The myristoylation consensus sequence is based on previous experiments (Utsumi *et al.* (32)).
(B) Full-length wild-type and mutant effectors were expressed in *S. cerevisiae* in the presence of 30 μCi/ml [^{3}H]myristic acid. After labeling for 4 hr, the Flag-tagged effectors were immunoprecipitated, subjected to SDS-PAGE, and analyzed by autoradiography (top panel) or western blotting with an anti-Flag antibody (bottom panel). An asterisk indicates non-specific bands.
(C) *S*-palmitoylation assays of full-length wild-type and mutant effectors were performed and analyzed as described in (B) using 50 μCi/ml [^{3}H]palmitic acid. Yeast radiolabeling experiments were performed twice with similar result

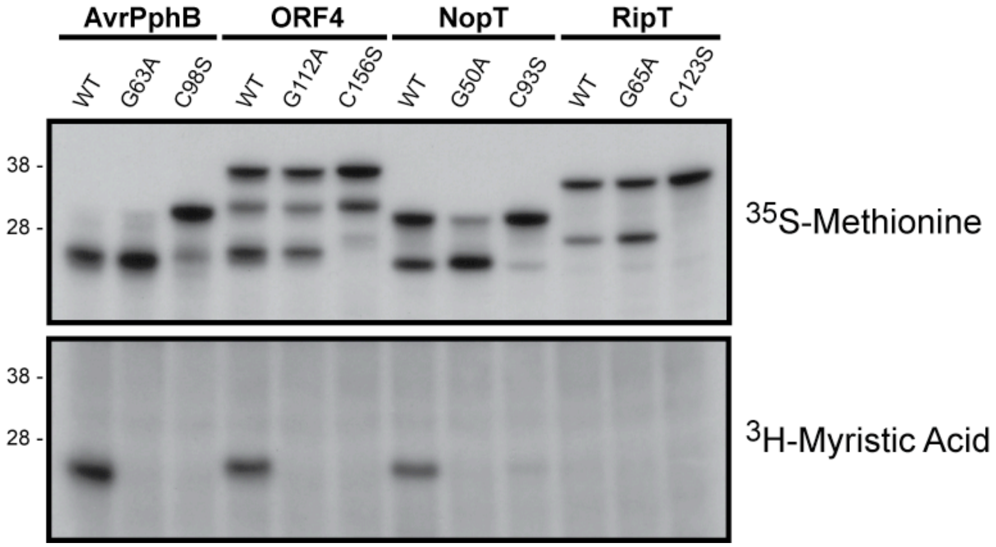


Figure 2.6 AvrPphB family members are *N*-myristoylated in wheat germ extract at embedded consensus sites.

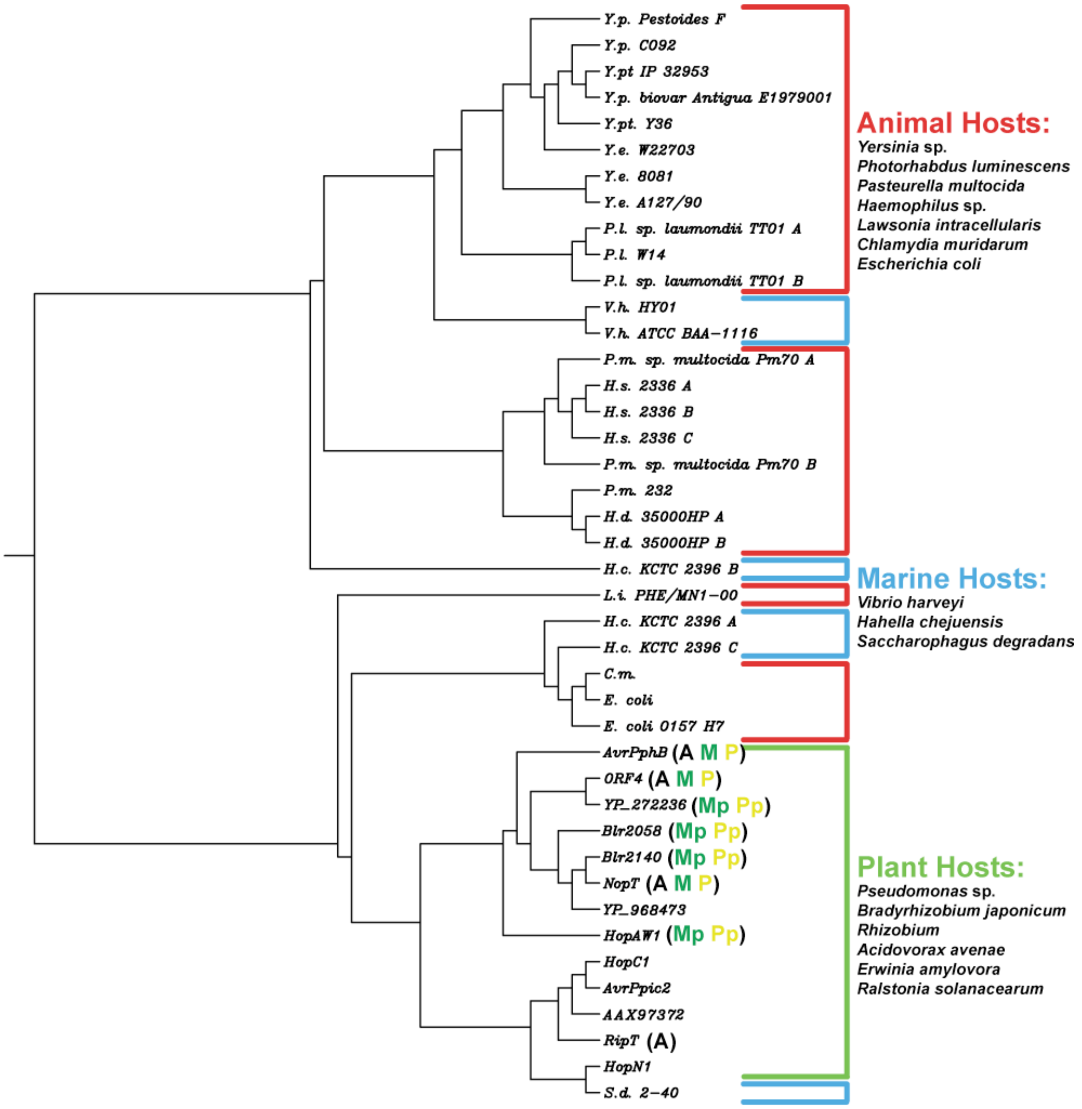
The indicated effectors (wild-type, myristoylation mutants, and catalytically inactive mutants) were produced by *in vitro* transcription and translation in the presence of either [^{35}S]methionine (top) to demonstrate protein synthesis or [^{3}H]myristic acid (bottom) to demonstrate myristoylation. After a 2 hr incubation, samples were quenched in 2× SDS sample buffer, subjected to SDS-PAGE, and analyzed by autoradiography. The experiment was repeated twice with similar results.

Figure 2.7 Phylogenetic analysis of the YopT family reveals evolutionary clustering of auto-proteolytic processing activity.

(A) Members of the YopT family were identified by PSI-BLAST using the YopT and AvrPphB sequences as the queries. The YopT phylogenetic tree was generated with the ClustalW server (http://align.genome.jp). Virulence factors from the indicated bacterial pathogens are classified according to their corresponding host specificity: Animal (red), Marine (blue), or Plant (green). The AvrPphB homologues are labeled according to the key as auto-processed, myristoylated, or palmitoylated.
(B) Embedded acylation sites are found in additional AvrPphB family members. Proteins were scanned visually for myristoylation consensus sites and then subjected to the Myristoylator prediction program (http://us.expasy.org/tools/myristoylator/; Bologna *et al.* (39)). Putative myristoylation sites are colored in green, palmitoylation sites in yellow, and residues preferred for myristoylation in gray.
.

A.

YopT Phylogenetic Tree



A = Auto-proteolytic processed (experimental)
M = Myristoylated by Eukaryotic Machinery (experimental)
P = Palmitoylated by Eukaryotic Machinery (experimental)
Mp = Putative embedded Myristoylation sites (Untested)
Pp = Putative embedded Palmitoylation sites (Untested)

B.

Putative Embeded Myristoylation/Palmitoylation Sites
of Untested Effector Proteins

HopAW1	(20)	T	I	G	G	C	S	S	S	T
HopAW1	(21)	I	G	G	C	S	S	S	T	A
YP_272236.1	(110)	E	R	G	C	S	S	S	K	A
Blr2058	(48)	R	M	G	L	C	F	S	K	P
Blr2058	(68)	T	S	G	L	S	T	S	S	L
Blr2140	(4)	R	I	G	G	S	S	T	R	T
Blr2140	(48)	K	M	G	A	C	C	S	K	P

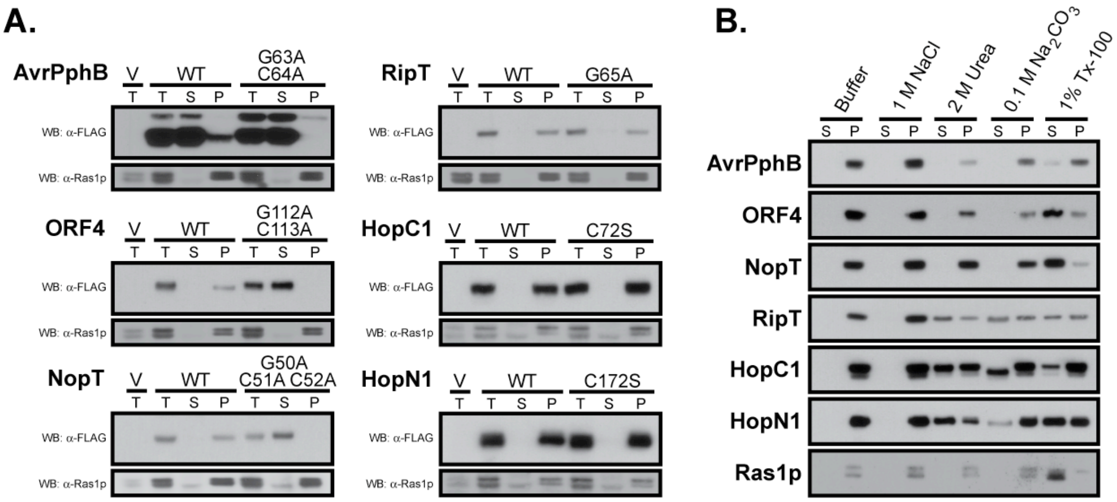


Figure 2.8 Acylated and non-acylated AvrPphB family members are differentially associated with *S. cerevisiae* membranes.

(A) Strains carrying the indicated Flag-tagged effectors or empty vector (V) were induced with galactose for 8 hr and homogenized. Total extracts (T) were fractionated into soluble (S) fractions and insoluble membrane pellets (P) by ultracentrifugation at 100,000×g. Equal volumes of each fraction were subjected to SDS-PAGE, blotted, and probed with anti-Flag or anti-v-H-Ras (plasma membrane marker) antibodies. (B) Membranes were isolated as in (A) and resuspended in either control lysis buffer, high salt buffer (1 M NaCl, 10 mM Tris-HCl, pH 7.4), denaturing buffer (2 M urea, 10 mM Tris-HCl, pH 7.4), high pH buffer (0.1 M Na_2CO_3, pH 11.5), or buffer containing detergent (1% Triton X-100, 10 mM Tris-HCl, pH 7.4). Treated samples were re-ultracentrifuged and equal volumes of the soluble (S) and pellet (P) fractions were subjected to SDS-PAGE and western blot analysis as in (A). Each experiment was performed twice with similar results.

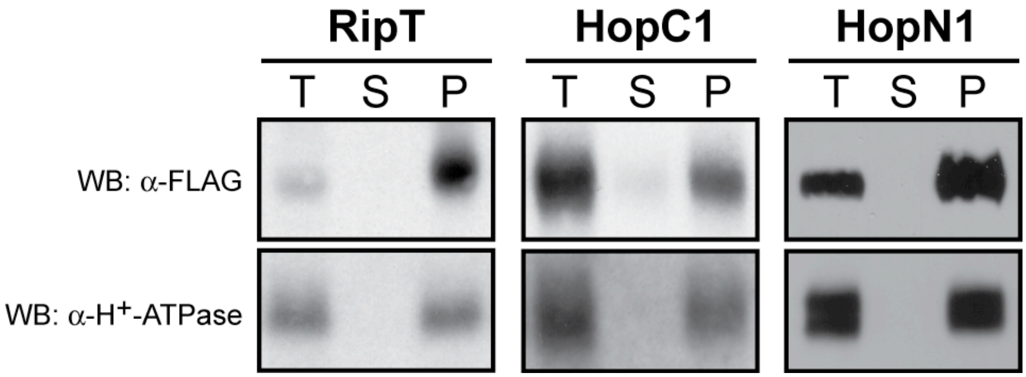


Figure 2.9 The non-acylated AvrPphB family members associate with *Arabidopsis* cell membranes.

The RipT, HopC1, and HopN1 effectors were expressed in *Arabidopsis* and leaves were harvested 8 hr after DEX treatment. Total (T) protein extracts were fractionated by ultracentrifugation (100,000×g for 1 hr) into soluble (S) and insoluble membrane pellets (P). Equal volumes of each fraction were subjected to SDS-PAGE, blotted, and probed with anti-Flag or anti-H+-ATPase (plasma membrane marker) antibodies. All apparent molecular weights are correct (RipT at 28 kDa, HopC1 at 29 kDa, HopN1 at 39 kDa, H+-ATPase at 25 kDa).

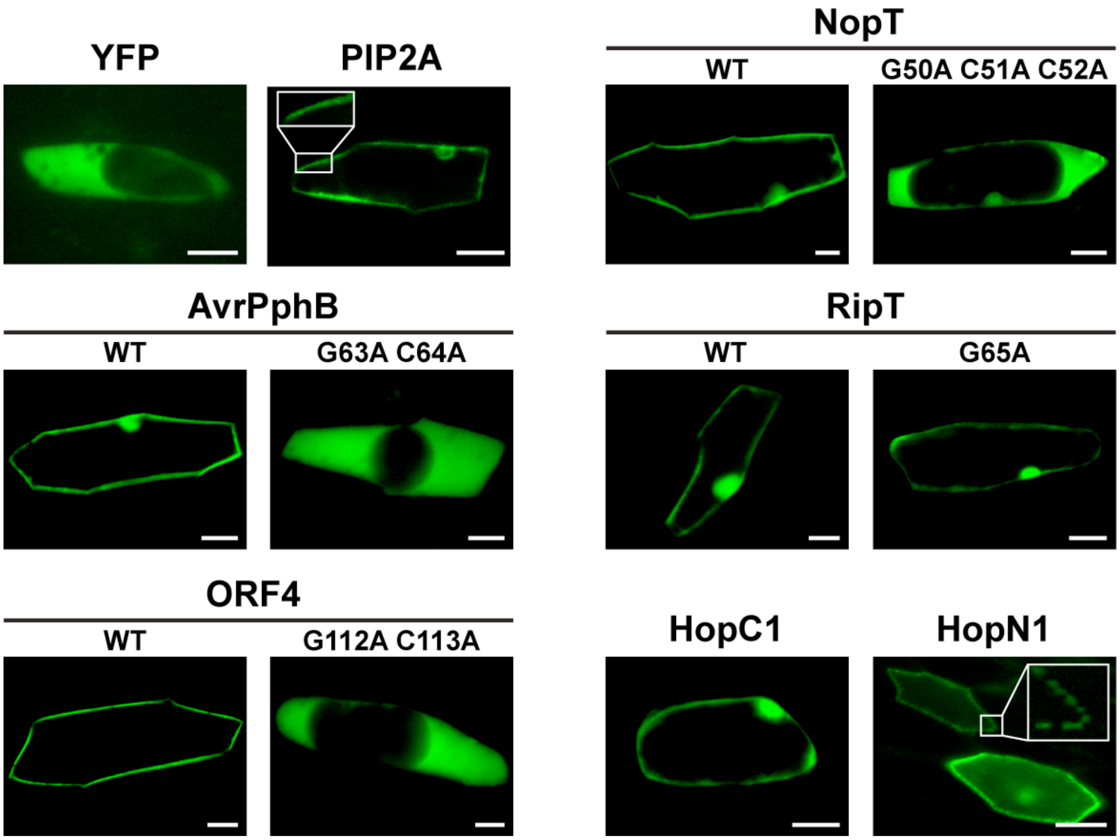
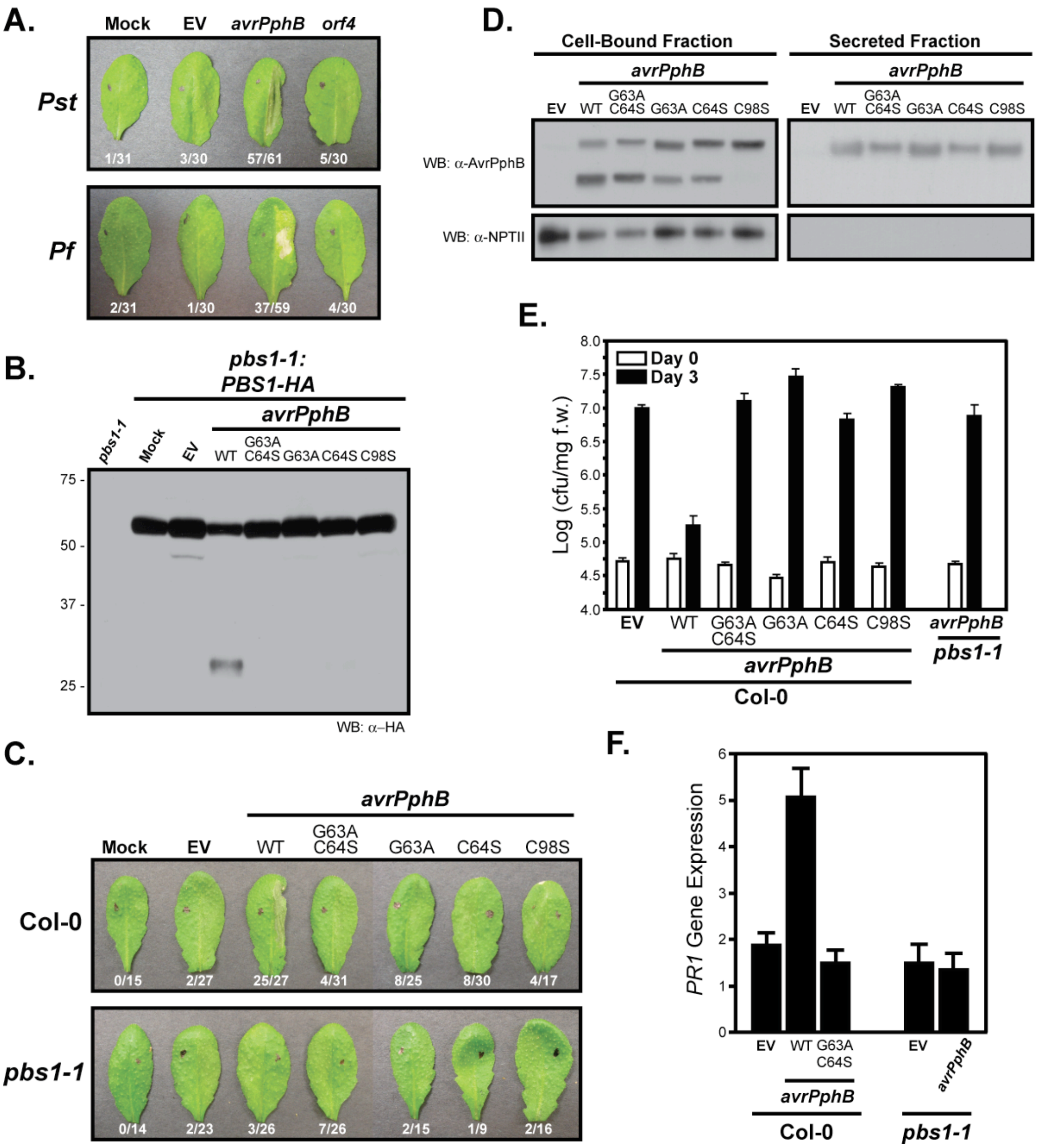


Figure 2.10 AvrPphB-like family members localize to the plasma membranes of Chinese cabbage cells.

C-terminally tagged YFP effector proteins were transiently expressed in Chinese cabbage epidermal cells using particle bombardment. Representative fluorescent images of cells expressing wild-type effectors or acylation-deficient mutants are indicated. Control bombardments were performed using the cytosolic YFP or plasma membrane localized PIP2A (Plasma membrane Intrinsic Protein 2A)-CFP proteins. Both YFP and CFP fluorescence is colored in green. Bar, 50 μm.

Figure 2.11 Host acylation of AvrPphB is required for avirulence activity in *Arabidopsis* plants carrying *PBS1.*

(A) Adult Col-0 leaves were syringe infiltrated (opposite to the marked leaf half) with ~3.75×10^7 cfu/ml *P. syringae* pv. *tomato* DC3000 (*Pst*) or *P. fluorescens* pLN1965 (*Pf*) strains expressing AvrPphB or ORF4. Also, plants were inoculated with 10 mM $MgCl_2$ (Mock) or strains carrying the pVSP61 empty vector (EV). Ratios below each leaf indicate the number of HR positive leaves/total number of leaves inoculated.
(B) Transgenic *pbs1-1:PBS1-HA* plants were inoculated as in (A) with the indicated *Pst* strains. Leaf tissue was harvested 14 hpi, homogenized, and 10 μg of total protein was subjected to SDS-PAGE. Blots were analyzed by anti-HA western blotting. Three individual T_2 plants were assayed for each infection condition and produced identical results.
(C) Col-0 or *pbs1-1* plants were inoculated as described in (A) with *Pst* strains carrying the indicated *avrPphB* alleles (G63, myristoylation site; C64, palmitoylation site; C98, catalytic cysteine). Data was collected 20 hpi and is representative of two independent experiments.
(D) *Pst* strains carrying the indicated alleles were grown in Hrp-inducing minimal media. Cultures were partitioned into cell-bound and secreted fractions by centrifugation. Protein samples were subjected to SDS-PAGE, blotted, and probed with antibodies against AvrPphB or NPTII (control for nonspecific lysis).
(E) *Arabidopsis* seedlings were inoculated by dipping with *Pst* strains (~2.5×10^7 cfu/ml) carrying the indicated effector alleles. At day 0 (white bars) or day 3 (black bars) the bacteria were extracted and quantified. Data are represented as the mean +/- SEM of four technical replicates. The experiment was repeated twice with similar results.
(F) *Arabidopsis* seedlings were inoculated as described in (E) with the indicated *Pst* strains. Tissue was harvested 24 hpi and RNA was subjected to RT-qPCR analysis using *PR1* and *TUBULIN3* specific primers. *PR1* mRNA levels (relative to *TUB3*) were calibrated to mock treated samples ($2^{-\Delta\Delta Ct}$). Data are represented as the mean +/- SEM of at least 5 biological replicates from two independent experiments.


A.
Mock
EV
avrPphB
orf4
Pst
1/31
3/30
57/61
5/30
Pf
2/31
1/30
37/59
4/30
B.
pbs1-1:
PBS1-HA
avrPphB
pbs1-1
Mock
EV
WT
G63A C64S
G63A
C64S
C98S
75 -
50 -
37 -
25 -
WB: α–HA
C.
avrPphB
Mock
EV
WT
G63A C64S
G63A
C64S
C98S
Col-0
0/15
2/27
25/27
4/31
8/25
8/30
4/17
pbs1-1
0/14
2/23
3/26
7/26
2/15
1/9
2/16
D.
Cell-Bound Fraction
Secreted Fraction
avrPphB
avrPphB
EV
WT
G63A C64S
G63A
C64S
C98S
EV
WT
G63A C64S
G63A
C64S
C98S
WB: α-AvrPphB
WB: α-NPTII
E.
Day 0
Day 3
Log (cfu/mg f.w.)
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
EV
WT
G63A C64S
G63A
C64S
C98S
avrPphB
avrPphB
pbs1-1
Col-0
F.
PR1 Gene Expression
6
5
4
3
2
1
0
EV
WT
G63A C64S
avrPphB
EV
avrPphB
Col-0
pbs1-1

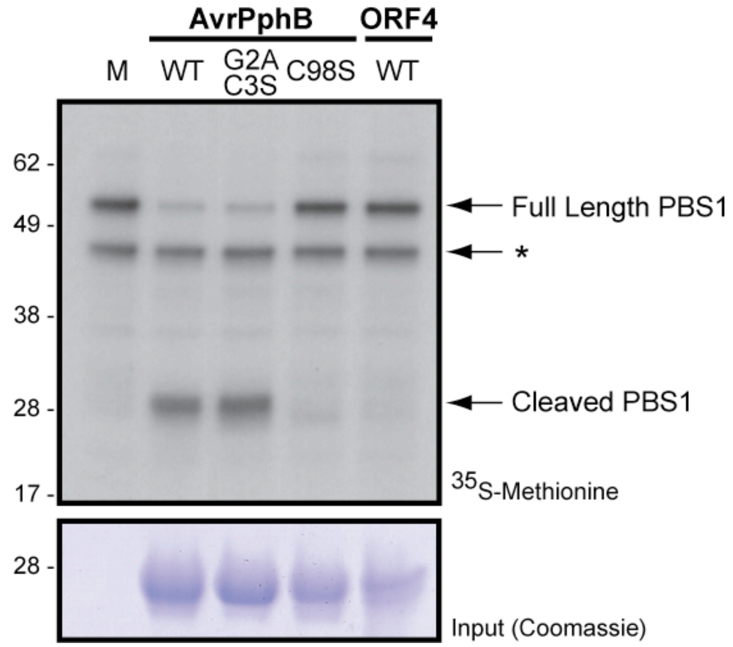


Figure 2.12 The AvrPphB acylation-deficient mutant, but not ORF4, cleaves PBS1 *in vitro*.

PBS1 radiolabeled protein was generated by *in vitro* transcription/translation in rabbit reticulocyte extract in the presence of [^{35}S]methionine. Cleavage reactions were performed by incubating recombinant AvrPphB(Δ62)-His$_6$ (wt or mutant proteins) or ORF4(Δ111)-His$_6$ at 0.2 μg/μL with radiolabeled PBS1 for 1 hr. The mock (M) reaction contained no effector protein. Samples were subjected to SDS-PAGE and analyzed by autoradiography (top) and by Coomassie blue staining of the recombinant proteins (bottom). The asterisk indicates a non-specific band generated during *in vitro* transcription and translation.

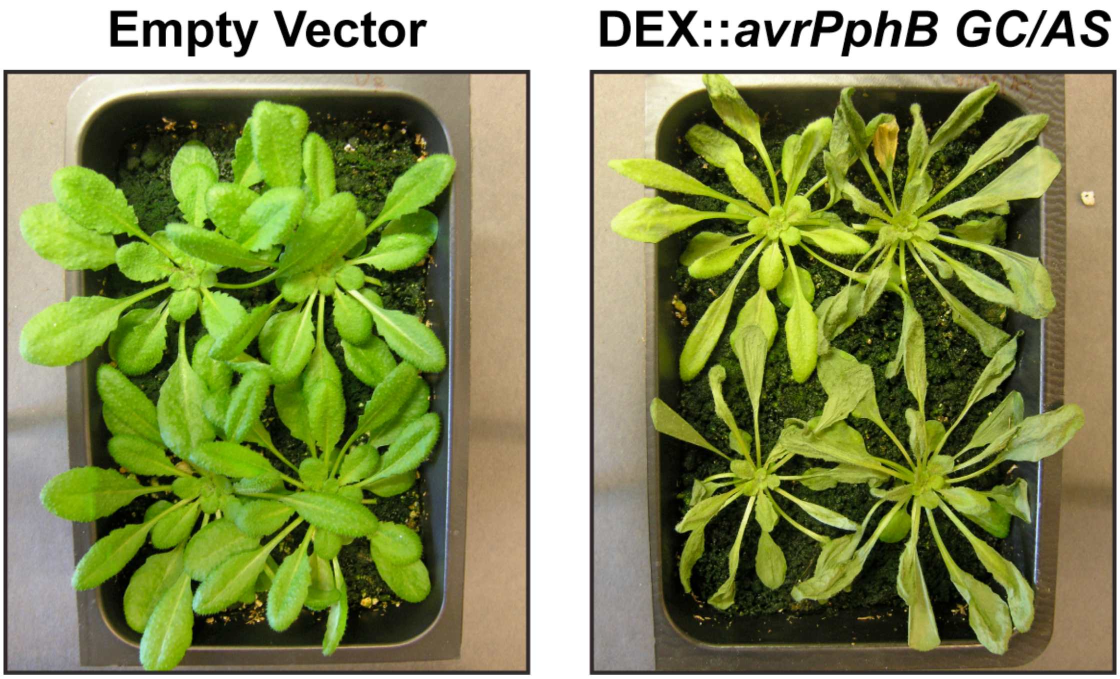


Figure 2.13 Over-expression of the acylation-deficient AvrPphB mutant effector in resistant plants induces HR.

T_2 transgenic plants carrying DEX::*avrPphB(Δ62)GC/AS*-2xFlag or the pTA7002 empty vector were sprayed with 20 μM DEX and plants were photographed 16 hr later. Two independent transgenic plant lines were assayed and produced similar results.

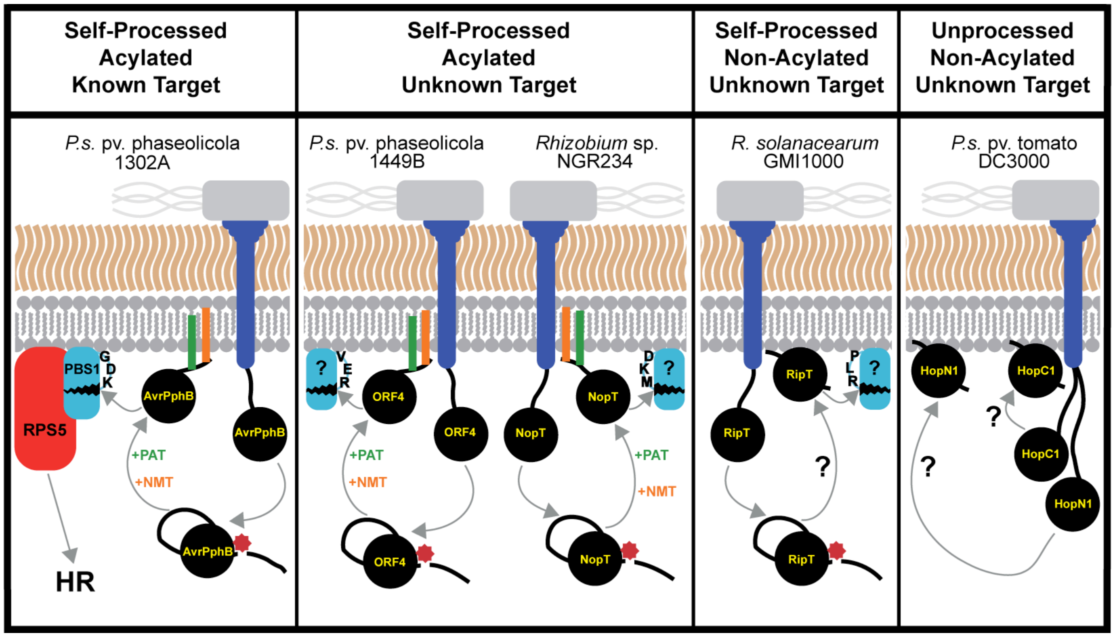

Self-Processed Acylated Known Target
Self-Processed Acylated Unknown Target
Self-Processed Non-Acylated Unknown Target
Unprocessed Non-Acylated Unknown Target
P.s. pv. phaseolicola 1302A
P.s. pv. phaseolicola 1449B
Rhizobium sp. NGR234
R. solanacearum GMI1000
P.s. pv. tomato DC3000
RPS5
PBS1
G D K
AvrPphB
+PAT
+NMT
HR
?
V E R
ORF4
NopT
D K M
RipT
P L R
HopN1
HopC1

Figure 2.14 A model for the sub-cellular localization strategies of the AvrPphB-like effector proteins in the plant cell.

The indicated strains are shown in gray, the TTSS in purple, and the effectors in black. Effectors are classified according to their ability to self-proteolytically process (red star in model), to be acylated by the host (*N*-myristoylation, orange; *S*-palmitoylation, green), and their biological function. AvrPphB, ORF4, and NopT are lipidated by the host machinery (NMT, *N*-myristoyl transferase; PAT, palmitoyl acyl transferase), while RipT, HopC1, and HopN1 are directed to the PM by an unknown mechanism. Host acylation of AvrPphB is essential for cleavage of PBS1 (blue) and initiation of RPS5 (red) defenses. Unknown targets of ORF4, NopT, and RipT are also included (blue) and contain putative target sequences.

REFERENCES

1. Mackey, D., and McFall, A. J. (2006) *Mol Microbiol* **61**(6), 1365-1371

2. Jones, J. D., and Dangl, J. L. (2006) *Nature* **444**(7117), 323-329

3. Chang, J. H., Urbach, J. M., Law, T. F., Arnold, L. W., Hu, A., Gombar, S., Grant, S. R., Ausubel, F. M., and Dangl, J. L. (2005) *Proc Natl Acad Sci U S A* **102**(7), 2549-2554

4. Roine, E., Wei, W., Yuan, J., Nurmiaho-Lassila, E. L., Kalkkinen, N., Romantschuk, M., and He, S. Y. (1997) *Proc Natl Acad Sci U S A* **94**(7), 3459-3464

5. Scheel, D. (1998) *Curr Opin Plant Biol* **1**(4), 305-310

6. Nimchuk, Z., Eulgem, T., Holt, B. F., 3rd, and Dangl, J. L. (2003) *Annu Rev Genet* **37**, 579-609

7. Deslandes, L., Olivier, J., Peeters, N., Feng, D. X., Khounlotham, M., Boucher, C., Somssich, I., Genin, S., and Marco, Y. (2003) *Proc Natl Acad Sci U S A* **100**(13), 8024-8029

8. Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Teh, T., Wang, C. I., Ayliffe, M. A., Kobe, B., and Ellis, J. G. (2006) *Proc Natl Acad Sci U S A* **103**(23), 8888-8893

9. Jia, Y., McAdams, S. A., Bryan, G. T., Hershey, H. P., and Valent, B. (2000) *Embo J* **19**(15), 4004-4014

10. Nimchuk, Z., Marois, E., Kjemtrup, S., Leister, R. T., Katagiri, F., and Dangl, J. L. (2000) *Cell* **101**(4), 353-363

11. Shao, F., Golstein, C., Ade, J., Stoutemyer, M., Dixon, J. E., and Innes, R. W. (2003) *Science* **301**(5637), 1230-1233

12. Shao, F., Merritt, P. M., Bao, Z., Innes, R. W., and Dixon, J. E. (2002) *Cell* **109**(5), 575-588

13. Ade, J., DeYoung, B. J., Golstein, C., and Innes, R. W. (2007) *Proc Natl Acad Sci U S A* **104**(7), 2531-2536

14. Holt, B. F., 3rd, Belkhadir, Y., and Dangl, J. L. (2005) *Science* **309**(5736), 929-932

15. Puri, N., Jenner, C., Bennett, M., Stewart, R., Mansfield, J., Lyons, N., and Taylor, J. (1997) *Mol Plant Microbe Interact* **10**(2), 247-256

16. Perret, X., Broughton, W. J., and Brenner, S. (1991) *Proc Natl Acad Sci U S A* **88**(5), 1923-1927

17. Wang, X., Li, X., Meisenhelder, J., Hunter, T., Yoshida, S., Asami, T., and Chory, J. (2005) *Dev Cell* **8**(6), 855-865

18. Bisgrove, S. R., Simonich, M. T., Smith, N. M., Sattler, A., and Innes, R. W. (1994) *Plant Cell* **6**(7), 927-933

19. Jamir, Y., Guo, M., Oh, H. S., Petnicki-Ocwieja, T., Chen, S., Tang, X., Dickman, M. B., Collmer, A., and Alfano, J. R. (2004) *Plant J* **37**(4), 554-565

20. Simonich, M. T., and Innes, R. W. (1995) *Mol Plant Microbe Interact* **8**(4), 637-640

21. Warren, R. F., Merritt, P. M., Holub, E., and Innes, R. W. (1999) *Genetics* **152**(1), 401-412

22. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res* **25**(17), 3389-3402

23. Aoyama, T., and Chua, N. H. (1997) *Plant J* **11**(3), 605-612

24. Clough, S. J., and Bent, A. F. (1998) *Plant J* **16**(6), 735-743

25. Lee, S. W., Mitchell, D. A., Markley, A. L., Hensler, M. E., Gonzalez, D., Wohlrab, A., Dorrestein, P. C., Nizet, V., and Dixon, J. E. (2008) *Proc Natl Acad Sci U S A*

26. Mitchell, D. A., and Marletta, M. A. (2005) *Nat Chem Biol* **1**(3), 154-158

27. Song, J., Hirschman, J., Gunn, K., and Dohlman, H. G. (1996) *J Biol Chem* **271**(34), 20273-20283

28. Tornero, P., and Dangl, J. L. (2001) *Plant J* **28**(4), 475-481

29. Huynh, T. V., Dahlbeck, D., and Staskawicz, B. J. (1989) *Science* **245**(4924), 1374-1377

30. Nobuta, K., Okrent, R. A., Stoutemyer, M., Rodibaugh, N., Kempema, L., Wildermuth, M. C., and Innes, R. W. (2007) *Plant Physiol* **144**(2), 1144-1156

31. Zhu, M., Shao, F., Innes, R. W., Dixon, J. E., and Xu, Z. (2004) *Proc Natl Acad Sci U S A* **101**(1), 302-307

32. Utsumi, T., Sato, M., Nakano, K., Takemura, D., Iwata, H., and Ishisaka, R. (2001) *J Biol Chem* **276**(13), 10505-10513

33. Resh, M. D. (1999) *Biochim Biophys Acta* **1451**(1), 1-16

34. Delaney, T. P., Uknes, S., Vernooij, B., Friedrich, L., Weymann, K., Negrotto, D., Gaffney, T., Gut-Rella, M., Kessmann, H., Ward, E., and Ryals, J. (1994) *Science* **266**(5188), 1247-1250

35. Guttman, D. S., Vinatzer, B. A., Sarkar, S. F., Ranall, M. V., Kettler, G., and Greenberg, J. T. (2002) *Science* **295**(5560), 1722-1726

36. Petnicki-Ocwieja, T., Schneider, D. J., Tam, V. C., Chancey, S. T., Shan, L., Jamir, Y., Schechter, L. M., Janes, M. D., Buell, C. R., Tang, X., Collmer, A., and Alfano, J. R. (2002) *Proc Natl Acad Sci U S A* **99**(11), 7652-7657

37. Coaker, G., Falick, A., and Staskawicz, B. (2005) *Science* **308**(5721), 548-550

38. Lopez-Solanilla, E., Bronstein, P. A., Schneider, A. R., and Collmer, A. (2004) *Mol Microbiol* **54**(2), 353-365

39. Bologna, G., Yvon, C., Duvaud, S., and Veuthey, A. L. (2004) *Proteomics* **4**(6), 1626-1632

40. Podell, S., and Gribskov, M. (2004) *BMC Genomics* **5**(1), 37

41. Shan, L., Thara, V. K., Martin, G. B., Zhou, J. M., and Tang, X. (2000) *Plant Cell* **12**(12), 2323-2338

42. Robert-Seilaniantz, A., Shan, L., Zhou, J. M., and Tang, X. (2006) *Mol Plant Microbe Interact* **19**(2), 130-138

43. Thieme, F., Szczesny, R., Urban, A., Kirchner, O., Hause, G., and Bonas, U. (2007) *Mol Plant Microbe Interact* **20**(10), 1250-1261

44. Lewis, J. D., Abada, W., Ma, W., Guttman, D. S., and Desveaux, D. (2008) *J Bacteriol* **190**(8), 2880-2891

45. Maurer-Stroh, S., and Eisenhaber, F. (2004) *Trends Microbiol* **12**(4), 178-185

46. Glover, C. J., Hartman, K. D., and Felsted, R. L. (1997) *J Biol Chem* **272**(45), 28680-28689

47. Zha, J., Weiler, S., Oh, K. J., Wei, M. C., and Korsmeyer, S. J. (2000) *Science* **290**(5497), 1761-1765

48. Berthiaume, L., and Resh, M. D. (1995) *J Biol Chem* **270**(38), 22399-22405

49. Dunphy, J. T., Greentree, W. K., Manahan, C. L., and Linder, M. E. (1996) *J Biol Chem* **271**(12), 7154-7159

50. Tampakaki, A. P., Bastaki, M., Mansfield, J. W., and Panopoulos, N. J. (2002) *Mol Plant Microbe Interact* **15**(3), 292-300

51. Dai, W. J., Zeng, Y., Xie, Z. P., and Staehelin, C. (2008) *J Bacteriol*

52. Nurnberger, T., Brunner, F., Kemmerling, B., and Piater, L. (2004) *Immunol Rev* **198**, 249-266

53. Shan, L., He, P., Li, J., Heese, A., Peck, S. C., Nurnberger, T., Martin, G. B., and Sheen, J. (2008) *Cell Host Microbe* **4**(1), 17-27

54. Xiang, T., Zong, N., Zou, Y., Wu, Y., Zhang, J., Xing, W., Li, Y., Tang, X., Zhu, L., Chai, J., and Zhou, J. M. (2008) *Curr Biol* **18**(1), 74-80

The text of Chapter 2 is a reprint of the material as it appears in the *Journal of Biological Chemistry,* 2009, Vol. 284, No. 23, Robert H. Dowen, James L. Engel, Feng Shao, Joseph R. Ecker, and Jack E. Dixon. The dissertation author was the primary researcher and the co-authors listed in the publication assisted and/or supervised the research that forms the basis of this chapter.

CHAPTER 3

Epigenetic regulation of the plant defense system against *Pseudomonas syringae*

ABSTRACT

For many higher eukaryotic organisms, DNA methylation is an essential and heritable regulatory component of a cell's transcriptional programming. Recent technologies have enabled high resolution profiling of the methylcytosines across the entire genome of the flowering plant *Arabidopsis thaliana* and have underscored the complex relationship between DNA methylation and transcriptional output. Here, we examine the role DNA methylation plays in protecting *Arabidopsis* against the bacterial phytopathogen *Pseudomonas syringae*. We found that mutants deficient in either maintenance or *de novo* DNA methylation pathways are markedly resistant to virulent bacterial pathogen, suggesting that these plants are genetically capable, but epigenetically repressed, to encode resistance against *Pseudomonas syringae.* Additionally, we have shown that the *de novo* methytransferases *DRM1*, *DRM2*, and *CMT3* are essential for a transgenerational memory of bacterial infection that normally encodes enhanced resistance to pathogen in wild-type plants. Using high-throughput deep sequencing technologies, we profiled the DNA methylome and transcriptome of *Arabidopsis* leaf tissue after infection with *Pseudomonas syringae*. These genomic approaches have revealed that pathogen-induced transient alterations in DNA methylation likely serve as important regulatory components at several loci across the genome, highlighting the complexity of the plant defense network.

INTRODUCTION

Epigenetic regulation of basic biological processes, including cellular responses to stress, is an important component of an organism's ability to sense and react to an environmental stimulus. Furthermore, alterations of DNA methylation patterns, histone modifications, and small RNA populations can result in stable, and heritable, changes in chromatin structure and transcriptional output. In higher eukaryotic organisms, DNA methylation encodes of additional layer of information on top of the genetic code and is known to be essential for proper regulation of several biological processes, including embryogenesis, stem cell patterning, genomic imprinting, X-chromosome inactivation, and tumorigenesis in mammals, as well as suppression of repetitive elements and gene regulation in plants (1-8). The DNA methylome of the flowering plant *Arabidopsis thaliana* has been resolved at single base resolution (4,9), and integrated with transcriptome and smRNAome profiling (4). These studies have uncovered a complex and dynamic relationship between smRNA directed DNA methylation, transposon silencing, and transcriptional regulation across the entire genome.

In *Arabidopsis*, DNA methylation is established by *DRM1/2* in all sequence contexts (CpG, CpHpG, CpHpH; where H = A, C, T), a process orthologous to mammalian *de novo* cytosine methylation performed by *DNMT3a/b*. Interestingly, generation of small RNAs (21-24 nt) through a *DICER-LIKE3*/*AGONAUTE4* pathway guides *de novo* methytransferases to specific regions throughout the plant genome, resulting in large deposits of non-CG methylation (10,11). This RNA-directed DNA methylation (RdDM) pathway clearly functions in plants to silence transposable elements and regulate a wide variety of endogenous transcripts (4), and it is likely that a functionally orthologous mechanism exists in mammals (12). Interestingly, non-

CG methylation in the CHG context, which is commonly found on both DNA strands in a symmetric pattern, is maintained by the plant-specific cytosine methyltransferase CMT3, and is directed by both smRNA molecules as well as proteinaceous chromatin binding factors (13,14). Maintenance of symmetric CpG methylation occurs during DNA strand replication by the *Arabidopsis* protein MET1, a homolog of the mammalian DNMT1 protein, through a process whereby the parental strand likely serves as a “methylation template” for subsequent methylation of the newly replicated DNA molecule (4,9,15-17). Although maintenance of CG methylation is clearly propagated by MET1, it remains unknown how asymmetrical non-CG methylation is maintained at specific regions in the genome, but at many genomic loci, this process likely involves rapid recognition of transcripts originating from repetitive elements, followed by processing of these transcripts into smRNAs, and initiation of RdDM.

The *Arabidopsis* genes *DME*, *DML2*, *DML3*, and *ROS1* encode DNA glycosylases that actively antagonize the action of the DNA methyltransferases at all sequence contexts (18-21). Although it remains unclear how DNA demethylases are regulated and specifically directed to genomic targets, it is likely that methylcytosines are removed in plants by a base excision mechanism that is essential for maintaining hundreds of loci in a demethylated state (20,21). Although somewhat controversial, active demethylation pathways also appear to be present in mammalian systems and have been implicated in a wide range of processes including transient regulation of specific loci, as well as post-fertilization demethylation of entire parental genomes (22-25). Interestingly, multiple groups have observed a rapid cycling of stand-specific promoter methylation/demethylation of the human oestrogen-responsive *pS2* gene in cells after addition of oestradiol (24,26). In addition, Bruniquel and colleagues demonstrated that a promoter-enhancer region of the interleukin-2 gene is actively demethylated in differentiated T cells upon stimulation (23). Together, these data indicate that transient methylation or demethylation of mammalian genomic elements is a likely mechanism for actively regulating gene transcription, and it is possible that analogous pathways exist in plant species.

Epigenetic regulation of *Arabidopsis* stress responses have been of particular interest due to the wide range of genetic and genomic tools available for probing these complex pathways, as well as pre-existing whole genome profiles of DNA methylation, small RNAs, and mRNAs (4). Despite recent technological advances, stress-induced transient alterations in DNA methylation have been difficult to resolve and have only been observed at a handful of different loci (27-32). To this end, we sought to examine the role of DNA methylation in directly regulating resistance against the bacterial phytopathogen *Pseudomonas syringae*. Remarkably, we found that *Arabidopsis* mutants deficient in either CG maintenance methylation or non-CG *de novo* methylation are substantially more resistant to pathogen. Wild-type plants, therefore, have the genetic potential to suppress pathogen growth, but are epigenetically restricted at one or more genomic loci, suggesting that plants may be primed to shift to a more resistant state after persistent pathogen stress. Indeed, we found that wild-type plants became more resistant to the virulent *P. syringae* pv. *tomato* DC3000 strain over multiple generations of persistent pathogen stress, a process that requires the *de novo* methyltransferases *DRM1/2* and *CMT3*. To identify any transient alterations in DNA methylation upon infection, we employed genome wide methylC-Seq (4) on untreated and pathogen infected *Arabidopsis* plants. We identified pathogen-induced hypo and hypermethylated regions within the *Arabidopsis* genome, indicating that environmental stimuli are capable of rapidly instilling transient DNA methylation changes, a mechanism that has been previously observed in mammalian systems. Together, these studies illustrate the epigenetic flexibility of the plant’s early response to pathogen stress, as well as the potential to coordinate a methylation-dependent transgeneration memory of infection.

EXPERIMENTAL PROCEDURES

Plant Lines and Bacterial Strains

Arabidopsis thaliana plants were grown in a Promix-HP:vermiculite (2:1) soil mix under short day conditions (9 hr photoperiod at 22°C). The mutant plant lines *met1* (*met1-3*) and *ddc* (*drm1-2 drm2-2 cmt3-11*, triple mutant) have been previously described (4,17,33). The pathogen strains used in this study were *Pseudomonas syringae* pv. *tomato* DC3000 (*Pst*, gift from Brian Staskawicz), *Pst(avrPphB)* (34), and *Pst(hrcC-)* (35). All *Pseudomonas* strains were grown in KB media at 28°C and antibiotic selection was carried out using the following concentrations (μg ml^{-1}): kanamycin, 50; rifampicin, 100.

Bacterial Quantification and Disease Progression Experiments

Short-day-grown wild-type Col-0 and *met1* adult plants were assayed for pathogen response at 4-6 weeks old. Seed abortion of many F_1 homozygous *met1* plants results in segregation rate of approximately 5% and prevented us from assaying the F_1 plants in large quantities (17). Therefore, the subsequent *met1* generation (F_2) was used in all pathogen assays. In contast, we utilized F_1 *ddc* triple homozygous mutant plants from a segregating population for infection assays. The segregating *ddc* plants, as well as the wild-type Col-0 control plants, were grown under long day conditions (16 hr photoperiod at 22°C) in large quantities. At approximately 2 weeks old, control and mutant seedlings displaying a curled leaf phenotype characteristic of the homozygous *ddc* triple mutant (36) were transplanted, moved into short day conditions, genotyped, and then assayed for pathogen response at 4-6 weeks old.

Quantification of bacterial growth assays was performed in adult plants infected via vacuum infiltration of bacteria at 1×10^5 cfu ml^{-1} (OD_{600} = 0.0002) as previously described (37). At least 15 individual plants of each genotype were assayed at each time point. Leaf tissue from two individuals, representing a single technical replicate, was pooled and leaf disks (8-10) were removed and ground in 10 mM $MgCl_2$. Bacterial colony forming units (CFUs) were calculated as previously described (37) and data was plotted as the mean and standard error of the decimal logarithm (log[cfu cm^{-2}]) of approximately 8 replicates. For qualitative measurement of disease progression, plants were infected with *Pseudomonas syringae* pv. *tomato* DC3000 at 1×10^5 cfu ml^{-1} by vacuum infiltration and representative photographs were taken of uninfected leaves or leaves at 1, 3, or 5 days post infection.

For the transgenerational experiments, wild-type Col-0 and F_1 triple mutant *ddc* plants (approximately 50 plants of each genotype) were grown under short day conditions, infected with *Pseudomonas syringae* pv. *tomato* DC3000, and bacterial quantification assays were performed as above. Following the infection, plants were moved to long day conditions to induce flowering. The plants were allowed to recover and bolt before re-infecting the plants, thereby maintaining the pathogen stress for a longer period of time. Seed from all plants of the same genotype was pooled, surface sterilized, and the next generation was planted, assayed, and re-stressed in an identical fashion.

Real Time PCR Assays

Leaf disks were removed at each time point (approximately 2 disks/leaf) from at least 15 individuals infected with *Pseudomonas syringae* pv. *tomato* DC3000 as described above. Tissue was sampled from uninfected tissue and at 1, 2, 3, 4, and 5 days post infection and total RNA was isolated using the RNeasy plant mini kit (Qiagen). cDNA was generated using the SuperScript III kit (Invitrogen) and oligo(dT) primers according to manufacturer's instructions. qPCR reactions were run on a Applied Biosystems 7500 Real-Time PCR System using the Power SYBR Green PCR Mastermix kit (Applied Biosystems). All primer pairs are listed in Table 3.1. Ct values were generated using default parameters and relative expression values (using a *TUBα2* control) were calculated using the formula $2^{-((Ct_{Target}\ \mathrm{Treatment} - Ct_{TUB\alpha2}\ \mathrm{Treatment}) - ((Ct_{Target}\ \mathrm{Untreated} - Ct_{TUB\alpha2}\ \mathrm{Untreated}))}$. Fold change values were then normalized to the untreated, wild-type sample and the data is presented as the mean and standard error of three technical replicates.

Isolation of *Arabidopsis* Nuclei

Preparation of *Arabidopsis* nuclei was performed as previously described with modifications (38,39). Briefly, leaf tissue from uninfected or infected plants was harvested (approximately 90 adult plants, 50-80 g fresh weight). Leaves were placed in a 2 L Erlenmeyer flask with 1 L of sterile water and shaken at 225 rpm for 30 min at 4°C on a floor shaker. This step removes approximately 80% of the bacteria. The tissue was chopped into small pieces, stirred in diethyl ether (5 min, 4°C), and washed thoroughly in cold sterile water. Three volumes of resuspension buffer (1 M sucrose, 10 mM Tris-HCl pH 7.2, 5 mM $MgCl_2$, 5 mM 2-mercaptoethanol) were added and the tissue was homogenized using a Polytron homogenizer. The sample was filtered through 4 layers of cheesecloth, 2 layers of Miracloth, and centrifuged at 9500 rcf for 15 min. Resuspension buffer containing 0.5% Triton X-100 was added to the pellet and the crude nuclei were resuspended using a dounce homogenizer. The sample was re-centrifuged at 6700 rcf for 10 min, the pellet resuspended as before, and the sample was spun at 3800 rcf for 10 min. After resuspending the pellet, the sample was laid on top of a 35% / 60% Percoll (Sigma) gradient and spun in an SW-28 ultracentrifuge rotor (5 min at 1900 rpm, then increase to 7500 rpm for 15 min). The nuclei were removed from the 35% / 60% interface, diluted in 5-10 volumes of resuspension buffer, and centrifuged at 6700 rcf for 10 min. The nuclei were further purified with two sequential rounds of 35% / 60% Percoll gradients as before, except a SW41-Ti ultracentrifuge (5 min at 1900 rpm, then increase to 7700 rpm for 15 min) was used for these additional purification steps. The final nuclear fraction from the 3rd Percoll gradient was diluted 5-10 volumes in resuspension buffer, centrifuged at 3800 rcf for 10 min, and then used for genomic DNA isolation.

Preparation of MethylC-Seq Libraries

Arabidopsis nuclear genomic DNA was extracted using the Plant DNeasy Mini Kit (Qiagen) and 25 ng unmethylated *c*I857 *Sam7* Lambda DNA (Promega) was added to 5 µg of *Arabidopsis* gDNA. The DNA was fragmented to 50-500 bp by sonication using a Bioruptor (Diagenode). The DNA fragments were end repaired with a nucleotide triphosphate mix free of dCTP (End-It DNA End-Repair Kit, Epicentre Biotechnologies). Cytosine-methylated adapters provided by Illumina were ligated to the sonicated DNA as per manufacturer’s instructions for genomic DNA library construction. Adapter-ligated DNA molecules of 175-225 bp were isolated by 2% agarose gel electrophoresis and purified using a minElute Gel Extraction Kit (Qiagen). A sodium bisulfite conversion was performed on each sample using the MethylEasy *Xceed* kit according to the manufacturer’s instructions (Human Genetic Signatures). One quarter of the bisulfite-converted, adapter-ligated DNA molecules were amplified by 4 cycles of PCR with the following reaction composition: 2.5 U of uracil-insensitive

PfuTurboC$_x$ Hotstart DNA polymerase (Stratagene), 5 µl 10X *PfuTurbo* reaction buffer, 1 µl of 25 µM dNTPs, 1 µl of Illumina's Primer 1.1, and 1 µl of Illumina's Primer 2.1 (50 µl final). The thermocycling parameters were: 95˚C 2 min, 98˚C 30 sec, then 4 cycles of 98˚C 15 sec, 60˚C 30 sec and 72˚C 4 min, ending with one 72˚C 10 min step. The PCR reaction was purified using the MinElute PCR purification kit (Qiagen), separated by 2% agarose gel electrophoresis, and the amplified product (175-225 bp) was purified from the gel using the MinElute gel purification kit (Qiagen). Quantitative PCR was used to measure the concentration of viable sequencing template molecules in the library prior to sequencing.

Preparation of Strand-Specific mRNA-Seq Libraries

Prior to harvesting tissue for *Arabidopsis* nuclei preparations, some leaves (5-8 g) were collected, flash frozen in liquid nitrogen, and ground into a fine powder for later use in RNA experiments. Here, we have generated and sequenced mRNAs from a single biological replicate using a similar strategy as previously described (4). Briefly, total RNA was isolated from approximately 250 mg of frozen leaf powder using the mirVana miRNA Isolation kit (Ambion) according to the manufacturer's instructions for isolation of total RNA from plant tissue. mRNA was purified out of the total RNA samples with two sequential poly(A) selections using a Oligotex mRNA Mini kit (Qiagen) as per manufacturer's instructions. The mRNA was ethanol precipitated after each poly(A) selection. The purified mRNA (100-150 ng) was fragmented by metal hydrolysis in 1X fragmentation buffer for 15 min at 70˚C and quenched by addition of 2 µl fragmentation stop solution (Life Technologies). Fragmented RNA was treated with 5 U Antarctic phosphatase (New England Biolabs) for 40 min at 37˚C in the presence of 40 U RNaseOut (Life Technologies), followed by phosphatase heat inactivation at 65˚C for 5 min. Re-phosphorylation of the fragmented RNA ends was performed by addition of 10 U PNK (New England Biolabs), 1 mM ATP, and 20 U RNaseOut and incubation at 37˚C for 1 h. The RNA was then purified using 66 µl SPRI beads (Agencourt) and eluted in 11 µl of TE buffer. One µl of 1:10 diluted pre-adenylated 3' RNA adapter oligonucleotide (5'-UCGUAUGCCGUCUUCUGCUUGidT-3') was added to the phosphorylated RNA and incubated at 70˚C for 2 min to denature any RNA secondary structure followed by placement on ice. The 3' RNA adapter ligation reaction was performed by addition of 2 µl 10x T4 RNA ligase 2 truncated ligation buffer, 1.6 µl 100 mM $MgCl_2$, 20 U RNaseOut, and 300 U T4 RNA ligase 2 truncated (New England Biolabs) and then incubated at 22˚C for 1 h. Ligation of the 5' RNA adapter was performed by addition to the 3' adapter-ligated reaction 1 µl of 1:1 diluted, heat denatured (70˚C, 2 min) 5' RNA adapter oligonucleotide (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3'), 1 µl 10 mM ATP, and 10 U T4 RNA ligase (Promega), and incubation at 20˚C for 1 h. The RNA was purified using 66 µl SPRI beads (Agencourt) and eluted in 10 µl of TE buffer. To the RNA ligation products, 2 µl of 1:5 diluted RT primer (5'-CAAGCAGAAGACGGCATACGA-3') was added and heat denatured (70˚C, 2 min), followed by incubation on ice. To the denatured RNA/primer solution 4 µl of 5x first strand buffer, 1 µl 12.5 mM dNTPs, 2 µl 100 mM DTT, and 40 U RNaseOut was added, followed by incubation at 48˚C for 1 min. To this, 200 U Superscript II reverse transcriptase (Life Technologies) was added, followed by incubation at 44˚C for 1 h. The entire RT reaction was used in a PCR amplification reaction containing 0.25 µM GEX1 (5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3') and 0.25 µM GEX2 (5'-CAAGCAGAAGACGGCATACGA-3') primers, 0.25 mM dNTPs, 1x Phusion polymerase buffer, and 4 U Phusion hot-start high fidelity DNA polymerase (New England Biolabs) in a 100 µl reaction using the following thermocycling parameters: 98˚C 30 sec, then 15 cycles of 98˚C 10 sec, 60˚C 30 sec and 72˚C 15 sec, ending with one 72˚C 10 min step. The PCR products were purified in two steps, first by purification using 180 µl SPRI beads and elution in 30 µl of TE buffer, followed by purification with 39 µl SPRI beads and elution in 10 µl of TE buffer. All oligonucleotides were obtained from Illumina. Quantitative PCR was used to measure the concentration of viable sequencing template molecules in the library prior to sequencing.

High-Throughput Sequencing

MethylC-Seq and RNA-Seq libraries were sequenced using the Illumina Genome Analyzer II (GA II) as per manufacturer’s instructions. Sequencing of MethylC-Seq libraries was performed up to 87 cycles to yield longer sequences that are more amenable for unambiguous mapping to the *Arabidopsis thaliana* Col-0 genome reference sequence (TAIR9). Image analysis and base calling were performed with the standard Illumina pipeline (Firecrest v1.4 and Bustard v1.4), performing automated matrix and phasing calculations on a PhiX or genomic DNA control.

Data Analysis

Processing and Alignment of MethylC-Seq Read Sequences

Read sequences produced by the Illumina pipeline (v1.4) in FastQ format were pre-processed in three steps. Firstly, reads were trimmed to before the first occurrence of a low quality base (PHRED score ≤ 2). Secondly, as a subset of reads contained all or part of the 3’ adapter oligonucleotide sequence, every read was searched for the adapter sequence, and if detected the read was trimmed to the preceding base. If the full adapter sequence was not detected, iterative searching of the k 3’ terminal bases of the read for the k 5’ bases of the adapter was performed, and if detected the read was trimmed to the preceding base. Thirdly, any cytosine base in a read was replaced with thymine. Following pre-processing, reads were sequentially aligned using v0.10.0 of the Bowtie algorithm (40) to two computationally converted TAIR9 reference sequences, the first in which cytosines were replaced with thymines, and the second in which guanines were replaced with adenines. The 48,502 bp *c*I857 *Sam7* Lambda genome was included in the reference sequence as an extra chromosome so that reads originating form the unmethylated control DNA could be aligned. As all cytosines in the reads were replaced with thymines, the methylation status of a particular genomic sequence has no bearing on its ability to map to the reference. Sequences originating from the Watson strand of the genome aligned to the cytosine-free reference sequence, whereas sequences originating from the Crick strand (complement) of the genome aligned to the guanine-free reference sequence after reverse complementation. The following parameters were used in the Bowtie alignment process: --solexa1.3-quals -e 80 -l 20 -n 0 -k 10 --best –strata –p –chunkmbs 1024 --nomaground. For each read, up to 10 of the most highly scoring alignment positions in the reference sequences were returned, tolerating a maximum sum quality score of 80 at mismatch positions. All results of aligning a read to both the Watson and Crick converted genome sequences were combined, and if more than one alignment position existed for a read it was categorized as ambiguously aligned and disregarded. For each sample type, the reads from two biological replicates were pooled to provide greater coverage for identification of the methylcytosines that are presented in this study. Whole lanes of aligned read sequences were combined in a manner based on the experimental setup. As up to two independent libraries from each biological replicate were sequenced, we first removed reads that shared the same 5’ alignment position within each library, referred to as “clonal” reads, leaving the read at that position that had the highest sum quality score. Subsequently, the reads from all libraries of a particular sample were combined. All unambiguous, or “unique”, read alignments were then subjected to post-processing, which consisted of 3 steps. Firstly, if a read contained more than 3 mismatches compared to the reference sequence, it was trimmed to the base preceding the fourth mismatch. Secondly, the cytosines that were originally removed from the read sequences prior to alignment were incorporated back into the aligned reads. Thirdly, to remove reads that were likely not bisulfite converted, reads that contained more than 3 cytosines in a non-CG context were discarded. Finally, the number of calls for each base at every reference sequence position and on each strand was calculated. Read number for each replicate before and after removal of clonal reads and post-processing is detailed in Table 3.2.

Identification of Methylated Cytosines

At each reference cytosine the binomial distribution was used to identify whether at least a subset of the genomes within the sample were methylated, using a 0.01 FDR corrected *P*-value. Each context of methylation was considered independently: CG, CHG or CHH (where H = A, C or T). We identified methylcytosines while keeping the number of false positives methylcytosine calls below 1% of the total number of methylcytosines we identified. The probability *p* in the binomial distribution B(n,*p*) was estimated from the number of cytosine bases sequenced in reference cytosine positions in the unmethylated Lambda genome (referred to as the error rate: non-conversion plus sequencing error frequency). The bisulfite conversion rates for all samples were approximately 99%, and the error rates were as follows: Untreated replicate 1, 0.0130; Untreated replicate 2, 0.0067; Untreated combined replicates, 0.0106; *Pst* 5 d.p.i. replicate 1, 0.0137; *Pst* 5 d.p.i. replicate 2, 0.0068; *Pst* 5 d.p.i. combined replicates, 0.0111. We interrogated the sequenced bases at each reference cytosine position one at a time, where read depth refers to the number of reads covering that position. For each position, the number of trials (n) in the binomial distribution was the read depth. For each possible value of n we calculated the number of cytosines sequenced (k) at which the probability of sequencing k cytosines out of n trials with an error rate of p was less than the value M, where (M * (number of unmethylated cytosines)) < (0.01 * (number of methylated cytosines)). In this way, we established the minimum threshold number of cytosines sequenced at each reference cytosine position at which the position could be called as methylated, so that out of all methylcytosines identified no more than 1% would be due to the error rate. The specifics of the mC analysis can be found in Table 3.3.

Identification of Hypo and Hypermethylated Blocks

A sliding window approach was used to identify blocks of DNA methylation enrichment throughout the genome. A window size of 100 bp was used and progressed at 100 additional bases per iteration. When a 100 bp window containing at least 4 mCs was identified, the block was extended in 300 bp increments until an increment was reached that contained no mCs. These blocks were then refined to start at the first mC of the block and end at the last mC. Blocks were identified using a union of the untreated (0 dpi) and treated (5 dpi) mC calls so that no regions were missed in the analysis. After block identification, the number of mCs contained in a given block were counted and summed for each individual data set. A ratio of each sum (number of 0dpi block mCs/number of 5dpi block mCs) was calculated to identify pathogen-induced hypo and hypermethylated DNA regions.

Mapping RNA-Seq Reads

Read sequences produced by the Illumina analysis pipeline were aligned with the ELAND algorithm to the TAIR9 reference sequence. Reads that aligned to multiple positions were discarded. Reads per kilobase of transcript per million reads (RPKM) were calculated to quantify transcriptional levels. To calculate a transcriptional fold change for a given gene, a ratio between the two RPKMs from the two samples was calculated.

Data Visualization in the AnnoJ Browser

MethylC-Seq and RNA-Seq sequencing reads, as well as all methylcytosines with respect to the TAIR9 reference sequence, gene models and functional genomic elements were visualized in the AnnoJ 2.0 browser, as described previously (4).

RESULTS

Resistance to *Pseudomonas syringae* is Enhanced by Global Loss of DNA Methylation

Direct regulation of transcriptional output by DNA methylation is well established, and there are multiple examples of DNA methylation-dependent developmental programs in plants. Temporal and structural floral development are, in part, regulated by cytosine methylation of the *FWA* and *SUP* genes, respectively, and mutant epi-alleles at these loci display clear flowering phenotypes (41,42). Additionally, global depletion of non-CG methylation in *drm1 drm2 cmt3* triple mutants results in a striking curled leaf phenotype that is solely attributed to loss of promoter methylation, and subsequent upregulation, of an F-box gene (36). Direct regulation of gene transcription by cytosine methylation is likely a wide spread phenomena, as loss of function mutations in *MET1* or *DRM1/DRM2/CMT3* result in misregulation of 3.0% or 1.2% of all *Arabidopsis* genes, respectively (4).

We sought to investigate the contribution of cytosine methylation in regulation of plant defense against the bacterial phytopathogen *Pseudomonas syringae*. Virulent strains, including *P. syringae* pv. *tomato* DC3000 (*Pst* DC3000), suppress PAMP-induced basal defense responses through the action of type III effectors, while non-host pathogens, like the type III secretion mutant *Pst* DC3000 (*hrcC-*), are effectively controlled by the plant's basal defenses. In contrast, avirulent strains, like *Pst* DC3000 (*avrPphB*), express an avirulence (*Avr*) gene that is recognized by the plant R protein defense system, resulting in an amplified resistance response. To resolve any DNA methylation-dependent regulation of the plant defense system, we tested plants globally depleted in CG methylation (*met1-3*) or non-CG methylation (*ddc*, *drm1-2 drm2-2 cmt3-11*) against infection with virulent, avirulent, or non-host pathogens (Figure 3.1). Remarkably, at 3 days post infection (dpi) the levels (or colony forming units, cfu) of the virulent *Pst* DC3000 strain are approximately 2.5 and 2 orders of magnitude lower in *met1* and *ddc* mutants compared to wild-type plants, respectively (Figure 3.1.A and Figure 3.1.B, left panels). The *met1* and *ddc* enhanced resistance to *Pst* DC3000 phenotype was observed throughout the time course of the infection, as illustrated by the lower levels of bacteria at each time point in the growth assays, as well as the differential progression of disease symptoms in the methylation mutants (asymptomatic) compared to wild-type plants (chlorosis/necrotic lesions at 3 and 5 dpi, Figure 3.1.C). Although difficult to resolve due to low levels of bacterial growth, we also detected enhanced resistance to avirulent (*Pst* DC3000 (*avrPphB*)) and non-pathogenic (*Pst* DC3000 (*hrcC-*)) strains of *P. syringae* in the *met1* and *ddc* mutants (Figure 3.1.A and Figure 3.1.B, center and right panels). Together, these data indicate that wild-type *Arabidopsis* plants possess a genetically functional, but epigenetically repressed, heightened defense response against virulent, avirulent, and non-pathogenic strains of *P. syringae*.

We have previously demonstrated that global disruption of CG or non-CG methylation results in a wide range of transcriptional changes, and it is possible that the enhanced resistance to *P. syringae* occurs independently of the traditional plant defense pathways. For example, both *met1* and *ddc* mutants show distinct developmental abnormalities that may be non-specifically responsible for the pathogen phenotype. To address this possibility, we performed real time PCR assays on infected *met1* and *ddc* leaf tissue and examined expression levels of genes that are traditionally involved in the plant defense response against bacterial pathogens (Figure 3.2). We observed strong activation of salicylic acid (*PR1*, Pathogenesis-related 1) and flagellin (*FRK1*, Flagellin Receptor Kinase 1) signaling pathways, as well as suppression of the Jasmonic acid/ethylene responsive plant defensin gene (PDF1.2), in wild-type, *met1*, and *ddc* plants in response to *Pst* DC3000. Interestingly, wild-type, *met1*, and *ddc* plants all display both temporal and amplitude of expression differences for each defense gene examined. Expression levels of the salicylic acid (SA) responsive gene *PR1*, a reliable indictor of plant defense levels, increases more rapidly in *met1* plants compared to wild-type (Figure 3.2, top left panel), suggesting that SA signaling is primed for rapid activation in *met1* plants. In contrast, *PR1* and *FRK1* transcript levels in *ddc* plants demonstrate a similar temporal, but a 2.5 fold higher, induction of gene expression compared to wild-type plants (Figure 3.2, bottom panels). Additionally, we observed an immediate suppression of the SA-antagonized *PDF1.2* gene in *met1* plants, while expression levels in *ddc* plants follow that of wild-type, indicating differential activation of SA signaling in the two mutants. While rapid SA induction at the onset of infection in *met1* plants is a likely indictor of a primed plant defense that results in rapid suppression of bacterial growth, *ddc* plants utilize an increased activation of SA signaling to control pathogen growth throughout infection. Interestingly, both DNA methylation mutants displayed prolonged upregulation of two heavily CG and non-CG methylated *R* genes, *RPP4* (Recognition of *Peronospora parasitica* 4) and *SNC1* (Suppressor of *npr1-1*, Constitutive 1), during infection (Figure 3.2, right panels and data not shown), a possible consequence of either hyper-activation of SA signaling or loss of regulatory gene body DNA methylation.

Transgenerational Memory of Pathogen Stress

Our examination of the DNA methylation mutants indicates that some defense genes are held under the strict control of cytosine methylation, suggesting that stable pathogen-induced alterations in DNA methylation may be passed to subsequent generations to encode a systemic enhanced resistance to bacterial pathogens. To examine DNA methylation-dependent transgenerational memory of pathogen stress, we performed *P. syringae* pv. *tomato* DC3000 infections in wild-type and *ddc* mutant plants throughout multiple generations, measuring the pathogen levels by bacterial enumeration at each individual generation (Figure 3.3). Importantly, we simultaneously grew control plants that were treated identically to the infected plants but were not exposed to pathogen. Interestingly, we did not detect a significant difference between control and stressed Col-0 plants in the second generation (S2); however, we found that the stressed *ddc* plants were approximately an order of magnitude more susceptible to pathogen than the control *ddc* plants (S2 plants, Figure 3.3). Remarkably, in the following generation (S3), we found that wild-type Col-0 plants were 3 fold, or ½ log, more resistant than the control plants (*p*=0.002), indicating that after multiple generations of infection, transgenerational memory of pathogen stress results in enhanced resistance to virulent bacteria. It is notable that we also observed hyper-susceptibility to *Pst* DC3000 in the S3 *ddc* plants at levels similar to those observed in the S2 generation, directly implicating DNA methylation in transgenerational memory. Together, these data support the notion that DNA methylation contributes to transgenertional memory of pathogen stress and, to our knowledge, represent the first direct example of heritable resistance to bacterial pathogens.

Transient DNA Methylation Changes in Response to Virulent *P. syringae*

Dynamic or transient changes in cytosine methylation levels have been observed in human cells (23,24,26) and it is likely that analogous mechanisms exist in plants (27-32). To resolve pathogen-induced transient alterations in DNA methylation, we performed genome wide bisulfite sequencing (methylC-Seq, Lister *et al.* (4)) on two independent biological replicates of untreated and pathogen-treated (*P. syringae* pv. *tomato* DC3000) *Arabidopsis* plants. Leaf tissue from wild-type Col-0 plants was harvested prior to infection (untreated) or 5 dpi (treated) and prepared for high-throughput sequencing. Infected leaf tissue, however, is heavily composed of bacterial cells (approximately 50 million cells per cm^2 leaf tissue). To enrich for *Arabidopsis* genomic DNA, we biochemically isolated crude nuclei by differential centrifugation, and subsequently purified the crude fraction with three consecutive rounds of discontinuous Percoll gradients. Using this enrichment strategy, sequencing reads of *P. syringae* origin represent only 0.08% of the total uniquely mapping reads in the *Pst*-treated sample (data not shown). We sequenced adaptor-ligated inserts up to 87 bases long using the Illumina Genome Analyzer II, and generated 58 and 66 million non-clonal reads, representing 4.3 and 4.8 gigabases (Gb), that uniquely map to the *Arabidopsis* genome for the untreated and treated samples, respectively (Pooled data of two biological replicates, Table 3.2). Our sequencing covered approximately 95% of the entire nuclear genome (94% of all nuclear cytosines) with at least two reads, and we reached an average read depth of 18 and 20 reads per nucleotide per strand for the untreated and treated samples, respectively (Table 3.3).

The binomial distribution was used at every cytosine, examining each sequence context (CG, CHG, CHH) independently, to determine if methylation was detected at significant levels using a mC false discovery rate of 1%. Using this approach, we identified 4,369,039 and 4,336,818 methylated cytosines, representing 10.9% and 10.8% of the nuclear cytosines covered by at least 2 reads, in the untreated and *Pst*-treated plants, respectively (Table 3.3). Furthermore, we observed the majority of DNA methylation in leaf tissue within the non-CG context (mCG, 37%; non-mCG, 63%) for both samples, and a strikingly large amount of CHH methylation (42% of total mCs, Table 3.3). Surprisingly, a much larger proportion of *Arabidopsis* leaf methylcytosines were found in the non-CG context compared to floral tissue (45% non-mCG; Lister *et al.* (4)), possibly representing global differences in DNA methylation patterns between the two tissue types.

Upon initial inspection of the DNA methylomes of the untreated and *Pst*-treated samples, we found no significant global differences in the number of methylcytosines detected or the sequence context in which they were found, indicating that pathogen-induced whole genome re-patterning of DNA methylation is an unlikely mechanism of epigenetic regulation. To further examine transient alterations of DNA methylation upon *P. syringae* treatment, we examined differential methylation across the genome in the context of DNA methylation-enriched blocks. Using a sliding window approach, we identified 29,003 blocks of DNA methylation in the *Arabidopsis* genome that ranged in size from 4 to 152,777 base pairs. We generated a DNA methylation ratio (untreated mCs/treated mCs) for each block and identified 157 pathogen-induced hypomethylated blocks and 147 hypermethylated blocks using a 3-fold-change threshold (Figure 3.4.A). These regions of differential DNA methylation are exemplified by an intronic region of the At4g011030 gene (encodes putative lipase with unknown function), where we observed regions of *Pst* DC3000-induced hyper and hypomethylation (Figure 3.4.B). Interestingly, we found differential methylcytosines (dmCs) in every sequence context (dmCG, dmCHG, dmCHH; Figure 3.4.B), indicating that transient DNA demethylation or hypermethylation within a given region is not restricted to a sequence-specific enzymatic activity. Although additional experiments are needed to validate these putative regions of pathogen-induced cytosine methylation changes, it is likely that we have identified several biologically relevant regions of DNA methylation re-patterning that may encode transcriptional regulation of proximal genes within the somatic tissue.

To examine the functional consequences of pathogen-induced alterations in DNA methylation patterns, we performed strand-specific sequencing of *Arabidopsis* mRNAs (mRNA-Seq; Lister *et al.* (4)) from untreated (0 dpi) and *Pst* DC3000-treated (5 dpi) plants. We generated approximately 27 and 19 million reads from mRNA-Seq libraries representing the untreated and pathogen-treated plants, respectively, and found 2,066 up-regulated and 4,219 down-regulated genes in plants infected with *Pst* DC3000 (greater than 3-fold change, data not shown). To resolve the consequences of transient DNA methylation changes on transcriptional output, we are applying integrative approaches to globally address this question. Simple visual scanning of the hypomethylated blocks, however, has revealed that a correlation between differentially methylated blocks and altered transcriptional output does exist in some cases, exemplified by At2g29460 (Figure 3.5), which encodes a pathogen-responsive glutathione S-transferase gene (*Pst* DC3000, 80 fold induction at 24 hours post infection, NASCArrays-120, Craigon *et al*. (43)). Interestingly, we observed a 4-fold loss of cytosine methylation within the putative promoter region (also a genomic repetitive element) of At2g29460 after *Pst* DC3000 infection, as well as a concomitant increase in transcript levels (23-fold, Figure 3.5). Together, our data support a model of active methylation/demethylation of specific genomic loci upon pathogen treatment, and suggest that alterations in methylation patterns may directly influence transcriptional output of nearby genes.

DISCUSSION

Single-base resolution of the *Arabidopsis* methylome has enabled unprecedented insight into the biology of DNA methylation. These recent discoveries include unraveling the dynamic genome wide relationship between cytosine methylation, small RNAs, and transcriptional output, as well as novel perspectives into specific developmental and evolutionary processes, from demethylation-dependent control of genomic imprinting in the endosperm to rationale for the striking depletion of cytosine methylated transposable elements within gene-rich regions (4,44-46). Here, we provide compelling evidence supporting a role for DNA methylation in controlling specific components of the *Arabidopsis* defense system against the plant pathogen *Pseudomonas syringae*. Remarkably, we found that mutants deficient in either CpG methylantion (*met1*) or non-CpG methylation (*ddc*, *drm1 drm2 cmt3*) are significantly more resistant, compared to wild-type plants, to a variety of *P. syringae* strains, indicating that some aspect of the plant defense system is under epigenetic control. Interestingly, wild-type plants suppress this unknown component, possibly because prolonged misregulation of this defense element has consequences of the plant's overall fitness, as has been observed in *Arabidopsis* hybrids that acquire deleterious epistatic interactions that result in a plant "autoimmune" response (47). Although it is possible that the enhanced disease resistance phenotype observed in the DNA methylation mutants is a function of multiple altered epistatic interactions or a result of polygenic contributions, it is also feasible that misregulation of a single gene could result in such a dramatic phenotype, a phenomena observed with the *ddc* curled leaf phenotype which is controlled by a single F-box gene (36).

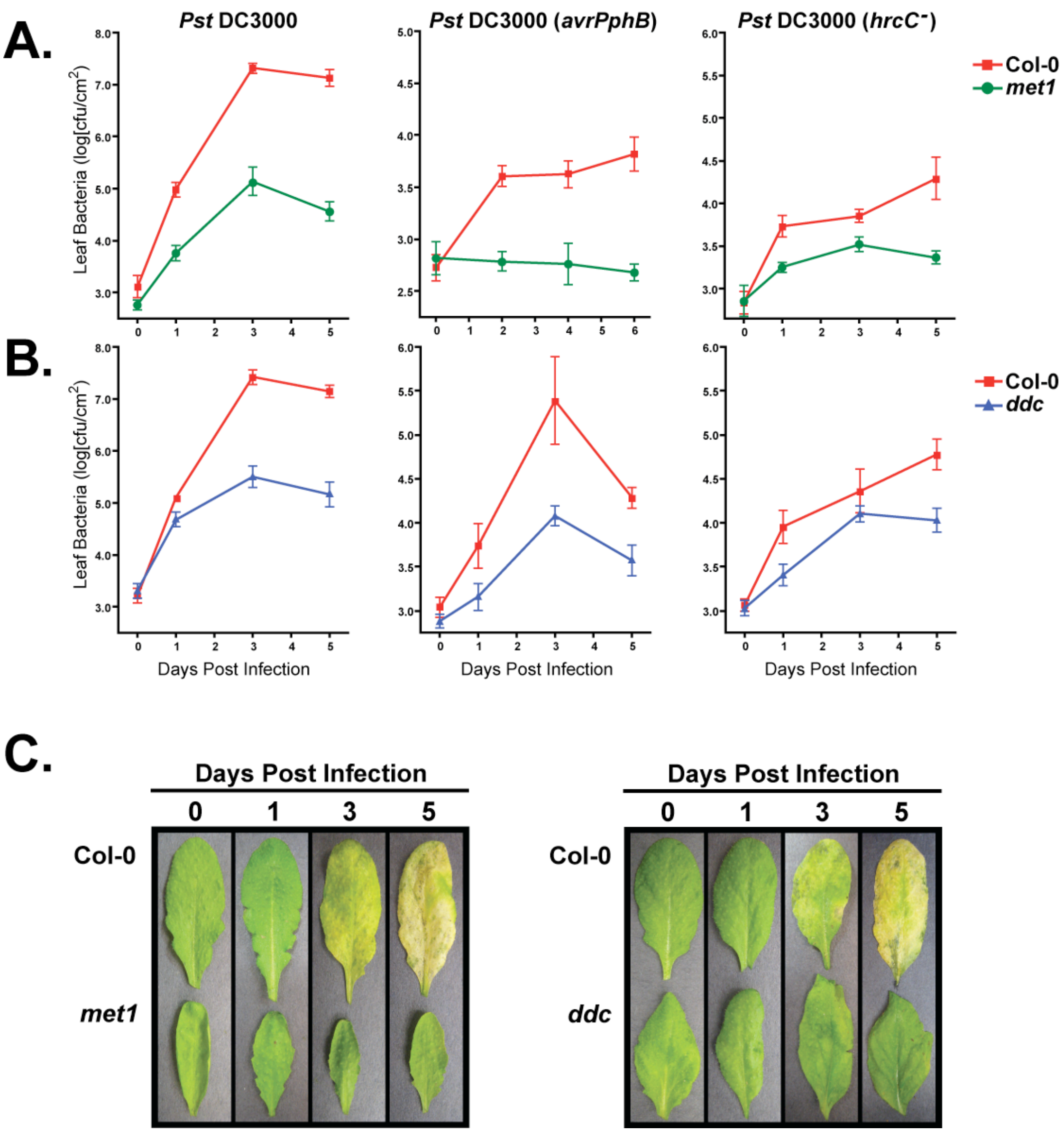
Interestingly, the *Arabidopsis* pathogen response has been recently investigated in a handful of mutants deficient in miRNA biogenesis and RNA-induced gene silencing. Post-transcriptional gene silencing, a *DCL1/HEN1/AGO1* dependent pathway in plants, functions in miRNA biogenesis and is required for full defense against non-pathogenic bacteria; and, not surprisingly, is a target of suppression by bacterial type III effector proteins (48). Bacterial pathogens likely inhibit these pathways because certain miRNAs are required for regulation of defense gene transcripts, as has been observed with PAMP-induced miRNA393, which contributes to resistance against virulent *P. syringae* strains (49). The RNA-directed DNA methylation (RdDM) pathways use non-overlapping, but functionally analogous, components, including the DCL3 and AGO4 proteins that are responsible for feeding small RNAs to the DRM1/DRM2/CMT3 cytosine methyltransferases. Although *dcl3* mutants show no detectable pathogen phenotype, a likely product of redundancy in the Dicer-like family, the *ago4* mutants are striking hyper-susceptible to the virulent *Pst* DC3000 strain (data not shown, Agorio *et al.* (50)). The discrepancy in pathogen phenotypes observed in *ago4* and *ddc* mutants suggests that these proteins act in a non-linear fashion to control resistance, and that there may additional contributing factors or pathways that direct the defense response. Indeed, AGO4 has been shown to interact with 24-nt siRNAs that are generated from hairpin RNAs independent of the traditional RdDM pathway (RNA Pol IV and *RDR2* dependent), as well as some specific miRNAs (21-nt) that preferentially bind AGO4 over AGO1 (51,52). We, as well as other groups, have demonstrated that RdDM and AGO4-dependent pathways play a key role in regulating the plant defense system, and consequently, may represent ideal targets for manipulation by bacterial type III effector proteins.

Remarkably, we found that wild-type plants utilize a transgenerational memory of virulent *P. syringae* infection to encode resistance in their progeny, a response that requires the activity of the DRM1, DRM2, and CMT3 cytosine methyltransferases. These data represent the first direct evidence supporting a cytosine methylation-dependent heritable resistance against bacterial pathogens. However, stress-induced genome instability, characterized by increased rates of DNA recombination in the somatic tissue, has also been linked to pathogen infection (53-55). Local activators of plant defense, including the PAMP elicitor flagellin, may induce systemic signals that result in increased levels of DNA recombination in uninfected tissues, as well as influence somatic recombination rates in subsequent generations (53,55). Furthermore, infection of tobacco leaves with tobacco mosaic virus (TMV) appears to induce systemic signals that result in DNA rearrangements surrounding several *R* gene loci, a process that may correlate with hypomethylation of these regions (56). The signals that propagate pathogen-induced alterations in DNA recombination rates and methylation patterning within the somatic tissues remains unclear, as well as the mechanism that transmits this information into the germline for use in subsequent generations; however, it is notable that our experimental approach to examine transgenerational memory exposes all aerial tissues, including the germline progenitor cells, to *P. syringae*. Further investigation of pathogen-induced systemic signals, using localized infection techniques, will be needed to determine how epigenetic information, including locus-specific DNA methylation changes, is transmitted into germline.

We have begun to apply sequencing-based genomic approaches to profile pathogen-induced transient DNA methylation changes (methylC-Seq), coupled with transcriptome analysis (mRNA-Seq), to resolve the dynamic relationship between cytosine methylation and mRNA output. Our initial analysis has identified a multitude of genomic regions displaying either hypo or hypermethylation upon *P. syringae* infection, however, further validation of these regions is needed. Additionally, we are currently applying methylC-Seq techniques to profile DNA methylation after activation of *R* gene defenses, which often initiates systemic acquired resistance (SAR) signaling pathways, using the avirulent strain *Pst* DC3000 (*avrPphB*). It is likely that activation of *R* gene defenses, or repetitive infection throughout consecutive generations, may result in stronger or more widespread changes in DNA methylation and provide additional insight into this complex component of the defense response. Finally, profiling of the *met1* and *ddc* transcriptomes during pathogen infection represents an alternative strategy for identification of both constitutive and pathogen-inducible defense genes that are under control of DNA methylation. We have utilized a combinatorial approach to examine the function of cytosine methylation in regulation of the plant defense system against *Pseudomonas syringae*, uncovering a role for DNA methylation in both transient and heritable epigenetic regulation of plant resistance response.

Figure 3.1 *Arabidopsis* mutants deficient in either maintenance or *de novo* DNA methylation are less susceptible to virulent, avirulent, and non-pathogenic bacteria.

(A) Adult wild-type Col-0 (red) or *met1* mutant (green) plants were infected with either virulent (*P. syringae* pv. *tomato* DC3000*, Pst* DC3000), avirulent (*Pst* DC3000 *(avrPphB)*), or non-host (*Pst* DC3000 *(hrcC-)*) bacteria at 1×10^5 cfu ml^{-1} by vacuum infiltration. At the indicated time points, infected leaf tissue was harvested and the bacterial colony forming units were quantified. Data is represented as the mean ± SE of the decimal logarithm (log[cfu cm^{-2}]) of at least 8 technical replicates. The experiment was repeated twice with similar results.
(B) Adult wild-type Col-0 (red) or *ddc* mutant (blue) plants were infected and analyzed as described in (A).
(C) Representative photographs of disease symptoms at the indicated time points in wild-type Col-0, *met1* (left panel), or *ddc* (right panel) leaves after infection with virulent *Pst* DC3000 (1×10^5 cfu ml^{-1}).


A.
Pst DC3000
Pst DC3000 (avrPphB)
Pst DC3000 (hrcC-)
Col-0
met1
Leaf Bacteria (log[cfu/cm^2])
3.0
4.0
5.0
6.0
7.0
8.0
2.5
3.5
4.5
5.5
0
1
2
3
4
5
6
B.
Col-0
ddc
Days Post Infection
C.
Days Post Infection
0
1
3
5
Col-0
met1
Col-0
ddc

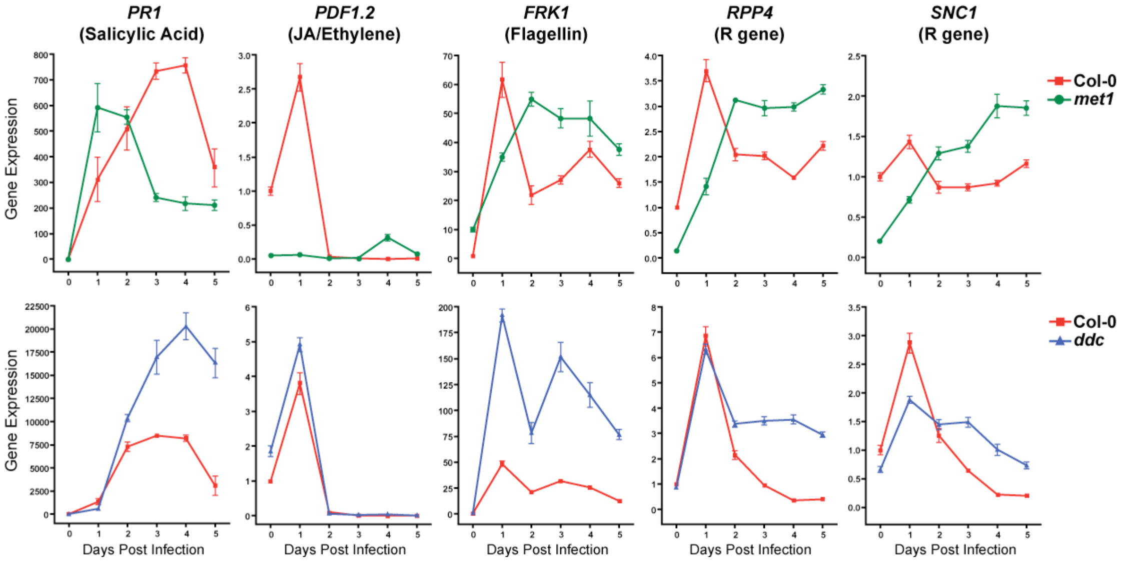


Figure 3.2 The *Arabidopsis met1* and *ddc* mutants show enhanced expression of certain defense genes.

Real time PCR analysis was performed on tissue from adult plants (wild-type Col-0, red; *met1*, green; *ddc*, blue) infected with virulent *P. syringae* pv. *tomato* DC3000 (1×10^5 cfu ml^{-1}) by vacuum infiltration. Genes regulated by the major plant defense signaling hormones (salicylic acid, jasmonic acid/ethylene, and flagellin), as well as two Resistance (*R*) genes that possess large amounts of gene body CG and non-CG methylation, were chosen for analysis. At least 15 individual plants were pooled and relative expression levels were calculated based on a *TUB*α*2* control, and all the values were normalized to the wild-type, untreated sample (0 Days Post Infection).

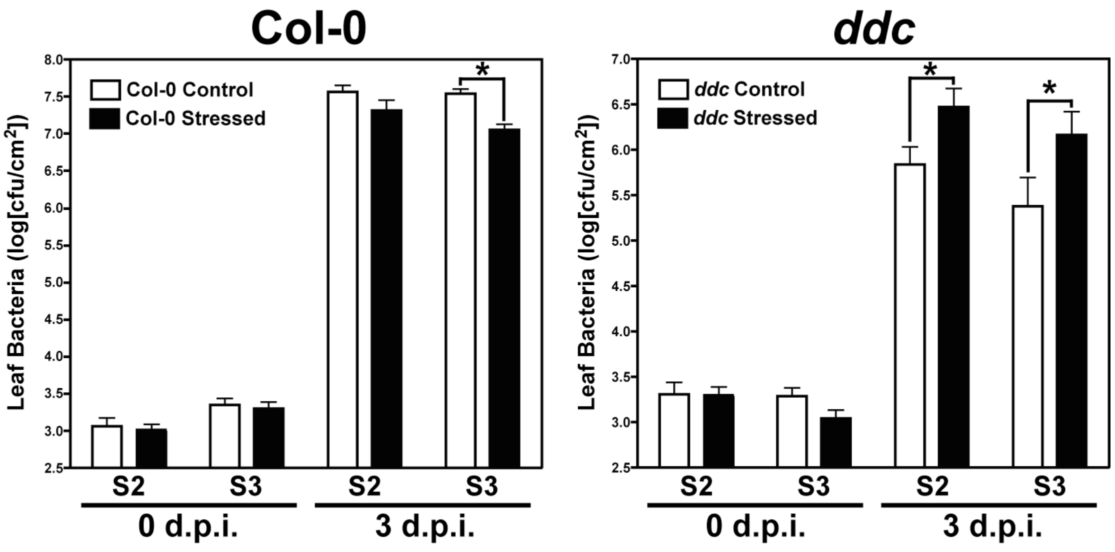
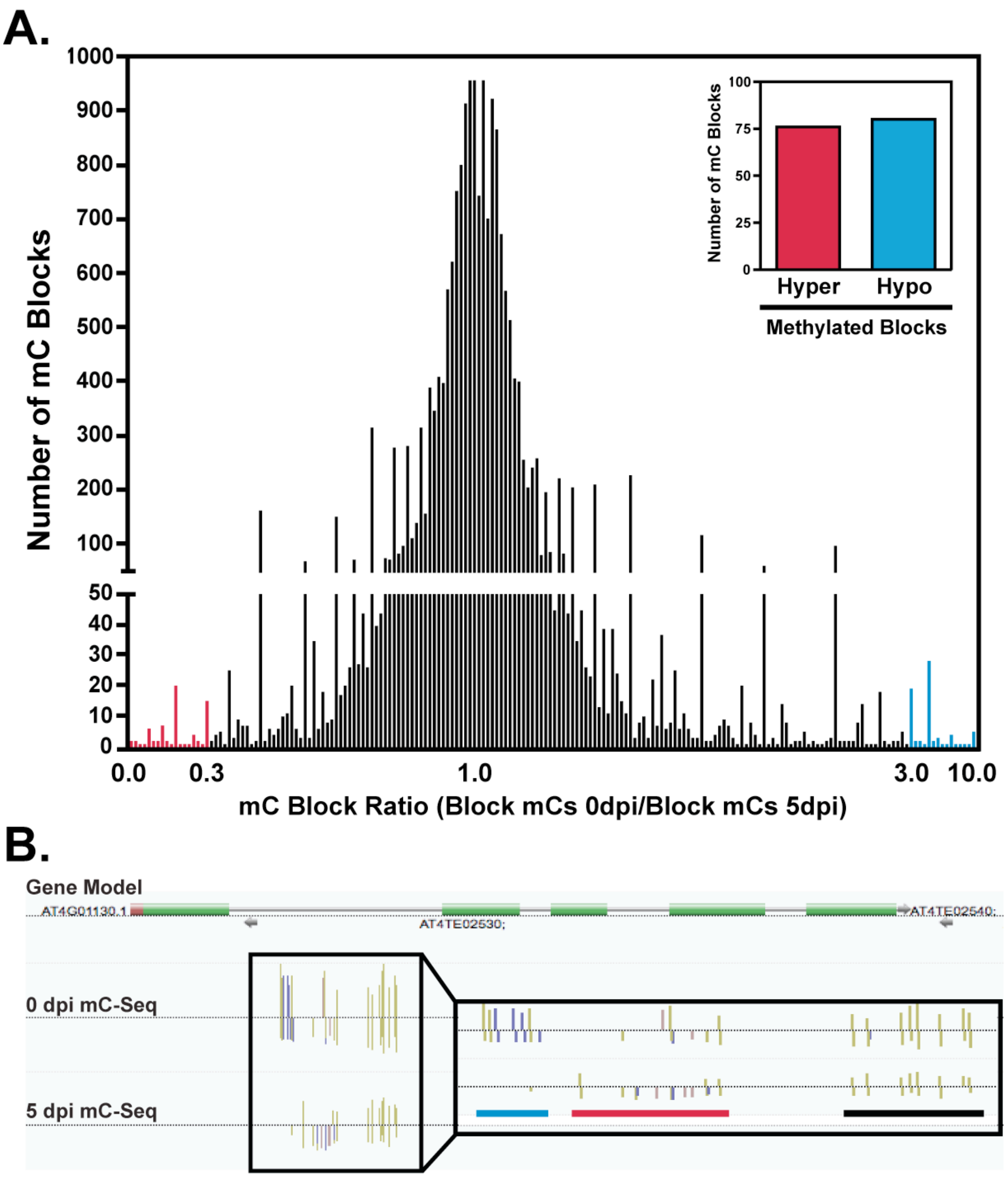


Figure 3.3 Transgenerational memory of virulent *P. syringae* infection requires the *de novo* DNA methytransferases.

Wild-type Col-0 (left panel) or *ddc* mutant (right panel) adult plants were infected with virulent *Pst* DC3000 bacteria (1×10^5 cfu ml^{-1}) by vacuum infiltration and bacteria were quantified at 0 or 3 days post infection as described in Figure 3.1.A. Plants were allowed to recover for approximately 2-3 weeks, re-infected with *Pst* DC3000 to maintain pathogen stress, and then allowed to fully senesce. The subsequent two generations (S2 and S3) were treated and assayed in an identical fashion. Control plants were grown in an identical fashion but were not exposed to bacterial pathogen. Plants stressed with pathogen at each generation (black bars) are compared to control plants (white bars) and an asterisk indicates p<0.05.

Figure 3.4 Detection of hypo and hypermethylated DNA regions in response to virulent *P. syringae* infection.

(A) Blocks of mC enrichment were identified across the genome (see Experimental Procedures) and mCs from the untreated (0 dpi) and treated (5 dpi) samples were counted and summed within each block. A ratio of the sums was calculated as 0 dpi mCs/5 dpi mCs for each block and the number of blocks having the same ratio were counted and plotted. The blocks were categorized as unchanged (black), hypomethylated (blue), or hypermethylated (red) based on a >3 fold change threshold. Also, blocks that displayed complete hypomethylation (blue, zero mCs in 5 dpi block) or hypermethylation (red, zero mCs in 0 dpi block) are shown in the inset graph.
(B) An example of an intronic region that shows active pathogen-induced hypomethylation (blue) and hypermethylation (red), as well as an unchanged block (black). mC calls: mCG, gold; mCHG, blue; mCHH, pink.


A.
1000
900
800
700
600
500
400
300
200
100
50
40
30
20
10
0
Number of mC Blocks
0.0
0.3
1.0
3.0
10.0
mC Block Ratio (Block mCs 0dpi/Block mCs 5dpi)
100
75
50
25
0
Number of mC Blocks
Hyper
Hypo
Methylated Blocks
B.
Gene Model
AT4G01130.1
AT4TE02530;
AT4TE02540;
0 dpi mC-Seq
5 dpi mC-Seq

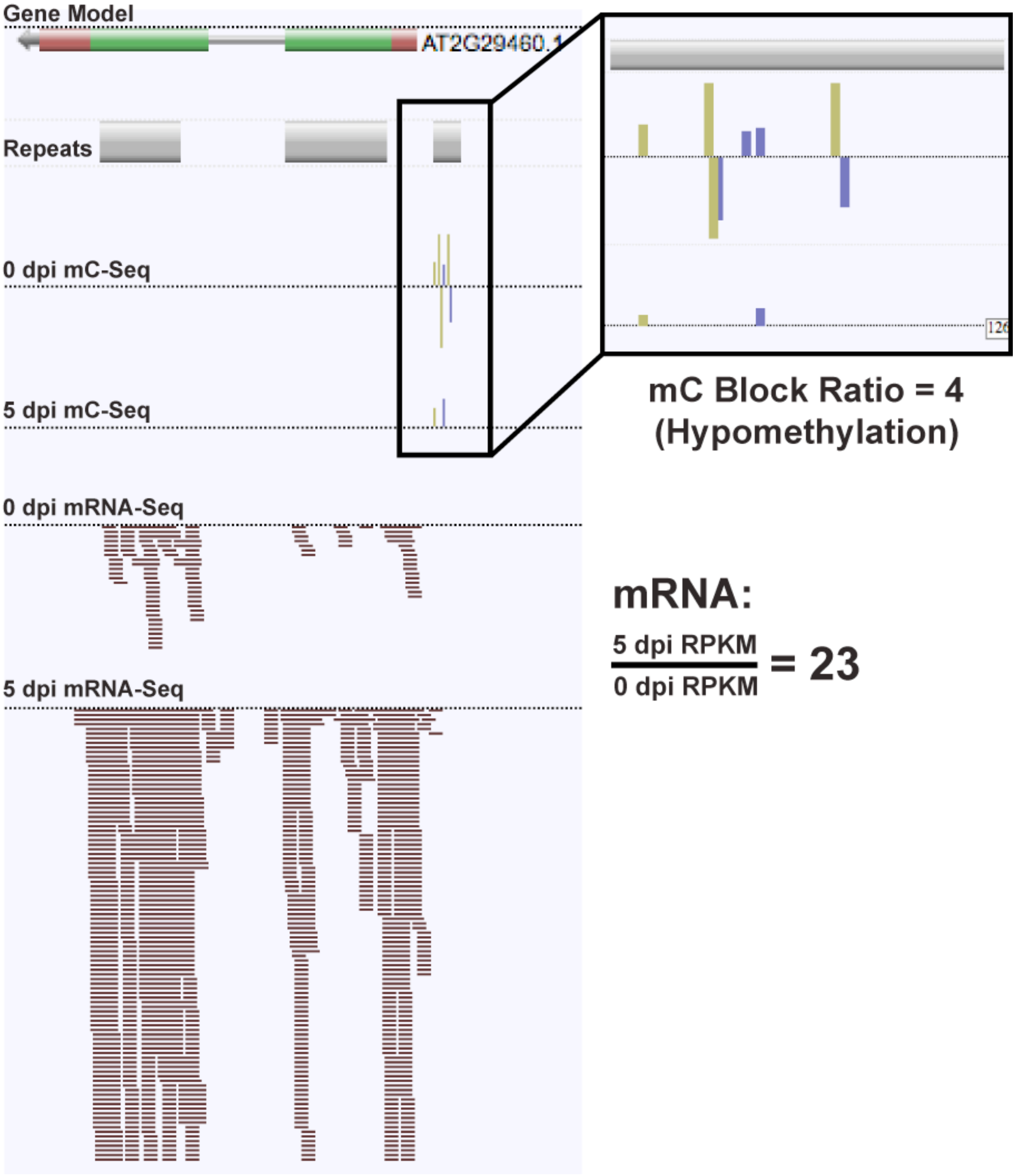


Figure 3.5 Pathogen-induced promoter hypomethylation correlates with transcriptional up-regulation of At2g29460.

The At2g29460 gene model and the surrounding genomic DNA repeats, as well as methylC-Seq data (mC calls) and mRNA-Seq data (reads) for each sample are displayed. An enlargement of the DNA methylation within the putative promoter region is shown (right) along with the calculated mC block ratio. mC calls: mCG, gold; mCHG, blue; mCHH, pink. A ratio of the RPKM (reads per kilobase of transcript per million reads) values for each mRNA-Seq sample was also calculated.

Table 3.1 Gene targets and primer sets used in Real Time PCR experiments.

Gene	***At* ID**	**Primer Name**	**Primer sequence (5' - 3')**
TUBa2	AT1G50010	5_TUBa2_qPCR	ATCTCTTGCTTGCGGTAG
-	-	3_TUBa2_qPCR	ACCCAGCTTAAATTCAGTTCTTGG
PR1	AT2G14610	5_PR1_qPCR	AGGCTAACTACAACTACGCTGCG
-	-	3_PR1_qPCR	GCTTCTCGTTCACATAATTCCCAC
PDF1.2	AT5G44420	5_PDF1.2_qPCR	GTTCTCTTTGCTGCTTTCGAC
-	-	3_PDF1.2_qPCR	GCAAACCCCTGACCATGT
FRK1	AT2G19190	5_FRK1_qPCR	GAGACTATTTGGCAGGTAAAAGGT
-	-	3_FRK1_qPCR	AGGAGGCTTACAACCATTGTG
RPP4	AT4G16860	5_RPP4_qPCR	GGGAGGATCTTCGGAACG
-	-	3_RPP4_qPCR	TTCCGACGAAGTCACCAAA
SNC1	AT4G16890	5_SNC1_qPCR	CCGGATATGATCTTCGGAAA
-	-	3_SNC1_qPCR	AACATCCTCGGCAAGCTCT

Table 3.2 MethylC-Seq library read numbers for each replicate before and after removal of clonal reads and post-processing.

Sample	Library	Mapped reads	Clonal reads removed	Post-processed
Untreated replicate 1	A	31,656,026	3,044,589	25,593,283
	B	11,540,210	1,070,572	9,458,437
Untreated replicate 2	A	30,057,886	3,472,057	23,180,770
Untreated combined	All	73,254,122	7,587,218	58,232,490
Total Bases				**4,297,500,701**
Pst 5 d.p.i. replicate 1	A	31,668,914	2,262,243	27,160,359
	B	13,303,844	1,071,665	11,147,914
Pst 5 d.p.i. replicate 2	A	32,842,126	2,336,495	28,173,371
Pst 5 d.p.i. combined	All	77,814,884	5,670,403	66,481,644
Total Bases				**4,807,642,269**

Table 3.3 Whole nuclear genome cytosine coverage and methylcytosine count.

		Untreated	Treated *(Pst* 5 dpi)
Lambda non-conversion + error frequency		1.06%	1.11%
% coverage of genome (>1 read)		94.88%	95.19%
% coverage of nuclear cytosines (>1 read)		93.60%	94.07%
Average read depth per base (Watson / Crick)		18.07 / 18.05	20.30 / 20.28
Number of nuclear methylcytosines (no depth restriction)		4,369,039	4,336,818
Context:	**CG (% of total mC)**	1,636,079 (37.4%)	1,635,910 (37.7%)
	CHG (% of total mC)	907,785 (20.8%)	906,410 (20.9%)
	CHH (% of total mC)	1,825,175 (41.8%)	1,794,498 (41.4%)
% genome methylated (vs all cytosines)		10.19%	10.12%
% genome methylated (vs covered cytosines, >1 read)		10.89%	10.76%

REFERENCES

1. Bestor, T. H. (2000) *Hum Mol Genet* **9**(16), 2395-2402

2. Li, E., Bestor, T. H., and Jaenisch, R. (1992) *Cell* **69**(6), 915-926

3. Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K. D., Carrington, J. C., Doerge, R. W., Colot, V., and Martienssen, R. (2004) *Nature* **430**(6998), 471-476

4. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008) *Cell* **133**(3), 523-536

5. Reik, W. (2007) *Nature* **447**(7143), 425-432

6. Rhee, I., Bachman, K. E., Park, B. H., Jair, K. W., Yen, R. W., Schuebel, K. E., Cui, H., Feinberg, A. P., Lengauer, C., Kinzler, K. W., Baylin, S. B., and Vogelstein, B. (2002) *Nature* **416**(6880), 552-556

7. Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., and Ecker, J. R. (2006) *Cell* **126**(6), 1189-1201

8. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007) *Nat Genet* **39**(1), 61-69

9. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008) *Nature* **452**(7184), 215-219

10. Cao, X., Aufsatz, W., Zilberman, D., Mette, M. F., Huang, M. S., Matzke, M., and Jacobsen, S. E. (2003) *Curr Biol* **13**(24), 2212-2217

11. Cao, X., and Jacobsen, S. E. (2002) *Curr Biol* **12**(13), 1138-1144

12. Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., Bestor, T., and Hannon, G. J. (2008) *Mol Cell* **31**(6), 785-799

13. Chan, S. W., Henderson, I. R., and Jacobsen, S. E. (2005) *Nat Rev Genet* **6**(5), 351-360

14. Jackson, J. P., Lindroth, A. M., Cao, X., and Jacobsen, S. E. (2002) *Nature* **416**(6880), 556-560

15. Finnegan, E. J., and Dennis, E. S. (1993) *Nucleic Acids Res* **21**(10), 2383-2388

16. Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddeloh, J. A., Riddle, N. C., Verbsky, M. L., and Richards, E. J. (2003) *Genetics* **163**(3), 1109-1122

17. Saze, H., Mittelsten Scheid, O., and Paszkowski, J. (2003) *Nat Genet* **34**(1), 65-69

18. Gehring, M., Huh, J. H., Hsieh, T. F., Penterman, J., Choi, Y., Harada, J. J., Goldberg, R. B., and Fischer, R. L. (2006) *Cell* **124**(3), 495-506

19. Gong, Z., Morales-Ruiz, T., Ariza, R. R., Roldan-Arjona, T., David, L., and Zhu, J. K. (2002) *Cell* **111**(6), 803-814

20. Morales-Ruiz, T., Ortega-Galisteo, A. P., Ponferrada-Marin, M. I., Martinez-Macias, M. I., Ariza, R. R., and Roldan-Arjona, T. (2006) *Proc Natl Acad Sci U S A* **103**(18), 6853-6858

21. Penterman, J., Zilberman, D., Huh, J. H., Ballinger, T., Henikoff, S., and Fischer, R. L. (2007) *Proc Natl Acad Sci U S A* **104**(16), 6752-6757

22. Barreto, G., Schafer, A., Marhold, J., Stach, D., Swaminathan, S. K., Handa, V., Doderlein, G., Maltry, N., Wu, W., Lyko, F., and Niehrs, C. (2007) *Nature* **445**(7128), 671-675

23. Bruniquel, D., and Schwartz, R. H. (2003) *Nat Immunol* **4**(3), 235-240

24. Metivier, R., Gallais, R., Tiffoche, C., Le Peron, C., Jurkowska, R. Z., Carmouche, R. P., Ibberson, D., Barath, P., Demay, F., Reid, G., Benes, V., Jeltsch, A., Gannon, F., and Salbert, G. (2008) *Nature* **452**(7183), 45-50

25. Rai, K., Huggins, I. J., James, S. R., Karpf, A. R., Jones, D. A., and Cairns, B. R. (2008) *Cell* **135**(7), 1201-1212

26. Kangaspeska, S., Stride, B., Metivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R. P., Benes, V., Gannon, F., and Reid, G. (2008) *Nature* **452**(7183), 112-115

27. Choi, C. S., and Sano, H. (2007) *Mol Genet Genomics* **277**(5), 589-600

28. Dyachenko, O. V., Zakharchenko, N. S., Shevchuk, T. V., Bohnert, H. J., Cushman, J. C., and Buryanov, Y. I. (2006) *Biochemistry (Mosc)* **71**(4), 461-465

29. Hashida, S. N., Uchiyama, T., Martin, C., Kishima, Y., Sano, Y., and Mikami, T. (2006) *Plant Cell* **18**(1), 104-118

30. Kovarik, A., Koukalova, B., Bezdek, M., and Opatrn, Z. (1997) *Theoretical and Applied Genetics* **95**, 301-306

31. Labra, M., Ghiani, A., Citterio, S., Sgorbati, S., Sala, F., Vannini, C., Ruffini-Castiglione, M., and Bracale, M. (2002) *Plant Biology* **4**(6), 694-699

32. Steward, N., Ito, M., Yamaguchi, Y., Koizumi, N., and Sano, H. (2002) *J Biol Chem* **277**(40), 37741-37746

33. Chan, S. W., Henderson, I. R., Zhang, X., Shah, G., Chien, J. S., and Jacobsen, S. E. (2006) *PLoS Genet* **2**(6), e83

34. Simonich, M. T., and Innes, R. W. (1995) *Mol Plant Microbe Interact* **8**(4), 637-640

35. Varet, A., Parker, J., Tornero, P., Nass, N., Nurnberger, T., Dangl, J. L., Scheel, D., and Lee, J. (2002) *Mol Plant Microbe Interact* **15**(6), 608-616

36. Henderson, I. R., and Jacobsen, S. E. (2008) *Genes Dev* **22**(12), 1597-1606

37. Katagiri, F., Thilmony, R., and He, S. Y. (2002) *The Arabidopsis Thaliana-Pseudomonas Syringae Interaction*, American Society of Plant Biologists, Rockville, MD

38. Feinbaum, R. L., and Ausubel, F. M. (1988) *Mol Cell Biol* **8**(5), 1985-1992

39. Hamilton, R. H., Kunsch, U., and Temperli, A. (1972) *Anal Biochem* **49**(1), 48-57

40. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) *Genome Biol* **10**(3), R25

41. Jacobsen, S. E., and Meyerowitz, E. M. (1997) *Science* **277**(5329), 1100-1103

42. Soppe, W. J., Jacobsen, S. E., Alonso-Blanco, C., Jackson, J. P., Kakutani, T., Koornneef, M., and Peeters, A. J. (2000) *Mol Cell* **6**(4), 791-802

43. Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004) *Nucleic Acids Res* **32**(Database issue), D575-577

44. Gehring, M., Bubb, K. L., and Henikoff, S. (2009) *Science* **324**(5933), 1447-1451

45. Hollister, J. D., and Gaut, B. S. (2009) *Genome Res* **19**(8), 1419-1428

46. Hsieh, T. F., Ibarra, C. A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. L., and Zilberman, D. (2009) *Science* **324**(5933), 1451-1454

47. Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., Dangl, J. L., and Weigel, D. (2007) *PLoS Biol* **5**(9), e236

48. Navarro, L., Jay, F., Nomura, K., He, S. Y., and Voinnet, O. (2008) *Science* **321**(5891), 964-967

49. Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., Voinnet, O., and Jones, J. D. (2006) *Science* **312**(5772), 436-439

50. Agorio, A., and Vera, P. (2007) *Plant Cell* **19**(11), 3778-3790

51. Qi, Y., He, X., Wang, X. J., Kohany, O., Jurka, J., and Hannon, G. J. (2006) *Nature* **443**(7114), 1008-1012

52. Zhang, X., Henderson, I. R., Lu, C., Green, P. J., and Jacobsen, S. E. (2007) *Proc Natl Acad Sci U S A* **104**(11), 4536-4541

53. Kovalchuk, I., Kovalchuk, O., Kalck, V., Boyko, V., Filkowski, J., Heinlein, M., and Hohn, B. (2003) *Nature* **423**(6941), 760-762

54. Lucht, J. M., Mauch-Mani, B., Steiner, H. Y., Metraux, J. P., Ryals, J., and Hohn, B. (2002) *Nat Genet* **30**(3), 311-314

55. Molinier, J., Ries, G., Zipfel, C., and Hohn, B. (2006) *Nature* **442**(7106), 1046-1049

56. Boyko, A., Kathiria, P., Zemp, F. J., Yao, Y., Pogribny, I., and Kovalchuk, I. (2007) *Nucleic Acids Res* **35**(5), 1714-1725

The text of Chapter 3, in part, is currently being prepared for submission for publication by Robert H. Dowen, Ryan Lister, Mattia Pelizzola, Joseph R. Nery, Jack E. Dixon, and Joseph R. Ecker. The dissertation author was the primary researcher and the co-authors listed either assisted and/or supervised the research that forms the basis of this chapter.

CHAPTER 4

Concluding remarks

The work presented here has contributed to a greater understanding of plant-pathogen interactions and represents new insights into the molecular mechanisms utilized by bacterial type III effector proteins to promote pathogenesis. Specifically, we have used several approaches to examine a unique family of effector proteins, the AvrPphB-like effectors, some of which utilize their own cysteine protease activity, in combination with the plant host lipidation machinery, to direct their specific sub-cellular localization. Remarkably, all AvrPphB family members localize to the plasma membrane where they likely target components of the plant defense network and contribute to bacterial pathogenesis. Although we have dramatically contributed to understanding the mechanisms that drive the function of AvrPphB family members, a number of interesting questions remain. In particular, identification of the host-specific *in planta* molecular target(s) of each effector protein is of great interest. We attempted to apply bioinformatic strategies to identify putative substrates based upon the amino acid sequence that defines the effector auto-processing site (see Chapter 2, Discussion); however, the limited amount of sequence specificity information (3 residues) prevented us from generating an experimentally testable number of possible substrates. Additional information about the sequence specificity would be needed to apply these types of bioinformatic approaches.

Identification of type III effector targets using genetic approaches has led to discovery of a number of *Avr*-*R* gene, or avirulent, genetic interactions; however, isolation of bona fide biochemical interactions, in particular virulence interactions, has been extremely difficult. It is possible that sophisticated biochemical or proteomicbased approaches could resolve some of the difficulties of identification of these protein-protein interactions. For example, using the AvrPphB-like family of effectors, one could conceivably screen for proteolytic cleavage events within a pool of possible proteinaceous substrates. Purified plant plasma membranes or protein libraries generated from cDNA pools could provide an appropriate collection of putative substrates for large scale biochemical analysis. Regardless of the approach, identification of type III effector targets represents one of the major hurdles in fully understanding the molecular mechanisms that drive bacterial pathogenesis.

The work presented here has also focused on plant-pathogen interactions from the opposing view, that of coordinating the plant's defenses against pathogen attack. We have employed a combined genetic and genomic approach to examine the role of DNA methylation in the regulation of the plant defense response. DNA methylation clearly contributes to the transcriptional regulation, either directly or indirectly, of defense genes, as well as encodes a transgenerational memory of infection. Based on our work and that of others, it seems likely that stress-induced DNA methylation changes serve as a mechanism of both transient and heritable regulation of transcriptional output, indicating that this process must be regulated and well refined. Although we have begun to examine genome wide cytosine methylation and transcriptional alterations upon pathogen infection, additional epigenetic information will be needed to fully deconvolute this dynamic system, including high-throughput sequencing of small RNAs, profiling of nucleosome positioning/content, and mapping of histone modifications. Furthermore, a global understanding of how somatic or germline-specific DNA methylation patterning may ultimately contribute to homologous recombination at specific DNA loci or at defined frequencies will be essential for understanding the true heritable nature of these genomic modifications and their role in the stress response. Finally, a major challenge of deciphering the epigenetic code will be to resolve how stress-induced alterations in DNA methylation patterning, some of which are likely transient by nature, are directed to specific genomic loci, as well as identification of the molecules that are responsible for this targeted response. Although our mechanistic understanding of these epigenetic processes is limited, approaches like whole genome profiling of DNA methylation, small RNAs, mRNAs, and histone modifications provide a strong starting point for further, in depth, dissection of these biological questions.

This dissertation has focused on expanding our understanding of the mechanisms that enhance, or restrict, pathogen growth in the plant, from the strategies employed by bacterial type III effectors to carry out their function to epigenetic regulation of host defense. Additionally, the techniques and scientific approaches described in Chapters 2 and 3 of this dissertation have also been successfully applied to investigate other diverse biological questions within the laboratory. Descriptions of this work, of which I was an essential secondary contributor, comprise the final portion of this dissertation (Appendices A and B).

APPENDIX A

The phosphatase laforin crosses evolutionary boundaries and links carbohydrate metabolism to neuronal disease

ABSTRACT

Lafora disease (LD) is a progressive myoclonic epilepsy resulting in severe neurodegeneration followed by death. A hallmark of LD is the accumulation of insoluble polyglucosans called Lafora bodies (LBs). LD is caused by mutations in the gene encoding the phosphatase laforin, which reportedly exists solely in vertebrates. We utilized a bioinformatics screen to identify laforin orthologues in five protists. These protists evolved from a progenitor red-alga and synthesize an insoluble carbohydrate, whose composition closely resembles LBs. Furthermore, we show that Kingdom Plantae, which lacks laforin, possesses a protein with laforin-like properties called SEX4. Mutations in the *Arabidopsis thaliana SEX4* gene results in a starch excess phenotype, reminiscent of LD. We demonstrate that *Homo sapiens* laforin complements the *sex4* phenotype, and propose that laforin and SEX4 are functional equivalents. Finally, we show that laforins and SEX4 dephosphorylate a complex carbohydrate, and form the only family of phosphatases with this activity. These results provide a molecular explanation for the etiology of LD.

INTRODUCTION

Lafora disease (LD; OMIM #254780) is an autosomal recessive neurodegenerative disorder resulting in severe epilepsy and death (1,2). It is one of five major progressive myoclonus epilepsies. LD presents as a single seizure in the second decade of the patient's life (3); this single event is followed by progressive central nervous system degeneration, culminating in death within ten years of the first seizure (4). A hallmark of LD is the accumulation of polyglucosan inclusion bodies called Lafora bodies (LBs; Lafora (5), Collins *et al*. (6)), located in the cytoplasm of cells in most organs (3,7,8). LB accumulation coincides with increased neuronal non-apoptotic cell death and number of seizures in LD patients. Thus, it is hypothesized that LBs are responsible for these symptoms and ultimately the death of the patient (9).

Recessive mutations in *EPM2B* (epilepsy of progressive myoclonus type 2 gene B)/*NHLRC1* (10), which encodes the E3 ubiquitin ligase malin (10,11), are responsible for ~40% of LD cases (12). Of the LD cases not attributed to mutations in *EPM2B*, ~48% result from recessive mutations in *EPM2A* (epilepsy of progressive myoclonus type 2 gene A; Minassian *et al*. (13), Serratosa *et al*. (14), Ianzano *et al*. (12)). *EPM2A* encodes laforin, which contains a carbohydrate binding module family 20 (CBM20; Wang *et al*. (15)) domain followed by the canonical dual specificity phosphatase (DSP) active site motif, HCXXGXXRS/T (CX_5R) (Figure A.1.A; Denu *et al*. (16), Yuvaniyama *et al*. (17), Minassian *et al*. (13)). The CBM of laforin binds complex carbohydrates *in vivo* and *in vitro* (15), and the DSP motif can hydrolyze phosphotyrosine and phosphoserine/threonine substrates *in vitro* (15,18). However, no group has detected endogenous laforin localization in tissue culture cells or in wildtype tissues, likely due to low levels of accumulation (Chan *et al*. (19), our unpublished results).

Of the 128 human phosphatases (20,21), only laforin possesses a CBM. CBM domains are predominantly found in glucosylhydrolases and glucotransferases of bacterial, fungal, or plant origin (22-24). The vast majority of enzymes containing a CBM use the domain to bind a specific type of carbohydrate and then enzymatically act on the sugar (23). Accordingly, we recently showed that laforin liberates phosphate from the complex carbohydrate amylopectin, while other phosphatases lack this activity (25).

Ganesh and coworkers disrupted the *EPM2A* locus in a mouse model (26). While this model faithfully recapitulated the disease, it yielded no molecular explanation for LD. Similarly, Chan and colleagues generated a transgenic mouse overexpressing inactivated laforin and this mouse model also developed LD (19). Despite the availability of these two LD mouse models, the molecular etiology of LD remained unexplained. These limitations demonstrate the need to develop alternative model systems to elucidate the biology of LD. Although a molecular mechanism to explain LD has remained elusive, data cumulatively place laforin in the context of being intimately, if not directly, involved in regulating glycogen metabolism. We, therefore, focused on this indisputable aspect of LD for clues to its molecular etiology.

Insoluble Glycogen, Starch, and Floridean Starch

Glycogen is produced in the cytoplasm of the majority of archaebacterial, bacterial, fungal and animal species. It is a water-soluble polymer composed of α-1,4-glycosidic linkages between glucose residues, with branches occurring in a continuous pattern every 12-14 residues via α-1,6-glycosidic linkages. Almost every report on LD refers to LBs as “insoluble glycogen”. However, definitive biochemical studies on LBs found that the arrangement and pattern of branching in LBs most closely resemble amylopectin (9,27,28).

Amylopectin, like glycogen, is composed of α-1,4-glycosidic linkages with α-1,6-glycosidic branches, but with branches arranged in a discontinuous pattern every 12-20 residues. This discontinuous and decreased amount of branching renders amylopectin insoluble. Amylopectin is one of the two components of starch, which is produced in the plastid of green plants (Viridaeplantae). Starch is an insoluble, semi-crystalline heterogeneous mixture of 10-25% amylose and 75-90% amylopectin. Plants synthesize starch in chloroplasts during daylight as a transient carbon store that is utilized during the dark cycle to generate a usable, reduced form of carbon in the absence of photosynthesis.

Floridean starch is another insoluble carbohydrate that has similar biochemical properties to amylopectin (29,30). Floridean starch is synthesized in the cytoplasm of a variety of protists (i.e. unicellular eukaryotes) and is utilized as an energy source during specific stages of their life cycle. Floridean starch, like LBs and amylopectin, is made of glucose polymers with branches every 12-20+ residues in a discontinuous pattern (30). Thus, floridean starch, amylopectin, and LBs have been described as possessing similar characteristics.

EXPERIMENTAL PROCEDURES

Cloning, Vectors, and Purification of Recombinant Proteins

The complete open reading frame of Cm-laforin was cloned from cDNA provided by T. Kuroiwa (Rikkyo University, Tokyo, Japan) and *SEX4* from SSP Consortium clone U14967 (31). Cm-laforin and *SEX4* were cloned into pET21a (Stratagene) according to standard protocols. A second pET21a *SEX4* construct was generated because the full-length protein is largely insoluble. We truncated the first 52 amino acids of SEX4 to generate pET21a ∆52-SEX4. pET21a VHR (32) and pET21a Hs-laforin (15) have been described previously. Hs-laforin, *SEX4*, and *sex4-C198S* were cloned in frame of a triple HA tag into pCHF1 (33), which is a modified version of pPZP221 (34). pCHF1 contains the 35S cauliflower mosaic virus promoter, the Rubisco terminator from pea, and confers gentamicin resistance for selection in plants. Because Kerk *et al*. (35) and Niittyla *et al*. (36) demonstrated that the cTP of SEX4 targets SEX4 to the chloroplast, we fused the cTP of SEX4 (nucleotides 1-213) in frame with Hs-laforin and the triple HA tag in pCHF to generate pCHF cTP-Hs-laforin. All point mutations were generated with the QuickChange Site-Directed Mutagenesis kit (Stratagene) according to the manufacturer's instructions. All constructs were verified by DNA sequencing. Recombinant proteins were expressed with a carboxy-terminal six-histidine tag in *Escherichia coli* BL21 (DE3) CodonPlus cells (Stratagene). Fusion proteins were expressed and purified from soluble bacterial extracts using Ni^{2+}-agarose affinity chromatography as described previously (11).

Phosphatase Assays

Hydrolysis of *p*-NPP was performed in 50 µl reactions containing 1x phosphate buffer (0.1 M sodium acetate, 0.05 M bis-Tris, 0.05 M Tris-HCl, and 2 mM

DTT at the appropriate pH), 50 mM *p*NPP, and 100-500 ng of enzyme at 37˚C for 1-5 min. The reaction was terminated by the addition of 200 μl of 0.25 M NaOH, and absorbance was measured at 410 nm. We tested the specific activity of each enzyme at pH 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, and 8.0. The optimal pH for each enzyme was as follows: Hs-laforin, pH 5.0; Cm-laforin, pH 5.5; SEX4, pH 6.0; and VHR, pH 6.0. Malachite green assays were performed as described previously (37) with the following modifications: 1x phosphate buffer, 100-500 ng of enzyme, and ~45 μg of amylopectin in a final volume of 20 μl. The reaction was stopped by the addition of 20 μl of 0.1 M *N*-ethylmaleimide and 80 µl of malachite green reagent. Absorbance was measured at 620 nm. We tested the specific activity of each enzyme at the same pH units as above. The optimal pH for each enzyme was: Hs-laforin, pH 7.0; Cm-laforin, pH 6.0; and SEX4, pH 8.0.

Phylogenetic Analyses and Sequence Alignment

The sequences of laforin and SEX4 orthologues were obtained by performing tBLASTn searches using the GenBank dbEST database or BLASTp and PSI-BLAST (38) searches using GenBank eukaryote genome and nr databases, the *C. merolae* genome project, Department of Energy Joint Genome Institute Resource, The Institute for Genomic Research, ToxoDB, GeneDB, Genoscope, and *Tetrahymena* Genome Database. Accession numbers are listed in Table A.3 and Table A.4. The web address for each database is listed in Table A.1. A list of each genome that we investigated and a reason why an organism's genome lacks laforin is listed in Table A.6. Amino acid sequences of laforin orthologues were aligned by ClustalW (39) and refined manually using MacVector. Small subunit ribosomal RNA sequences were obtained by performing BLASTn using GenBank from all organisms and nr databases, and accession numbers are listed in Table A.5. The phylogenetic tree was generated from a ClustalW (39) multiple sequence alignment using PROTDIST and FITCH from the PHYLIP 3.65 software package and was displayed using HYPERTREE 1.0.0 (Pfizer; Bingham *et al*. (40)).

Plant Material

Homozygous *sex4-3* plants (T-DNA insertion line SALK_102567; Alonso *et al*. (41)) were isolated by PCR. Stable transgenic plant lines were generated by *Agrobacterium*-mediated floral dipping (42), and seeds were sterilized, plated on standard growth medium (43), and selected for using 100 µg/ml gentamycin per standard protocols (42,43). Plants were grown in Promix-HP soil at 22˚C with supplemental lighting conditions of 16-h days. To stain starch in leaves, leaves were decolorized in 80% (vol/vol) ethanol, stained with an iodine solution, and destained in water. Starch content was quantified as previously described (44). mRNA was obtained using a RNeasy Plant Mini kit (QIAGEN), and first-strand synthesis was performed using SuperScript III First-Strand Synthesis SuperMix (Invitrogen) according to the manufacturer's recommendations. Four primer sets were used to test for the presence of transcripts in wild-type (Columbia-0) and *sex4-3* plants. Three primer sets to the *SEX4* transcript and a positive control to *UBC5*, the *UBC5* primer set, was included in each PCR tube. Plant whole leaf lysate was obtained as described previously (45).

Antibodies and Western Analysis

The α-Hs-laforin and α-Cm-laforin antibodies were generated by immunizing rabbits with recombinant Hs-laforin or Cm-laforin, and antibodies were affinity purified from the serum with a HiTrap NHS-activated HP affinity column (GE Healthcare) of Hs-laforin or Cm-laforin protein, respectively. Recombinant Hs-laforin and Cm-laforin were detected with their respective primary antibodies followed by goat α-rabbit HRP (GE Healthcare). Recombinant VHR and SEX4 were detected with α-HIS HRP (Santa Cruz Biotechnology, Inc.). Protein expression of *A. thaliana* transgenes was monitored by Western analysis using rat α-HA (clone 3F10, Roche) and goat α-rat HRP (Chemicon).

C. merolae Cell Culture, Immunofluorescence, and Immunogold Electron Microscopy

C. merolae 10D-14 (46) was provided by T. Kuroiwa and grown asynchronously at pH 2.5 in 2x Allens's medium at 42˚C under continuous illumination as described previously (47). For immunofluorescence, cells were fixed, washed, and blocked as previously described (48). Cells were then probed with 1:100 preimmune serum or 1:1,000 α-Cm-laforin antibody followed by 1:1,000 AlexaFluor488 goat α-rabbit antibody (Invitrogen). Chloroplasts were visualized by their autofluorescence. Immunofluorescence was performed using a light microscope (DMR, Leica) with a PL APO 63x 1.32 NA oil objective (Leica) at room temperature, and images were captured with a CCD camera (C4742-95, Hamamatsu) using OpenLab 4.0.1 software (Improvision). For immunogold EM, cells were fixed, washed, sectioned, and blocked as previously described (49). Sections were immunostained with 1:50 preimmune serum or 1:250 α-Cm-laforin antibody and with 10 nm of gold particle-conjugated goat α-rabbit antibody. Grids were viewed using a transmission electron microscope (1200EX II, JEOL), and images were collected using digital camera (ORIUS SC600, Gatan) and Digital Micrograph software (Gatan).

Photoshop (Adobe) and Illustrator (Adobe) were used to generate figures of all images.

RESULTS

Discovery of Laforin Orthologues

One protist that accumulates floridean starch (also called amylopectin granules) in its cytoplasm is *Toxoplasma gondii* (30,50-52). *T. gondii* is an obligate intracellular parasite that can infect nearly any nucleated cell from a warm-blooded animal. Like most members of Apicomplexa, *T. gondii* has a complex life cycle; in its intermediate hosts it exists as a rapidly dividing tachyzoite or an encysted bradyzoite, depending on the host immune response. The bradyzoite forms floridean starch in its cytoplasm that is used as an energy source (50). Recent reports characterized the biochemical composition of *T. gondii* floridean starch (30,51). We noted that the biochemical composition of *T. gondii* floridean starch was remarkably similar to that of LBs described nearly 40 years ago (9,27-28). Although *EPM2A* has been reported to be present only in vertebrates (53,54), the similarity between *T. gondii* floridean starch and LBs led us to explore the partially completed *T. gondii* genome for a laforin orthologue(s).

The sequence of the *T. gondii* genome, like the genome of many protists, was not accessible via GenBank when we initiated this study. Therefore, we searched the *T. gondii* database (ToxoDB; Kissinger *et al*. (55)) for a laforin orthologue. We used the criteria that a laforin orthologue must contain both an amino-terminal CBM and a carboxy-terminal DSP domain (Figure A.1.A). DSP domains are readily recognized by the protein families database (pfam; Bateman *et al*. (56)) and NCBI's conserved domain database (CDD; Marchler-Bauer *et al*. (57)). However, CBMs are very degenerate at the primary amino acid level and neither database consistently recognizes any of the 45 CBM families. Because CDD and pfam do not reliably recognize CBMs, we devised a multi-tiered search strategy to identify laforin orthologues (Figure A.1.B). First, we performed BLASTp (38) searches using the DSP motif HCXXGXXR as an index sequence and identified 20 *T. gondii* proteins containing this motif. Because laforin contains an amino-terminal CBM and CBMs contain 80-100 amino acids, we eliminated two of these proteins because their HCXXGXXR motif was within the first 80 amino acids. We next performed a secondary BLAST using the NCBI nonredundant (nr) database with each of the remaining 18 proteins minus their DSP domain. If the protein contained a CBM, then the BLAST identified other CBM containing proteins. Using this strategy, we identified one protein, which we refer to as *T. gondii* laforin (Tg-laforin), that met the aforementioned criteria. Tg-laforin and *Homo sapiens* laforin (Hs-laforin) are 37% identical (Figure A.1.C). Importantly, Tg-laforin contains all of the residues important for carbohydrate binding as well as the signature residues of a DSP (Figure A.1.A).

With the discovery of a putative laforin orthologue in *T. gondii*, we extended our search methods to identify additional orthologues using a variety of genome databases (Table A.1). Utilizing this strategy, we identified laforin orthologues in the four classes of vertebrates with sequenced genomes (mammals, aves, amphibians, and osteichthyes; Figure A.1.A and A.1.D). In addition, we identified putative laforin orthologues in four additional protests: *Eimeria tenella*, *Tetrahymena thermophila*, *Paramecium tetraurelia*, and *Cyanidioschyzon merolae* (Figure A.1.A and A.1.C). While Hs-laforin contains 331 amino acids, the putative protist orthologues varied in predicted size from 323-727 amino acids. However, each putative orthologue contained the “signature” amino acids of a CBM20 and DSP; that is, four invariant aromatic amino acids (Hs-laforin F5, W32, W60, and W99) as well as $DX_{30}CX_2GX_2R$, respectively (Figure A.1.A). Despite exhaustive efforts (we searched ~170 eukaryotic genomes and ~670 bacterial and archaeal genomes), we did not identify any other putative laforin orthologues. Thus, laforin is absent in all “traditional” non-vertebrate model organisms (e.g. yeast, fly, and worms). Laforin orthologues exist in all classes of vertebrates where sequence data is available, and in the five protists that we identified (Figure A.1.A, A.1.C, and A.1.D).

Biochemical Properties and Subcellular Localization of Laforin Orthologues

C. merolae laforin (Cm-laforin) shares the least identity with Hs-laforin (Figure A.1.C). As such, we reasoned that if it exhibited similar *in vitro* characteristics as Hs-laforin, then the other putative orthologues were likely to as well. To test whether the identified protist proteins had similar biochemical characteristics as Hs-laforin and were thus laforin orthologues, we cloned the putative orthologue from *C. merolae* (Cm-laforin) and purified recombinant protein from bacteria (Figure A.2.A). Characteristic of all DSPs, Hs-laforin exhibits phosphatase activity against the artificial substrate para-nitrophenylphosphate (*p*-NPP; Figure A.3.A; Ganesh *et al.* (18)). Cm-laforin also utilized *p*-NPP as an artificial substrate with similar kinetics as Hs-laforin (Table A.2) and displayed a similar specific activity (Figure A.3.A). In addition to activity against *p*-NPP, we recently showed that recombinant Hs-laforin releases phosphate from amylopectin (25), and that this activity is unique to laforin (25). Additionally, we fused the CBM of laforin to DSP VH1 related (VHR) and demonstrated that although this fusion protein was an active phosphatase, it did not liberate phosphate from amylopectin (25). Figure A.3.B shows that like Hs-laforin, Cm-laforin displays a robust ability to release phosphate from amylopectin, while VHR does not hydrolyze phosphate from amylopectin. As predicted, the catalytically inactive Cm-laforin-C/S mutant displayed no activity against either substrate (Figure

A.3.A and A.3.B). Additionally, Tg-laforin also displayed activity against both *p*-NPP and amylopectin (unpublished data).

Hs-laforin is the only phosphatase in the human genome that contains a CBM and, as such, is predicted to be the only phosphatase that binds carbohydrates. Cm-laforin and Tg-laforin bound amylopectin to the same extent as Hs-laforin (Figure A.3.C and not depicted). Conversely, VHR did not bind amylopectin (Figure A.3.C). Wang and coworkers previously demonstrated that conserved tryptophan and lysine residues (Figure A.1.A) that participate in binding to the sugar are necessary for Hs-laforin to bind amylopectin (Figure A.3.C; Wang *et al*. (15)). Accordingly, mutation of these corresponding residues in Cm-laforin also abolished its ability to bind amylopectin (Figure A.3.C). These mutations also significantly reduced the ability of Cm-laforin to release phosphate from amylopectin (Figure A.4.A), while only minimally affecting its *p*-NPP activity (Figure A.4.B). These data suggest that Cm-laforin must be “positioned” correctly via the CBM in order for the DSP domain to dephosphorylate amylopectin or that the CBM binding to the carbohydrate is needed to “activate” the DSP.

While laforin from all three species binds α-glucans *in vitro*, this result may not reflect the biological localization of laforin. Moreover, the localization of Hs-laforin has never been determined in wild-type cells or tissues (Chan *et al*. (19), our unpublished data). Because we identified multiple new systems to study laforin, we investigated laforin’s localization in *C. merolae*. A *C. merolae* cell contains a chloroplast, mitochondrion, and nucleus, and when grown in continuous light accumulates vast storages of floridean starch (Figure A.3.D, schematic; Viola *et al*. (58)). We fixed *C. merolae* cells and probed them with an affinity-purified α-Cm-laforin antibody. We found that endogenous Cm-laforin localized in punctate accumulations throughout the cytoplasm of cells (Figure A.3.D). To further define the localization of Cm-laforin, we performed immunogold electron microscopy staining. Ultra-thin sections of *C. merolae* cells were probed with the affinity-purified α-Cm-laforin antibody and a 10-nm gold particle-conjugated goat α-rabbit secondary antibody. Positive staining was observed surrounding the floridean starch granules (Figure A.3.E, arrowheads). No Cm-laforin was observed within the granules because prior to sectioning no protein would have access to this region. In addition, no background staining was observed with the secondary antibody alone (Figure A.5). Thus, as we hypothesized, endogenous laforin binds the outer region of insoluble carbohydrates.

Cm-laforin and Tg-laforin possess the same three *in vitro* properties as Hs-laforin: both use *p*-NPP as an artificial substrate, bind amylopectin, and release phosphate from amylopectin. Accordingly, the laforin orthologues in vertebrates and the five mentioned protists contain the critical signature primary amino acid structure of a CBM20 and DSP. Thus, our integrated bioinformatics searches for combined CBM and DSP domains correctly predicted the biochemical properties of Cm-laforin. Because the laforin orthologues are the only proteins in any of these genomes that contain a CBM and DSP, we hypothesized that these organisms may have acquired laforin from a common ancestor.

Evolutionary Lineage of Laforin

The key to the evolutionary lineage of laforin lies in the origin of the aforementioned five protists. The chromalveolate hypothesis postulates (59) that a distinct sequence of events led to the evolution of kingdom Plantae and to subsequent progeny, including the five aforementioned protists. As illustrated in Figure A.6.A, a mitochondriate protist engulfed a cyanobacterium (60,61) and eventually gave rise to kingdom Plantae (62). Once Plantae was established, a second endosymbiosis involving red algae (63) gave rise to the chromalveolates (Figure A.6.B; Cavalier-Smith *et al*. (59)). These engulfments were accompanied with the co-evolution of “various manifestations of mitochondria” (64) and various forms of carbohydrate storage (58). These combined evolutionary events resulted in organisms possessing a mitosome, a hydrogenosome, or a true mitochondrion; and some organisms evolved floridean starch as their storage carbohydrate. We hypothesized that interspersed with these evolutionary events, organisms maintained, gained, or lost laforin.

To trace the lineage of laforin, we generated a phylogeny derived from the small-subunit ribosomal RNA gene of organisms belonging to diverse evolutionary niches, and highlighted the organisms whose genome contains laforin (Figure A.6.C). This phylogenetic analysis revealed that each of the five protists containing a laforin orthologue is of red-algal descent. However, the genome of some organisms of red-algal descent lack laforin (Figure A.6.C). To determine why some organisms of red-algal descent lack laforin, we analyzed the biology of each of the organisms in Figure A.6.C. We discovered that each organism of red-algal descent that contained laforin also contained a true mitochondrion and produced floridean starch. Conversely, organisms of red-algal descent lacking laforin either lacked a true mitochondrion or did not produce floridean starch. For example, *Plasmodium falciparum* is of red-algal descent and possesses mitochondria; however, it does not produce floridean starch and thus lacks laforin (Figure A.6.C). Similarly, *Cryptosporidium parvum* is of red-algal descent and produces floridean starch, but it has mitosomes instead of mitochondria and thus lacks laforin (Figure A.6.C). Conversely, *C. merolae* is a red alga that produces floridean starch and contains a single mitochondrion, and, in agreement with our established criteria, contains laforin. Additionally, glaucophytes and green algae/land plants lack a laforin orthologue because they evolved as contemporaries of red algae and not as descendents (Figure A.6.A). Thus, our analyses generated three criteria to predict if a protist's genome possesses laforin: the organism must 1) be of red-algal descent, 2) possess a true mitochondrion, and 3) produce floridean starch. To determine if our criteria correctly predicted the presence of laforin, we investigated the biology of each organism from the 168 eukaryotic genomes we probed. We found that in each case our criteria correctly predicted the presence or absence of laforin (Table A.3).

A Laforin-like Protein in Plants

Protists such as *T. gondii* use insoluble floridean starch as an energy source when transitioning from inactive/hibernating life cycle stages to active/replicative stages (50). Likewise, *C. merolae*, a red alga that contains laforin, synthesizes insoluble floridean starch during the day and uses it as a source of energy at night. Plants have a similar diurnal cycle, producing insoluble carbohydrate in the form of starch during the day and catabolizing it during the night. Because Hs-laforin has been implicated in carbohydrate metabolism and we show that Cm-laforin binds and releases phosphate from amylopectin, we hypothesized that laforin plays a vital role in insoluble carbohydrate metabolism. Thus, we predicted that plants would also have a "laforin-like" activity; however, we were unable to identify a laforin orthologue in plants. Recently, several starch excess mutants which accumulate starch have been described in plants (65-67); one of these is attributed to mutations in the *starch excess 4* (*SEX4*) gene (At3g52180; Niittyla *et al.* (36)). Kerk and co-workers (35) and Niittyla and colleagues (36) demonstrated that the *Arabidopsis thaliana SEX4* gene (previously identified as a phosphatase and called *AtPTPKIS1*; Fordham-Skelton *et al.* (68)) encodes a protein containing a chloroplast targeting peptide (cTP) and DSP domain at its amino-terminus followed by a CBM-like domain at its carboxy-terminus (Figure A.7.A), suggesting that SEX4 might be a “laforin-like” phosphatase (36).

The DSP of SEX4 shares the key $DX_{30}CX_2GX_2R$ catalytic residues with the DSP of Hs-laforin and is 24% identical to Hs-laforin (Figure A.7.B and Figure A.7.D). Conversely, the CBM of SEX4 lacks many of the invariant CBM20 residues (Figure A.7.C vs. Figure A.1.A) and shares only 18% identity with the CBM of Hs-laforin (Figure A.7.D). Instead, a sequence search using the CBM of SEX4 shows that it is most similar to another class of CBM, the AMP-activated protein kinase β-glycogen-binding domain (AMPKβ-GBD) family (69), and not to CBM20 (Figure A.7.C and Figure A.7.D). Despite their structural differences, both CBM20 and the AMPKβ-GBD domains interact with individual glycan chains of carbohydrates (23,69), suggesting that SEX4 could bind starch via its AMPKβ-GBD. Thus, SEX4 contains similar domains to laforin, but the domains are arranged in the opposite orientation (Figure A.1.A vs. Figure A.7.A). We next performed BLASTp searches of various databases (Table A.1) and found that SEX4 is conserved in all land plants and in *Chlamydomonas reinhardtii*, a single-cell green alga closely related to the progenitor of land plants (Figure A.7.C and Figure A.7.D). Thus, SEX4 likely evolved before or during the establishment of green algae and performs a kingdom-wide function in Plantae.

To ascertain whether SEX4 possesses biochemical properties similar to laforin, we cloned *A. thaliana SEX4* and assayed purified recombinant SEX4 protein (At-SEX4, Figure A.2.B). Because the cTP of SEX4 is highly hydrophobic and renders the protein insoluble, we deleted the first 52 amino acids and used purified recombinant HIS-tagged Δ52-SEX4 for our assays (Figure A.2.B). We found that Δ52-SEX4 has a similar specific activity and possesses similar kinetics as Hs-laforin against *p*-NPP (Figure A.7.E and Table A.2) and efficiently liberates phosphate from amylopectin (Figure A.7.F). Conversely, mutation of the active site cysteine to serine abolished these activities (Figure A.7.E and Figure A.7.F). Additionally, wild-type (Δ52-SEX4) and catalytically inactive SEX4 (Δ52-SEX4-C198S) bind amylopectin similar to Hs-laforin (Figure A.7.G). Importantly, mutations in key conserved AMPKβ-GBD residues that form essential hydrogen bonds with the sugar (69,70) abolish this interaction (Figure A.7.G) while minimally affecting the phosphatase activity of SEX4 (Figure A.8.A). These mutations significantly reduced the ability of SEX4 to release phosphate from amylopectin (Figure A.8.B). Thus, like Cm-laforin, SEX4 must also be "positioned" correctly via the CBM in order for the DSP domain to dephosphorylate amylopectin.

Clearly, SEX4 and the laforins contain both a functional CBM and a DSP domain highly specific for dephosphorylating amylopectin. Additionally, we speculate that they are involved in insoluble carbohydrate metabolism. Because carbohydrate metabolism evolved independently in the kingdom Plantae and kingdom Animalia, the use of similar protein modules to regulate a key feature of carbohydrate metabolism in these lineages is a striking example of convergent evolution and strongly suggests that laforin and SEX4 might be functional equivalents.

SEX4 is a Functional Equivalent of Laforin

The *SEX4* locus was recently mapped in *A. thaliana* to At3g52180, and multiple mutations in this gene display a starch excess phenotype (36,71). One characterized mutation is the *sex4-3* allele that contains an *Agrobacterium* transferred

DNA (T-DNA) insertion in the sixth exon (36) and leads to disruption of *SEX4* expression (Figure A.9.A). Because laforin and SEX4 are the only reported proteins in any kingdom that contain both functional CBM and DSP domains and because mutations in the gene expressing either protein results in aberrant carbohydrate accumulation, we postulated that SEX4 and laforin could be functional equivalents.

To test this hypothesis, we transformed *sex4-3* plants to generate stable lines expressing SEX4, sex4-C/S, Hs-laforin, and Hs-laforin fused behind a cTP (cTP-Hs-laforin) to target Hs-laforin to the chloroplast (like SEX4), and monitored protein expression of the transgenes (Figure A.9.B). We then assayed starch accumulation in wild-type, *sex4-3*, and *sex4-3* transgenic plants. As per our prediction, transformants expressing SEX4 and cTP-Hs-laforin no longer displayed the starch excess phenotype, whereas the catalytically inactive sex4-C/S mutant and Hs-laforin transformants still accumulated excess starch (Figure A.9.C and Figure A.9.D; Figure A.10). Thus, the cTP-Hs-laforin fusion rescued the starch excess phenotype both qualitatively and quantitatively. Conversely, Hs-laforin lacking the cTP did not rescue any portion of the phenotype. Therefore, Hs-laforin is a functional equivalent of SEX4 that must be targeted to the chloroplast, just like SEX4, in order to perform the equivalent function.

DISCUSSION

Our studies probe the molecular mechanism of LD. We identified laforin orthologues in specific protists and further showed that Hs-laforin and plant SEX4 are functional equivalents. Our results provide compelling evidence that a laforin-like activity is required to regulate the metabolism of amylopectin-like material across multiple kingdoms. Additionally, we demonstrate the nature of this activity; that is, the dephosphorylation of the carbohydrate itself, thus providing a molecular explanation for LD. Although there are examples of DSPs that dephosphorylate nonproteinacious substrates (such as phosphate and tensin homologue, the myotubularin family, and Sac domain phosphatases that dephosphorylate the inositol head group of phospholipids; 72-77), ours is the first example of a family of phosphatases that dephosphorylate complex carbohydrates.

We demonstrate that laforin is not merely restricted to the genomes of vertebrates but is well conserved in the protists *T. gondii*, *E. tenella*, *T. thermophila*, *P. tetraurelia*, and *C. merolae*. Laforin's evolutionary lineage shows that it originated in a primitive red alga during early eukaryotic evolution. Despite its early origin, laforin was only maintained by organisms that synthesize floridean starch (such as the aforementioned five protists) and organisms that inhibit the production of insoluble carbohydrates (i.e., all vertebrates). Organisms that no longer performed either of these processes lost laforin. Conversely to laforin, we show that although SEX4 contains similar domains as laforin, its lineage differs in that SEX4 is conserved in all land plants as well as in *C. reinhaardtii*, a close descendent of primitive green algae. Despite their different lineages, Hs-laforin performs the same function as the plant protein SEX4; thus, we propose that laforin and SEX4 are functional equivalents.

It must be noted that although laforin and SEX4 share a common function and similar domains, they are not orthologous proteins. They are not orthologues because (1) although they share similar CBMs, the CBMs belong to different classes and differ considerable with respect to the primary amino acids that are important for binding carbohydrates, and (2) the DSP and CBM of laforin and SEX4 are arranged in opposite orientations. Thus, it is likely that red and green algae independently evolved a phosphatase via convergent evolution that utilizes a similar mechanism to regulate insoluble carbohydrate metabolism.

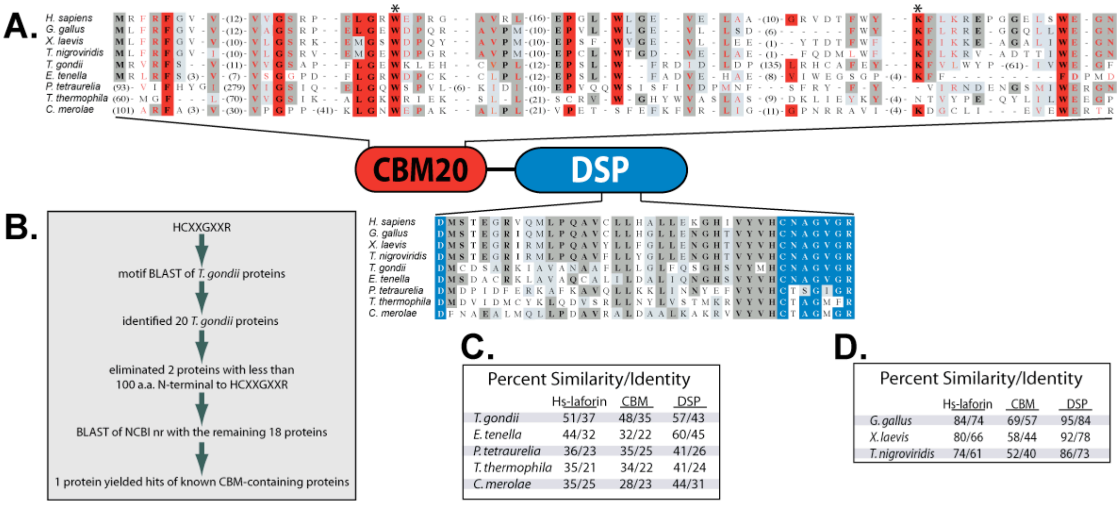
Despite the independent means by which laforin and SEX4 evolved, they both dephosphorylate the same carbohydrate substrate and constitute a unique family of phosphatases. In addition, we demonstrate that endogenous Cm-laforin localizes around the floridean starch granules. Although most studies thus far suggest a carbohydrate substrate for laforin and SEX4, it is possible that they bind their respective amylopectin-like material (insoluble glycogen and starch, respectively) and dephosphorylate a proteinacious substrate. This proteinacious substrate would likely be involved in regulating carbohydrate metabolism, a process controlled by multiple levels of phosphorylation (78). Although the overall carbohydrate machinery differs substantially between mammals and plants, both systems contain common phosphoproteins that share conserved functions (30,79-80). These proteins would be likely substrate candidates for laforin and SEX4. To address this hypothesis, we tested the majority of the mammalian candidates, but none of them served as a substrate for laforin (Worby *et al*. (25), our unpublished data).

It is interesting that laforin and SEX4 are functional equivalents that dephosphorylate a complex carbohydrate and that the mutation of either gene results in the accumulation of insoluble carbohydrates in vertebrates and plants, respectively.

Our understanding of the metabolism of insoluble carbohydrates in vertebrate systems is still in its infancy. In contrast, the plant community has made significant progress in understanding the metabolism of starch (66,67). In plants, it is clear that the phosphorylation of glucose residues within starch is required for its proper accumulation and degradation (65-67). In *A. thaliana*, glucan water dikinase (81) and phosphoglucan water dikinase (44,82) phosphorylate glucose monomers within amylopectin at the C6 and C3 position (83), respectively. As with *SEX4*, mutations in the genes encoding glucan water dikinase and phosphoglucan water dikinase also yield a starch excess phenotype (44,82,84). Phosphorylation is necessary for both starch accumulation and degradation; however, the timing of these phosphorylation and dephosphorylation events is unknown (66,67). Intriguingly, although glycogen, the soluble storage carbohydrate in vertebrates, contains little to no phosphate, detrimental insoluble carbohydrates like LBs are highly phosphorylated, just like amylopectin in plant starch (28,85). Therefore, it appears logical that laforin and SEX4 evolved to perform the critical role of dephosphorylating insoluble carbohydrates to allow their proper degradation.

This basic function of insoluble carbohydrate metabolism provides an intriguing explanation for both the existence of a laforin-like activity in protists and plants and the role of laforin in preventing LD. In protists and plants, carbohydrate dephosphorylation would be necessary for utilization of insoluble carbohydrates as an energy source. When this activity is absent, these organisms accumulate unusable starch as in the *sex4* mutants. In vertebrates, laforin would dephosphorylate nascent insoluble carbohydrates to inhibit the formation of detrimental LBs. In the absence of laforin, these nascent molecules increase in size and number and eventually cause LD.

Our work clearly demonstrates that a laforin-like activity is necessary for the proper metabolism of insoluble carbohydrates. This activity is required throughout multiple kingdoms and regulates an overlooked aspect of carbohydrate metabolism. It is striking that protists and plants have provided new insights into a human neurodegenerative disease involving aberrant carbohydrate metabolism that was described almost 100 years ago by Lafora and Gluck (1,5).


A.
H. sapiens
G. gallus
X. laevis
T. nigroviridis
T. gondii
E. tenella
P. tetraurelia
T. thermophila
C. merolae
CBM20
DSP
B.
HCXXGXXR
motif BLAST of T. gondii proteins
identified 20 T. gondii proteins
eliminated 2 proteins with less than 100 a.a. N-terminal to HCXXGXXR
BLAST of NCBI nr with the remaining 18 proteins
1 protein yielded hits of known CBM-containing proteins
C.
D.

C. Percent Similarity/Identity

	Hs-laforin	CBM	DSP
T. gondii	51/37	48/35	57/43
E. tenella	44/32	32/22	60/45
P. tetraurelia	36/23	35/25	41/26
T. thermophila	35/21	34/22	41/24
C. merolae	35/25	28/23	44/31

D. Percent Similarity/Identity

	Hs-laforin	CBM	DSP
G. gallus	84/74	69/57	95/84
X. laevis	80/66	58/44	92/78
T. nigroviridis	74/61	52/40	86/73

Figure A.1 Laforin orthologues.

(A) An alignment of the vertebrate and protist laforin orthologues. Residues highlighted in red are highly conserved CBM20 residues as defined by the CBM20 family (Wang *et al*. (15)), residues boxed in red are invariant CBM20 residues (Wang *et al*. (15)), and residues boxed in blue are part of the DSP catalytic site. Residues boxed in dark grey are identical, and those boxed in light grey are conserved substitutions. Asterisks mark residues necessary for binding to carbohydrates (Wang *et al*. (15)). Accession numbers are listed in Table A.4.
(B) Strategy to identify laforin orthologues in *T. gondii*.
(C) Percent similarity and identity of full-length Hs-laforin and both domains compared to those of protists.
(D) Percent similarity and identity of full-length Hs-laforin and both domains compared to those of other vertebrates.

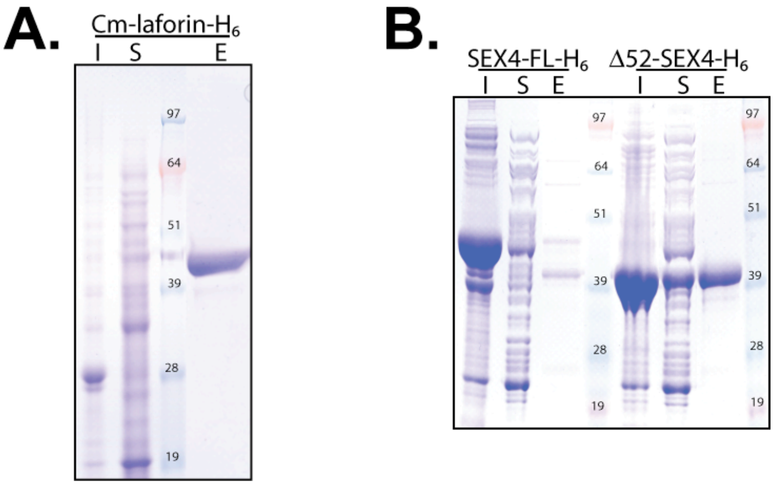


Figure A.2 Purification of recombinant Cm-laforin and recombinant SEX4.

(A) Cm-laforin-HIS$_6$ was purified from soluble *E. coli* lysate via Ni^{2+}-agarose affinity chromatography.
(B) Full length SEX4-HIS$_6$ (SEX4-FL-H$_6$) and an *N*-terminal 52 amino acid deletion of SEX4-HIS$_6$ (Δ52-SEX4-H$_6$) were purified from soluble *E. coli* lysate via Ni^{2+}-agarose affinity chromatography. I, insoluble fraction; S, soluble fraction; and E, eluate fraction.

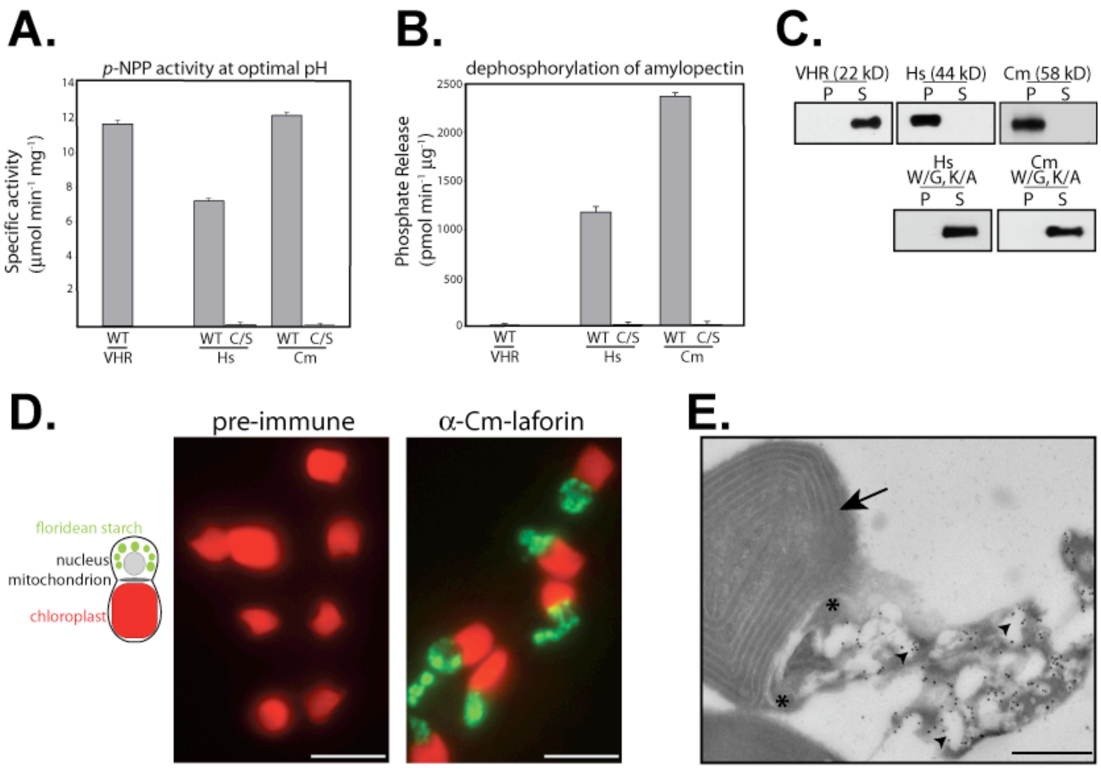


Figure A.3 Biochemical characterization of laforin orthologues.

(A) Specific activity of VHR, Hs-laforin, Hs-laforin-C/S, Cm-laforin, and Cm-laforin-C/S against *p*-NPP at their respective optimal pH. Error bars indicate mean ± SD.
(B) Phosphate release measured by malachite green assays using VHR, Hs-laforin, Hs-laforin-C/S, Cm-laforin, and Cm-laforin-C/S against amylopectin at their respective optimal pH. Error bars indicate mean ± SD.
(C) Recombinant 6x histidine-tagged proteins were incubated with 5 mg/ml amylopectin, amylopectin was pelleted by ultracentrifugation and proteins in the pellet (P) and supernatant (S) were visualized by Western analysis as described. VHR, *H. sapiens* VHR; Hs, Hs-laforin; Cm, Cm-laforin. Mutated residues are marked with an asterisk as in Figure A.1.A.
(D) A schematic of a non-dividing *C. merolae* cell, defining the position of the chloroplast, mitochondrion, nucleus, and floridean starch. Cells were fixed, incubated with pre-immune serum (left) or α-Cm-laforin antibody (right), probed with a FITC conjugated α-rabbit secondary antibody. Chloroplast were visualized via their autofluoresence. Bar, 3 μm.
(E) *C. merolae* cells were sectioned, probed with α-Cm-laforin and α-rabbit 10 nM gold conjugated secondary antibodies and visualized at 6,000x magnification. An arrow defines the chloroplast, asterisks mark the distal ends of a mitochondrion, and arrowheads mark three (of the many) floridean starch granules. Bar, 500 nm.

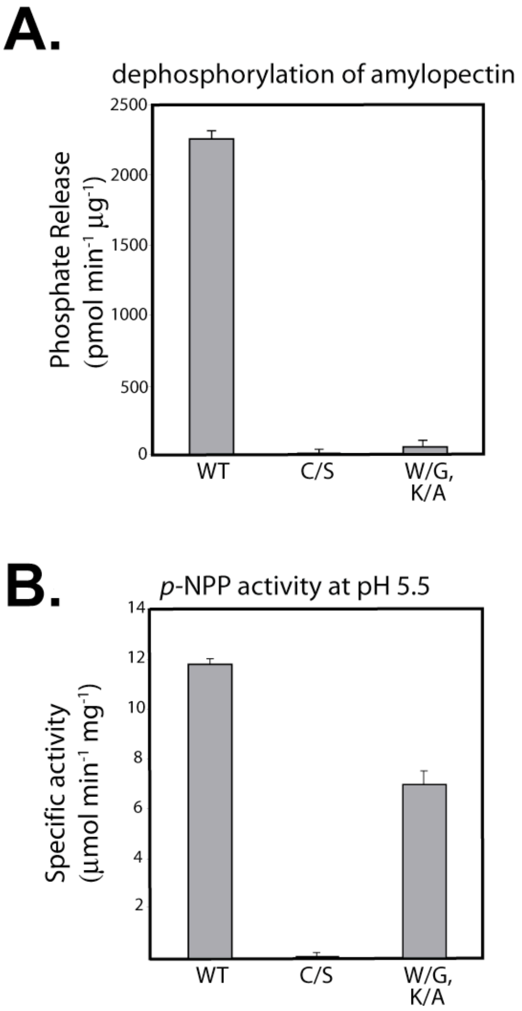


Figure A.4 Phosphatase activity of Cm-laforin mutants.

(A) Specific activity of wild-type Cm-laforin (WT), Cm-laforin-C/S (C/S), and Cm-laforin-W/G, K/A (W/G, K/A) against *p*-NPP at pH 5.5. Error bars indicate mean ± SD. (B) Phosphate release measured by malachite green assays using wild-type Cm-laforin (WT), Cm-laforin-C/S (C/S), and Cm-laforin-W/G, K/A (W/G, K/A) against amylopectin at pH 6.0. Error bars indicate mean ± SD.

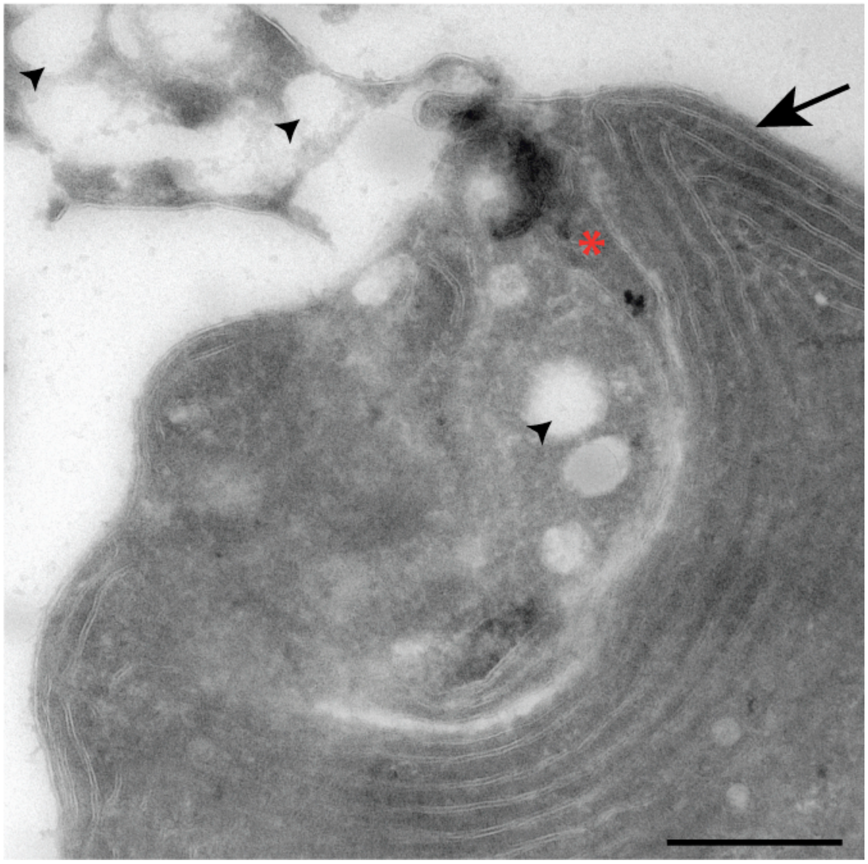


Figure A.5 Immuno-EM of a *C. merolae* cell probed with the secondary antibody alone.

C. merolae cells were treated as in Figure A.3.E, but the primary antibody was pre-immune serum and not α-Cm-laforin antibody. An arrow defines the chloroplast, an asterisk marks the mitochondrion, and arrowheads mark three (of many) floridean starch granules. This cell is visualized at 10,000x magnification. Bar, 500 nm.

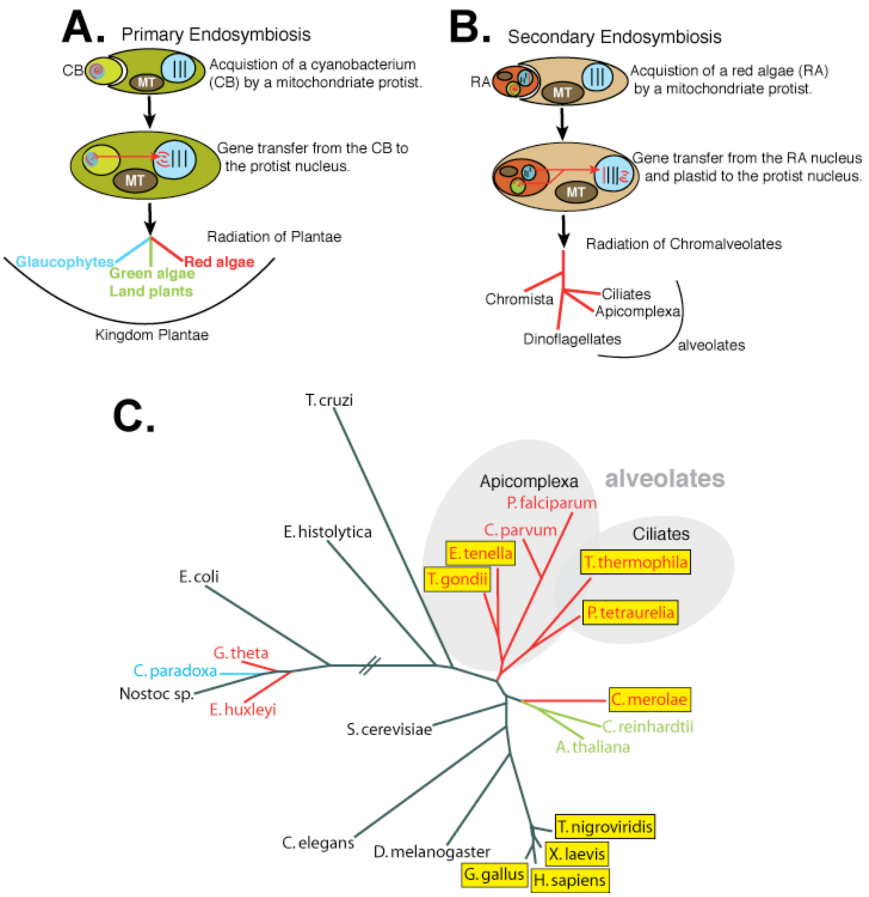
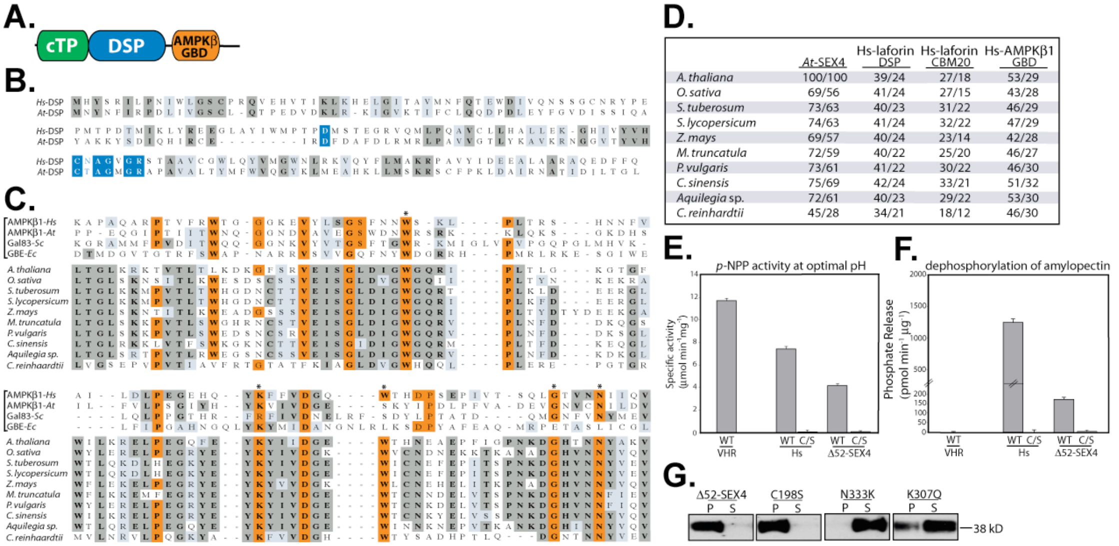


Figure A.6 Evolutionary lineage of laforin.

(A) Primary endosymbiosis hypothesized by the chromalveolate hypothesis (Cavalier-Smith (59)). A cyanobacterium (CB) was engulfed by a mitochondriate protist (Cavalier-Smith (60), Bhattacharya and Medlin (61)). MT, mitochondrion. Over generations, gene transfer occurred between the engulfed CB and protist, the CB was reduced to a plastid bound by two membranes, and the plastid-containing protist radiated into the founding members of kingdom Plantae (Cavalier-Smith (62)).
(B) Secondary endosymbiosis hypothesized by the chromalveolate hypothesis (Cavalier-Smith (59)). A red alga (RA) was engulfed by a mitochondriate protist (Gillott and Gibbs (63)). Over generations, gene transfer occurred from the RA nucleus and plastid to the nucleus of the protist, the RA was reduced to a plastid bound by three or four membranes, and the new protist radiated into the kingdom Chromista and the alveolates, which are collectively referred to as the chromalveolates (Cavalier-Smith (59)). The figure expands on the work of Weber *et al*. (86).
(C) Conservation of laforin orthologues. A phylogeny of the small subunit ribosomal RNA sequences was generated as described in Experimental Procedures, and accession numbers are listed in Table A.5. Organisms containing laforin are boxed in yellow. Organisms from green algal descent are in green, organisms from glaucophyte descent are in blue, and organisms from red algal descent are in red. Alveolates are shaded with a grey background.



	At-SEX4	Hs-laforin DSP	Hs-laforin CBM20	Hs-AMPKβ1 GBD
A. thaliana	100/100	39/24	27/18	53/29
O. sativa	69/56	41/24	27/15	43/28
S. tuberosum	73/63	40/23	31/22	46/29
S. lycopersicum	74/63	41/24	32/22	47/29
Z. mays	69/57	40/24	23/14	42/28
M. truncatula	72/59	40/22	25/20	46/27
P. vulgaris	73/61	41/22	30/22	46/30
C. sinensis	75/69	42/24	33/21	51/32
Aquilegia sp.	72/61	40/23	29/22	53/30
C. reinhardtii	45/28	34/21	18/12	46/30

Figure A.7 Conservation and biochemical properties of SEX4.

(A) Domain topography of SEX4. cTP, chloroplast-targeting peptide; DSP, dual specific phosphatase; GBD, glycogen binding domain.
(B) Alignment of the DSP of Hs-laforin (*Hs*-DSP) and the DSP of SEX4 (*At*-DSP). Residues boxed in blue are part of the DSP catalytic site, those boxed in dark grey are identical, and those boxed in light grey are conserved substitutions.
(C) Alignment of the AMPKβ-GBD of four founding members of the family (top four, marked by a bracket; Polekhina *et al*. (69)) and the AMPKβ-GBD of SEX4 orthologues. Residues boxed in orange are highly conserved amino acids amongst multiple AMPKβ-GBD (Polekhina *et al*. (69)) proteins. Residues boxed in dark grey are identical, and those boxed in light grey are conserved substitutions. Asterisks mark residues necessary for carbohydrate binding (Polekhina *et al*. (69,70)). Accession numbers are listed in Table A.3.
(D) Percent similarity and identity of full-length At-SEX4 compared with SEX4 orthologues, the DSP of Hs-laforin to the DSP of each SEX4 orthologue, and the CBM20 of Hs-laforin to the glycogen-binding domain (GBD) of each SEX4 orthologue as well as the percent similarity of the GBD of Hs-AMPKβ1 to the GBD of each SEX4 orthologue.
(E) Specific activity of VHR, Hs-laforin, Hs-laforin-C/S, Δ52-SEX4, and Δ52-SEX4-C/S against *p*-NPP at their respective optimal pH. Error bars indicate mean ± SD.
(F) Phosphate release measured by malachite green assays using VHR, Hs-laforin, Hs-laforin-C/S, Δ52-SEX4, and Δ52-SEX4-C/S against amylopectin at their respective optimal pH. Error bars indicate mean ± SD.
(G) At-SEX4 binding to amylopectin. Recombinant 6x histidine-tagged proteins were incubated with 5 mg/ml amylopectin, amylopectin was pelleted by ultracentrifugation, and proteins in the pellet (P) and supernatant (S) were visualized by Western analysis as described. Δ52-SEX4, amino terminal truncation of SEX4; C198S, Δ52-SEX4-C198S; N333K, Δ52-SEX4-N333K; K307Q, Δ52-SEX4-K307Q.

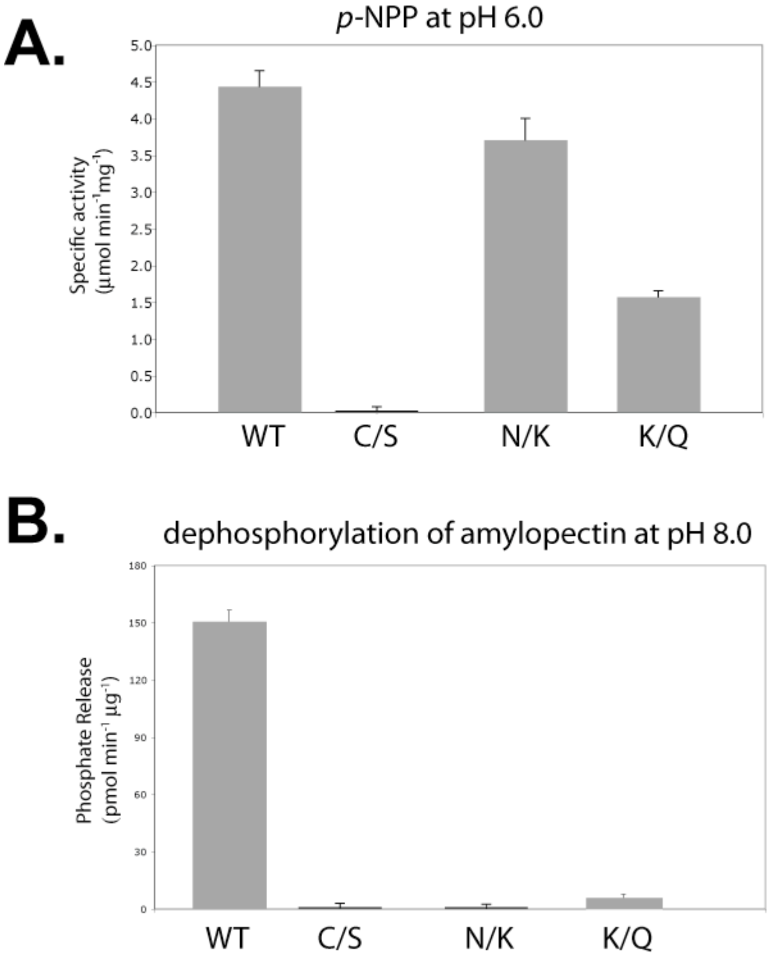
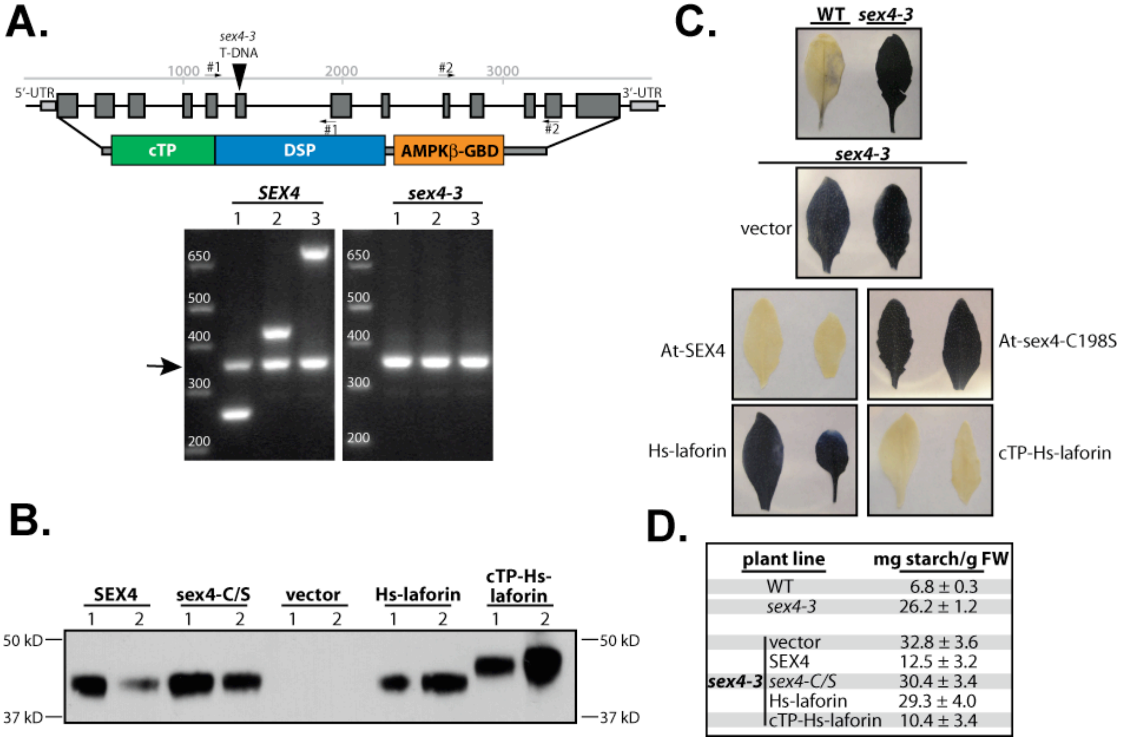


Figure A.8 Phosphatase activity of SEX4 mutants**.**

(A) Specific activity of wild-type SEX4 and SEX4 mutants against *p*-NPP at pH 6.0. (B) Phosphate release measured by malachite green assays using SEX4 and SEX4 mutants against amylopectin at pH 8.0. Δ52-SEX4-H_6, WT; Δ52-SEX4-H_6-C198S, C/S; Δ52-SEX4-H_6-N333K, N/K; Δ52-SEX4-H_6-K307Q, K/Q. Error bars indicate mean ± SD.



	plant line	mg starch/g FW
	WT	6.8 ± 0.3
	sex4-3	26.2 ± 1.2
sex4-3	vector	32.8 ± 3.6
	SEX4	12.5 ± 3.2
	sex4-C/S	30.4 ± 3.4
	Hs-laforin	29.3 ± 4.0
	cTP-Hs-laforin	10.4 ± 3.4

Figure A.9 Hs-laforin is a functional equivalent of SEX4.

(A) *sex4-3* disruption and gene expression. A schematic of the At3g51280 locus with exons represented by dark grey boxes, the T-DNA insertion depicted as a black arrowhead, and a schematic of the SEX4 protein. Small arrows denote the primer sets used to confirm the absence of a *SEX4* transcript in *sex4-3* plants. RT-PCR results from wild-type and *sex4-3* isolated mRNA. Lane 1, primer set 1 from the schematic; lane 2, primer set 2; lane 3, the 5' primer from 1 and the 3' primer from 2. The arrow indicates the size of the positive control *UBC5*.
(B) Protein expression of transgenes. Tissue samples for Western analysis were taken from two independent T_2 plants transformed with empty vector or an HA epitope-tagged transgene as indicated. 30 μg of whole tissue lysate was loaded in each lane.
(C) Complementation of the starch excess phenotype. One leaf from the same two independent T_2 plants as in (B) were decolorized with hot ethanol and stained with iodine. Wild-type (WT) leaves contain little starch and do not stain; conversely, *sex4-3* leaves contain high amounts of starch and are darkly stained. Similarly, transgenes that complement the phenotype do not stain, whereas those that do not complement stain dark.
(D) Quantitation of the complementation. Starch was quantified from leaves of wild-type, *sex4-3*, and *sex4-3* transgenic plants. Each sample is the mean of replicate samples ± standard error. For *sex4-3* transgenic plants, six independent T_2 plants isogenic to those used in (B) and (C) were used to quantitate the amount of starch in leaves. FW, fresh weight.

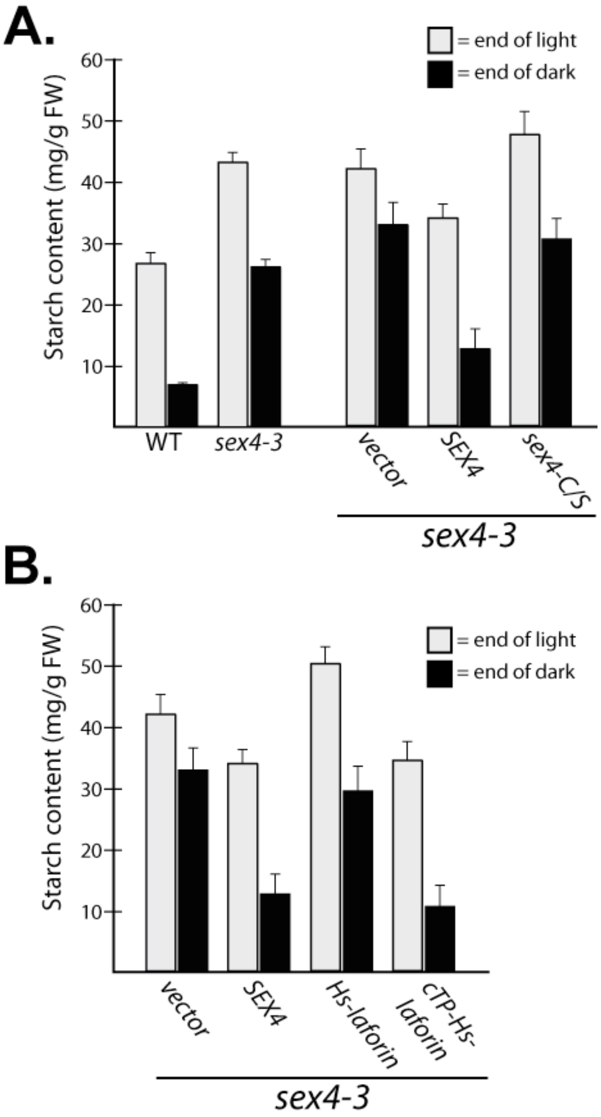
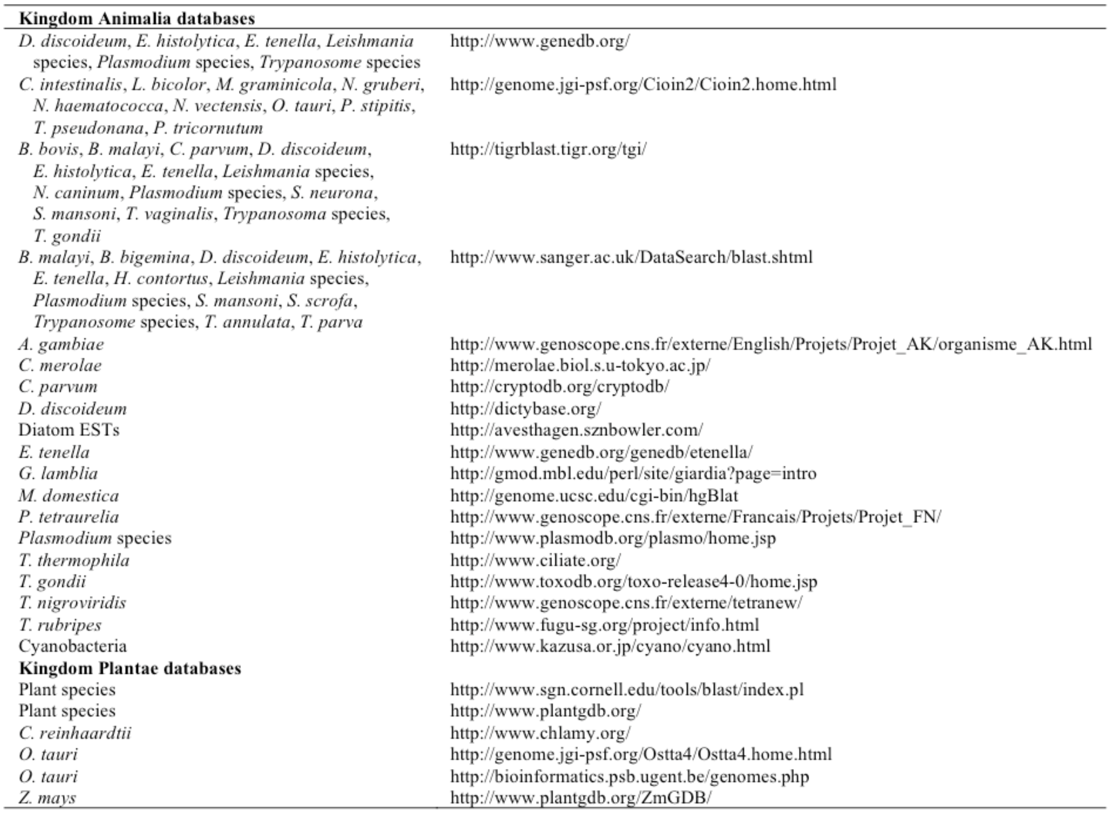


Figure A.10 Quantitation of starch content.

(A) Six independent T_2 plants isogenic to those used in Figure A.9.B-D were used to quantitate the amount of starch in leaves at the end of the light cycle (light grey bars) and at the end of the dark cycle (black bars). Starch content in wild-type (WT), *sex4-3*, and *sex4-3* transgenic plants stably transformed with vector, *SEX4* or *sex4-C/S*. (B) Starch content in *sex4-3* transgenic plants stably transformed with vector, *SEX4*, *Hs-laforin*, or *cTP-Hs-laforin*. Each sample is the mean of six replicate samples ± standard error. FW, fresh weight.

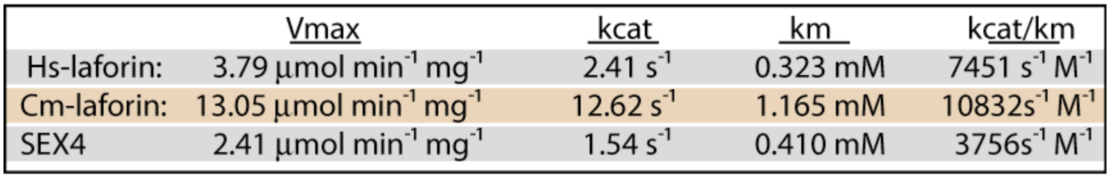
Table A.1 Non-NCBI databases.



Kingdom Animalia databases	
D. discoideum, *E. histolytica*, *E. tenella*, *Leishmania* species, *Plasmodium* species, *Trypanosome* species	http://www.genedb.org/
C. intestinalis, *L. bicolor*, *M. graminicola*, *N. gruberi*, *N. haematococca*, *N. vectensis*, *O. tauri*, *P. stipitis*, *T. pseudonana*, *P. tricornutum*	http://genome.jgi-psf.org/Cioin2/Cioin2.home.html
B. bovis, *B. malayi*, *C. parvum*, *D. discoideum*, *E. histolytica*, *E. tenella*, *Leishmania* species, *N. caninum*, *Plasmodium* species, *S. neurona*, *S. mansoni*, *T. vaginalis*, *Trypanosoma* species, *T. gondii*	http://tigrblast.tigr.org/tgi/
B. malayi, *B. bigemina*, *D. discoideum*, *E. histolytica*, *E. tenella*, *H. contortus*, *Leishmania* species, *Plasmodium* species, *S. mansoni*, *S. scrofa*, *Trypanosome* species, *T. annulata*, *T. parva*	http://www.sanger.ac.uk/DataSearch/blast.shtml
A. gambiae	http://www.genoscope.cns.fr/externe/English/Projets/Projet_AK/organisme_AK.html
C. merolae	http://merolae.biol.s.u-tokyo.ac.jp/
C. parvum	http://cryptodb.org/cryptodb/
D. discoideum	http://dictybase.org/
Diatom ESTs	http://avesthagen.sznbowler.com/
E. tenella	http://www.genedb.org/genedb/etenella/
G. lamblia	http://gmod.mbl.edu/perl/site/giardia?page=intro
M. domestica	http://genome.ucsc.edu/cgi-bin/hgBlat
P. tetraurelia	http://www.genoscope.cns.fr/externe/Francais/Projets/Projet_FN/
Plasmodium species	http://www.plasmodb.org/plasmo/home.jsp
T. thermophila	http://www.ciliate.org/
T. gondii	http://www.toxodb.org/toxo-release4-0/home.jsp
T. nigroviridis	http://www.genoscope.cns.fr/externe/tetranew/
T. rubripes	http://www.fugu-sg.org/project/info.html
Cyanobacteria	http://www.kazusa.or.jp/cyano/cyano.html
Kingdom Plantae databases	
Plant species	http://www.sgn.cornell.edu/tools/blast/index.pl
Plant species	http://www.plantgdb.org/
C. reinhaardtii	http://www.chlamy.org/
O. tauri	http://genome.jgi-psf.org/Ostta4/Ostta4.home.html
O. tauri	http://bioinformatics.psb.ugent.be/genomes.php
Z. mays	http://www.plantgdb.org/ZmGDB/

The non-NCBI databases used to search for laforin orthologues and the organism's genome in each database are listed. Many of the databases were found by performing a Google search of the organism's genus name and genome.

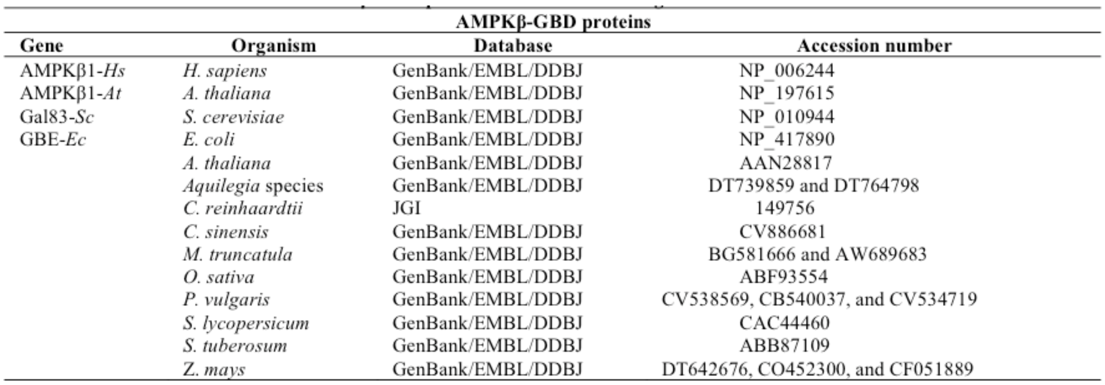
Table A.2 Kinetics of laforin orthologues and SEX4 utilizing *p*-NPP.



	Vmax	kcat	km	kcat/km
Hs-laforin:	3.79 µmol min^{-1} mg^{-1}	2.41 s^{-1}	0.323 mM	7451 s^{-1} M^{-1}
Cm-laforin:	13.05 µmol min^{-1} mg^{-1}	12.62 s^{-1}	1.165 mM	10832s^{-1} M^{-1}
SEX4	2.41 µmol min^{-1} mg^{-1}	1.54 s^{-1}	0.410 mM	3756s^{-1} M^{-1}

Kinetics of *H. sapiens* laforin (Hs-laforin), *C. merolae* laforin (Cm-laforin), and *A. thaliana*-Δ52-SEX4 (SEX4) utilizing *p*-NPP at pH 5.5, pH 6.0, and pH 6.0, respectively.

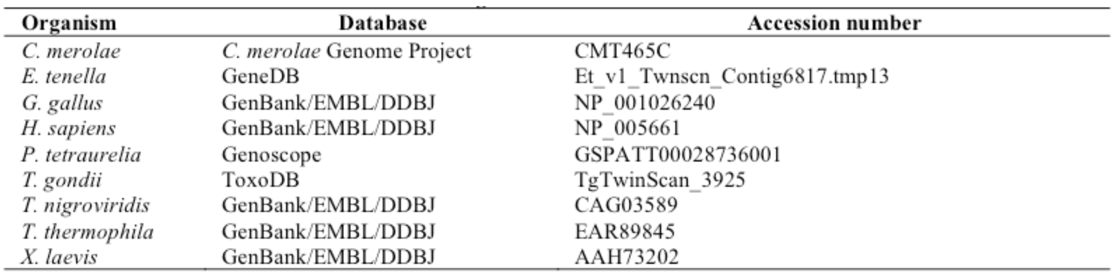
Table A.3 Accession numbers for AMPKβ-GBD proteins and SEX4 orthologues.



AMPKβ-GBD proteins			
Gene	**Organism**	**Database**	**Accession number**
AMPKβ1-*Hs*	*H. sapiens*	GenBank/EMBL/DDBJ	NP_006244
AMPKβ1-*At*	*A. thaliana*	GenBank/EMBL/DDBJ	NP_197615
Gal83-*Sc*	*S. cerevisiae*	GenBank/EMBL/DDBJ	NP_010944
GBE-*Ec*	*E. coli*	GenBank/EMBL/DDBJ	NP_417890
	A. thaliana	GenBank/EMBL/DDBJ	AAN28817
	Aquilegia species	GenBank/EMBL/DDBJ	DT739859 and DT764798
	C. reinhaardtii	JGI	149756
	C. sinensis	GenBank/EMBL/DDBJ	CV886681
	M. truncatula	GenBank/EMBL/DDBJ	BG581666 and AW689683
	O. sativa	GenBank/EMBL/DDBJ	ABF93554
	P. vulgaris	GenBank/EMBL/DDBJ	CV538569, CB540037, and CV534719
	S. lycopersicum	GenBank/EMBL/DDBJ	CAC44460
	S. tuberosum	GenBank/EMBL/DDBJ	ABB87109
	Z. mays	GenBank/EMBL/DDBJ	DT642676, CO452300, and CF051889

Some protein sequences were obtained from one or multiple cDNA sequences, and in these cases, the cDNA accession numbers are listed.

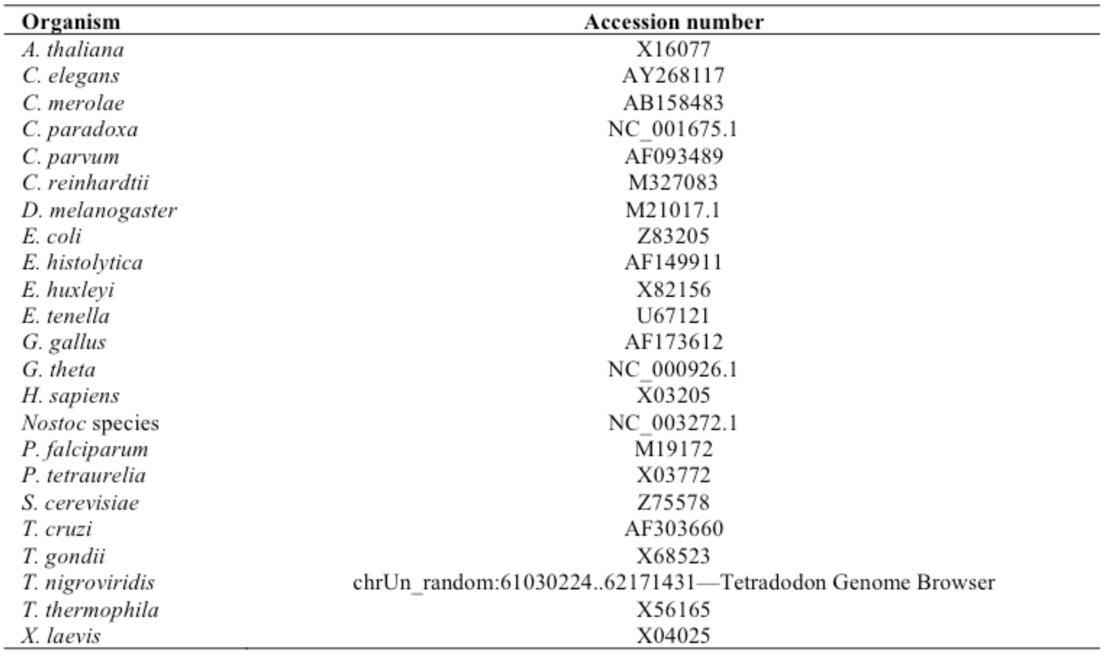
Table A.4 Accession numbers for laforin orthologues.



Organism	Database	Accession number
C. merolae	*C. merolae* Genome Project	CMT465C
E. tenella	GeneDB	Et_v1_Twnscn_Contig6817.tmp13
G. gallus	GenBank/EMBL/DDBJ	NP_001026240
H. sapiens	GenBank/EMBL/DDBJ	NP_005661
P. tetraurelia	Genoscope	GSPATT00028736001
T. gondii	ToxoDB	TgTwinScan_3925
T. nigroviridis	GenBank/EMBL/DDBJ	CAG03589
T. thermophila	GenBank/EMBL/DDBJ	EAR89845
X. laevis	GenBank/EMBL/DDBJ	AAH73202

Listed on the left are the organisms from the alignment in Figure A.1.A, and on the right are the accession numbers for the laforin orthologues. All of the accession numbers are from NCBI GenBank/EMBL/DDBJ unless otherwise noted.

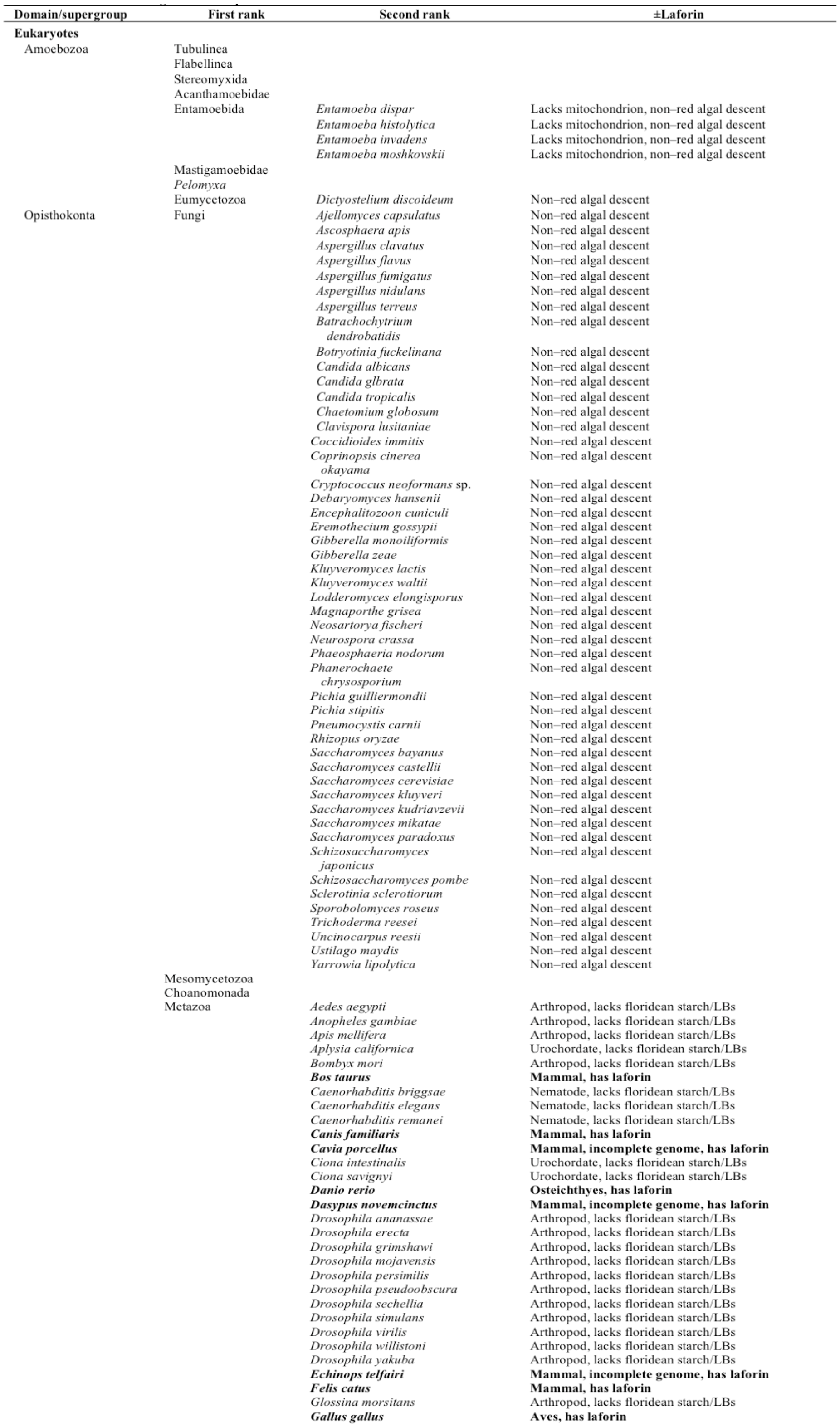
Table A.5 Small subunit ribosomal RNA accession numbers.



Organism	Accession number
A. thaliana	X16077
C. elegans	AY268117
C. merolae	AB158483
C. paradoxa	NC_001675.1
C. parvum	AF093489
C. reinhardtii	M327083
D. melanogaster	M21017.1
E. coli	Z83205
E. histolytica	AF149911
E. huxleyi	X82156
E. tenella	U67121
G. gallus	AF173612
G. theta	NC_000926.1
H. sapiens	X03205
Nostoc species	NC_003272.1
P. falciparum	M19172
P. tetraurelia	X03772
S. cerevisiae	Z75578
T. cruzi	AF303660
T. gondii	X68523
T. nigroviridis	chrUn_random:61030224..62171431—Tetradodon Genome Browser
T. thermophila	X56165
X. laevis	X04025

Listed on the left are the organisms from the phylogeny in Figure A.6.C, and on the right are the accession numbers for the small subunit ribosomal RNA genes. All of the accession numbers are from NCBI GenBank/EMBL/DDBJ unless otherwise noted.

Table A.6 Genomes investigated for the presence of laforin.

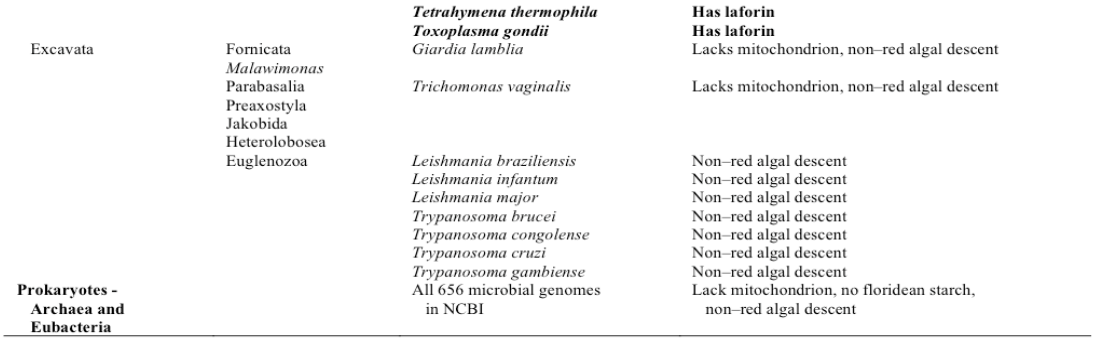


Domain/supergroup	First rank	Second rank	±Laforin
Eukaryotes			
Amoebozoa	Tubulinea		
	Flabellinea		
	Stereomyxida		
	Acanthamoebidae		
	Entamoebida	*Entamoeba dispar*	Lacks mitochondrion, non–red algal descent
		Entamoeba histolytica	Lacks mitochondrion, non–red algal descent
		Entamoeba invadens	Lacks mitochondrion, non–red algal descent
		Entamoeba moshkovskii	Lacks mitochondrion, non–red algal descent
	Mastigamoebidae		
	Pelomyxa		
	Eumycetozoa	*Dictyostelium discoideum*	Non–red algal descent
Opisthokonta	Fungi	*Ajellomyces capsulatus*	Non–red algal descent
		Ascosphaera apis	Non–red algal descent
		Aspergillus clavatus	Non–red algal descent
		Aspergillus flavus	Non–red algal descent
		Aspergillus fumigatus	Non–red algal descent
		Aspergillus nidulans	Non–red algal descent
		Aspergillus terreus	Non–red algal descent
		Batrachochytrium dendrobatidis	Non–red algal descent
		Botryotinia fuckelinana	Non–red algal descent
		Candida albicans	Non–red algal descent
		Candida glbrata	Non–red algal descent
		Candida tropicalis	Non–red algal descent
		Chaetomium globosum	Non–red algal descent
		Clavispora lusitaniae	Non–red algal descent
		Coccidioides immitis	Non–red algal descent
		Coprinopsis cinerea okayama	Non–red algal descent
		Cryptococcus neoformans sp.	Non–red algal descent
		Debaryomyces hansenii	Non–red algal descent
		Encephalitozoon cuniculi	Non–red algal descent
		Eremothecium gossypii	Non–red algal descent
		Gibberella monoiliformis	Non–red algal descent
		Gibberella zeae	Non–red algal descent
		Kluyveromyces lactis	Non–red algal descent
		Kluyveromyces waltii	Non–red algal descent
		Lodderomyces elongisporus	Non–red algal descent
		Magnaporthe grisea	Non–red algal descent
		Neosartorya fischeri	Non–red algal descent
		Neurospora crassa	Non–red algal descent
		Phaeosphaeria nodorum	Non–red algal descent
		Phanerochaete chrysosporium	Non–red algal descent
		Pichia guilliermondii	Non–red algal descent
		Pichia stipitis	Non–red algal descent
		Pneumocystis carnii	Non–red algal descent
		Rhizopus oryzae	Non–red algal descent
		Saccharomyces bayanus	Non–red algal descent
		Saccharomyces castellii	Non–red algal descent
		Saccharomyces cerevisiae	Non–red algal descent
		Saccharomyces kluyveri	Non–red algal descent
		Saccharomyces kudriavzevii	Non–red algal descent
		Saccharomyces mikatae	Non–red algal descent
		Saccharomyces paradoxus	Non–red algal descent
		Schizosaccharomyces japonicus	Non–red algal descent
		Schizosaccharomyces pombe	Non–red algal descent
		Sclerotinia sclerotiorum	Non–red algal descent
		Sporobolomyces roseus	Non–red algal descent
		Trichoderma reesei	Non–red algal descent
		Uncinocarpus reesii	Non–red algal descent
		Ustilago maydis	Non–red algal descent
		Yarrowia lipolytica	Non–red algal descent
	Mesomycetozoa		
	Choanomonada		
	Metazoa	*Aedes aegypti*	Arthropod, lacks floridean starch/LBs
		Anopheles gambiae	Arthropod, lacks floridean starch/LBs
		Apis mellifera	Arthropod, lacks floridean starch/LBs
		Aplysia californica	Urochordate, lacks floridean starch/LBs
		Bombyx mori	Arthropod, lacks floridean starch/LBs
		Bos taurus	**Mammal, has laforin**
		Caenorhabditis briggsae	Nematode, lacks floridean starch/LBs
		Caenorhabditis elegans	Nematode, lacks floridean starch/LBs
		Caenorhabditis remanei	Nematode, lacks floridean starch/LBs
		Canis familiaris	**Mammal, has laforin**
		Cavia porcellus	**Mammal, incomplete genome, has laforin**
		Ciona intestinalis	Urochordate, lacks floridean starch/LBs
		Ciona savignyi	Urochordate, lacks floridean starch/LBs
		Danio rerio	**Osteichthyes, has laforin**
		Dasypus novemcinctus	**Mammal, incomplete genome, has laforin**
		Drosophila ananassae	Arthropod, lacks floridean starch/LBs
		Drosophila erecta	Arthropod, lacks floridean starch/LBs
		Drosophila grimshawi	Arthropod, lacks floridean starch/LBs
		Drosophila mojavensis	Arthropod, lacks floridean starch/LBs
		Drosophila persimilis	Arthropod, lacks floridean starch/LBs
		Drosophila pseudoobscura	Arthropod, lacks floridean starch/LBs
		Drosophila sechellia	Arthropod, lacks floridean starch/LBs
		Drosophila simulans	Arthropod, lacks floridean starch/LBs
		Drosophila virilis	Arthropod, lacks floridean starch/LBs
		Drosophila willistoni	Arthropod, lacks floridean starch/LBs
		Drosophila yakuba	Arthropod, lacks floridean starch/LBs
		Echinops telfairi	**Mammal, incomplete genome, has laforin**
		Felis catus	**Mammal, has laforin**
		Glossina morsitans	Arthropod, lacks floridean starch/LBs
		Gallus gallus	**Aves, has laforin**

Table A.6 (Continued) Genomes investigated for the presence of laforin.

		Haemonchus contortus	Roundworm, lacks floridean starch/LBs
		Homo sapiens	**Mammal, has laforin**
		Loxodonta africana	**Mammal, incomplete genome, has laforin**
		Macaca mulatta	**Mammal, has laforin**
		Monodelphis domestica	**Mammal, incomplete genome, has laforin**
		Mus musculus	**Mammal, has laforin**
		Myotis lucifugus	**Mammal, incomplete genome, has laforin**
		Ornithorhynchus anatinus	**Mammal, incomplete genome, has laforin**
		Oryctolagus cuniculus	**Mammal, incomplete genome, has laforin**
		Oryzias latipes	**Osteichthyes, has laforin**
		Otolemur garnettii	**Mammal, incomplete genome, has laforin**
		Pan troglodytes	**Mammal, incomplete genome, has laforin**
		Pongo pygmaeus	**Mammal, incomplete genome, has laforin**
		Rattus norvegicus	**Mammal, has laforin**
		Schistosoma mansoni	Flatworm, lacks floridean starch/LBs
		Sorex araneus	**Mammal, incomplete genome, has laforin**
		Strongylocentrotus purpuratus	Urochordate, lacks floridean starch/LBs
		Sus scrofa	**Mammal, has laforin**
		Takifugu rubripes	**Osteichthyes, has laforin**
		Tetraodon nigroviridis	**Osteichthyes, has laforin**
		Tribolium castaneum	Arthropod, lacks floridean starch/LBs
		Xenopus laevis	**Amphibian, has laforin**
Rhizaria	Cercozoa	*Phytophthora infestans*	Plant pathogen, lacks floridean starch
		Phytophthora ramorum	Plant pathogen, lacks floridean starch
		Phytophthora sojae	Plant pathogen, lacks floridean starch
	Haplosporidia		
	Foraminifera		
	Gromia		
	Radiolaria		
Archaeplastida	Glaucophyta		
	Rhodophyceae	***Cyanidioschyzon merolae***	**Has laforin**
		Galdieria sulphuraria	Incomplete genome, likely has laforin
	Chloroplastida	*Arabidopsis thaliana*	Land plant, has SEX4
		Aquilegia sp.	Land plant, has SEX4
		Chlamydomonas reinhardtii	Land plant, has SEX4
		Citrus sinensis	Land plant, has SEX4
		Medicago truncatula	Land plant, has SEX4
		Oryza sativa	Land plant, has SEX4
		Ostreococcus tauri	Land plant, has SEX4
		Phaseolus vulgaris	Land plant, has SEX4
		Solanum lycopersicum	Land plant, has SEX4
		Solanum tuberosum	Land plant, has SEX4
		Sorghum bicolor	Land plant, has SEX4
		Triticum aestivum	Land plant, has SEX4
		Zea mays	Land plant, has SEX4
Chromalveolata	Cryptophyceae	*Guillardia theta*	Nucleomorph sequenced, nuclear genome not sequenced, likely has laforin
	Haptophyta	*Emiliania huxleyi*	Phytoplankton, lacks floridean starch
	Stramenopiles	*Thalassiosira pseudonana*	Diatom, lacks floridean starch
		Phaeodactylum tricornutum	Diatom, lacks floridean starch
	Alveolata	*Babesia bovis*	Lacks floridean starch
		Babesia bigemina	Lacks floridean starch
		Babesia malayi	Lacks floridean starch
		Cryptosporidium parvum	Lacks mitochondrion
		Cryptosporidium hominis	Lacks mitochondrion
		Eimeria tenella	**Has laforin**
		Neospora caninum	Incomplete genome, likely has laforin
		Paramecium tetraurelia	**Has laforin**
		Plasmodium berghei	Lacks floridean starch
		Plasmodium chabaudi	Lacks floridean starch
		Plasmodium falciparum	Lacks floridean starch
		Plasmodium gallinaceum	Lacks floridean starch
		Plasmodium knowlesi	Lacks floridean starch
		Plasmodium reichenowi	Lacks floridean starch
		Plasmodium vivax	Lacks floridean starch
		Plasmodium yhoelii yoelii	Lacks floridean starch
		Sarcocystis neurona	Incomplete genome, likely has laforin
		Theileria annulata	Lacks floridean starch
		Theileria parva	Lacks floridean starch

Table A.6 (Continued) Genomes investigated for the presence of laforin.



		Tetrahymena thermophila	**Has laforin**
		Toxoplasma gondii	**Has laforin**
Excavata	Fornicata	*Giardia lamblia*	Lacks mitochondrion, non–red algal descent
	Malawimonas		
	Parabasalia	*Trichomonas vaginalis*	Lacks mitochondrion, non–red algal descent
	Preaxostyla		
	Jakobida		
	Heterolobosea		
	Euglenozoa	*Leishmania braziliensis*	Non–red algal descent
		Leishmania infantum	Non–red algal descent
		Leishmania major	Non–red algal descent
		Trypanosoma brucei	Non–red algal descent
		Trypanosoma congolense	Non–red algal descent
		Trypanosoma cruzi	Non–red algal descent
		Trypanosoma gambiense	Non–red algal descent
Prokaryotes - Archaea and Eubacteria		All 656 microbial genomes in NCBI	Lack mitochondrion, no floridean starch, non–red algal descent

The genome of each organism was searched for laforin using the appropriate database (Table A.1). If laforin was absent, an extensive literature search was performed on the organism to determine which of the three criteria it lacked: red algal descent, mitochondrion, and/or floridean starch/LBs. If the organism lacked laforin, at least one of the three criteria that it lacks is presented beside its name. The organism names of genomes containing laforin are bold. The organism names of genomes that are nearing completion and that contain laforin based on our predictions are underlined. The phrase “incomplete genome, has laforin” refers to organisms with incomplete genomes but in which a partial CBM and DSP corresponding to laforin was found. Organism classification is based on Adl *et al*. (87).

REFERENCES

1. Lafora, G. R., and Gluck, B. (1911) *Z Ges Neurol Psychiatr* **6**, 1-14

2. Van Hoof, F., and Hageman-Bal, M. (1967) *Acta Neuropathol* **7**(4), 315-326

3. Schwarz, G. A., and Yanoff, M. (1965) *Arch Neurol* **12**, 172-188

4. Van Heycop Ten Ham, M. W. (1975) Lafora disease, a form of progressive myoclonus epilepsy. In: Vinken, P. J. a. Bruyn, G. W. (ed). *Handbook of Clinical Neurology*, North Holland Publishing Company, Holland, Amsterdam

5. Lafora, G. R. (1911) *Virchows Arch. f. Path. Anat.* **205**, 295

6. Collins, G. H., Cowden, R. R., and Nevis, A. H. (1968) *Arch Pathol* **86**(3), 239-254

7. Harriman, D. G., Millar, J. H., and Stevenson, A. C. (1955) *Brain* **78**(3), 325-349

8. Carpenter, S., and Karpati, G. (1981) *Neurology* **31**(12), 1564-1568

9. Yokoi, S., Austin, J., Witmer, F., and Sakai, M. (1968) *Arch Neurol* **19**(1), 15-33

10. Chan, E. M., Young, E. J., Ianzano, L., Munteanu, I., Zhao, X., Christopoulos, C. C., Avanzini, G., Elia, M., Ackerley, C. A., Jovic, N. J., Bohlega, S., Andermann, E., Rouleau, G. A., Delgado-Escueta, A. V., Minassian, B. A., and Scherer, S. W. (2003) *Nat Genet* **35**(2), 125-127

11. Gentry, M. S., Worby, C. A., and Dixon, J. E. (2005) *Proc Natl Acad Sci U S A* **102**(24), 8501-8506

12. Ianzano, L., Zhang, J., Chan, E. M., Zhao, X., Lohi, H., Scherer, S. W., and Minassian, B. A. (2005) *Hum Mutat* **26**(4), 397

13. Minassian, B. A., Lee, J. R., Herbrick, J. A., Huizenga, J., Soder, S., Mungall, A. J., Dunham, I., Gardner, R., Fong, C. Y., Carpenter, S., Jardim, L., Satishchandra, P., Andermann, E., Snead, O. C., 3rd, Lopes-Cendes, I., Tsui, L. C., Delgado-Escueta, A. V., Rouleau, G. A., and Scherer, S. W. (1998) *Nat Genet* **20**(2), 171-174

14. Serratosa, J. M., Gomez-Garre, P., Gallardo, M. E., Anta, B., de Bernabe, D. B., Lindhout, D., Augustijn, P. B., Tassinari, C. A., Malafosse, R. M., Topcu, M., Grid, D., Dravet, C., Berkovic, S. F., and de Cordoba, S. R. (1999) *Hum Mol Genet* **8**(2), 345-352

15. Wang, J., Stuckey, J. A., Wishart, M. J., and Dixon, J. E. (2002) *J Biol Chem* **277**(4), 2377-2380

16. Denu, J. M., Stuckey, J. A., Saper, M. A., and Dixon, J. E. (1996) *Cell* **87**(3), 361-364

17. Yuvaniyama, J., Denu, J. M., Dixon, J. E., and Saper, M. A. (1996) *Science* **272**, 1328-1331

18. Ganesh, S., Agarwala, K. L., Ueda, K., Akagi, T., Shoda, K., Usui, T., Hashikawa, T., Osada, H., Delgado-Escueta, A. V., and Yamakawa, K. (2000) *Hum Mol Genet* **9**(15), 2251-2261

19. Chan, E. M., Ackerley, C. A., Lohi, H., Ianzano, L., Cortez, M. A., Shannon, P., Scherer, S. W., and Minassian, B. A. (2004) *Hum Mol Genet* **13**(11), 1117-1129

20. Zolnierowicz, S. (2000) *Biochem Pharmacol* **60**(8), 1225-1235

21. Alonso, A., Sasin, J., Bottini, N., Friedberg, I., Friedberg, I., Osterman, A., Godzik, A., Hunter, T., Dixon, J., and Mustelin, T. (2004) *Cell* **117**(6), 699-711

22. Coutinho, P. M., and Henrissat, B. (1999) Carbohydrate-active enzymes: an integrated database approach. In: H.J. Gilbert, G.J. Davies, B. Henrissat and B. Svensson (ed). *Recent Advances in Carbohydrate Bioengineering*, The Royal Society of Chemistry, Cambridge

23. Boraston, A. B., Bolam, D. N., Gilbert, H. J., and Daview, G. J. (2004) *Biochem J* **382**, 769-781

24. Rodriguez-Sanoja, R., Oviedo, N., and Sanchez, S. (2005) *Curr Opin Microbiol* **8**(3), 260-267

25. Worby, C. A., Gentry, M. S., and Dixon, J. E. (2006) *J Biol Chem* **281**(41), 30412-30418

26. Ganesh, S., Delgado-Escueta, A. V., Sakamoto, T., Avila, M. R., Machado-Salas, J., Hoshii, Y., Akagi, T., Gomi, H., Suzuki, T., Amano, K., Agarwala, K. L., Hasegawa, Y., Bai, D. S., Ishihara, T., Hashikawa, T., Itohara, S., Cornford, E. M., Niki, H., and Yamakawa, K. (2002) *Hum Mol Genet* **11**(11), 1251-1262

27. Yokoi, S., Austin, J., and Witmer, F. (1967) *J Neuropathol Exp Neurol* **26**(1), 125-127

28. Sakai, M., Austin, J., Witmer, F., and Trueb, L. (1970) *Neurology* **20**(2), 160-176

29. Peat, S., Turvey, J. R., and Evans, J. M. (1959) *J Chem Soc* 3341-3344

30. Coppin, A., Varré, J., Lienard, L., Dauvillée, D., Guérardel, Y., Soyer-Gobillard, M., Buléon, A., Ball, S., and Stanislas Tomavo. (2005) *J Mol Evol* **60**(2), 257-267

31. Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., Wu, H. C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S. X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H. L., Tripp, M., Chang, C. H., Lee, J. M., Toriumi, M., Chan, M. M. H., Tang, C. C., Onodera, C. S., Deng, J. M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A. D., Gurjal, M., Hansen, N. F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V. W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P. X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E. K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R. W., Theologis, A., and Ecker, J. R. (2003) *Science* **302**(5646), 842-846

32. Denu, J. M., Zhou, G., Wu, L., Zhao, R., Yuvaniyama, J., Saper, M. A., and Dixon, J. E. (1995) *J Biol Chem* **270**(8), 3796-3803

33. Neff, M. M., Nguyen, S. M., Malancharuvil, E. J., Fujioka, S., Noguchi, T., Seto, H., Tsubuki, M., Honda, T., Takatsuto, S., Yoshida, S., and Chory, J. (1999) *Proc Natl Acad Sci U S A* **96**(26), 15316-15323

34. Hajdukiewicz, P., Svab, Z., and Maliga, P. (1994) *Plant Mol Biol* **25**(6), 989-994

35. Kerk, D., Conley, T.R., Rodriguez, F.A., Tran, H.T., Nimick, M., Muench, D.G., Moorhead, G.B. (2006) *Plant J* **46**(3), 400-413

36. Niittyla, T., Comparot-Moss, S., Lue, W.-L., Messerli, G., Trevisan, M., Seymour, M. D. J., Gatehouse, J. A., Villadsen, D., Smith, S. M., Chen, J., Zeeman, S. C., and Smith, A. M. (2006) *J Biol Chem* **281**(17), 11815-11818

37. Harder, K. W., Owen, P., Wong, L. K., Aebersold, R., Clark-Lewis, I., and Jirik, F. R. (1994) *Biochem J* **298**, 395-401

38. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res* **25**(17), 3389-3402

39. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res* **22**(22), 4673-4680

40. Bingham, J., and Sudarsanam, S. (2000) *Bioinformatics* **16**(7), 660-661

41. Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C. C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D. E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W. L., Berry, C. C., and Ecker, J. R. (2003) *Science* **301**(5633), 653-657

42. Clough, S. J., and Bent, A. F. (1998) *Plant J* **16**(6), 735-743

43. Valvekens, D., Montagu, M. V., and Lijsebettens, M. V. (1988) *Proc Natl Acad Sci U S A* **85**(15), 5536-5540

44. Kotting, O., Pusch, K., Tiessen, A., Geigenberger, P., Steup, M., and Ritte, G. (2005) *Plant Physiol* **137**(1), 242-252

45. Nimchuk, Z., Marois, E., Kjemtrup, S., Leister, R. T., Katagiri, F., and Dangl, J. L. (2000) *Cell* **101**(4), 353-363

46. Toda, K., Takano, H., Miyagishima, S., Kuroiwa, H., and Kuroiwa, T. (1998) *Biochim Biophys Acta* **1403**(1), 72-84

47. Minoda, A., Sakagami, R., Yagisawa, F., Kuroiwa, T., and Tanaka, K. (2004) *Plant Cell Physiol* **45**(6), 667-671

48. Nishida, K., Misumi, O., Yagisawa, F., Kuroiwa, H., Nagata, T., and Kuroiwa, T. (2004) *J Histochem Cytochem* **52**(7), 843-849

49. Nishida, K., Takahara, M., Miyagishima, S.-y., Kuroiwa, H., Matsuzaki, M., and Kuroiwa, T. (2003) *Proc Natl Acad Sci U S A* **100**(4), 2146-2151

50. Coppin, A., Dzierszinski, F., Legrand, S., Mortuaire, M., Ferguson, D., and Tomavo, S. (2003) *Biochimie* **85**(3-4), 353-361

51. Guérardel, Y., Leleu, D., Coppin, A., Liénard, L., Slomianny, C., Strecker, G., Ball, S., and Tomavo, S. (2005) *Microbes Infect* **7**(1), 41-48

52. Dubey, J.P., Lindsay, D.S., and Speer, C.A. (1998) *Clin Microbiol Rev* **11**(2), 267-299

53. Ganesh, S., Agarwala, K. L., Amano, K., Suzuki, T., Delgado-Escueta, A. V., and Yamakawa, K. (2001) *Biochem Biophys Res Commun* **283**(5), 1046-1053

54. Ganesh, S., Tsurutani, N., Suzuki, T., Hoshii, Y., Ishihara, T., Delgado-Escueta, A. V., and Yamakawa, K. (2004) *Biochem Biophys Res Commun* **313**(4), 1101-1109

55. Kissinger, J. C., Gajria, B., Li, L., Paulsen, I. T., and Roos, D. S. (2003) *Nucleic Acids Res* **31**(1), 234-236

56. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) *Nucleic Acids Res* **32**, D138-141

57. Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Yamashita,

R. A., Yin, J. J., Zhang, D., and Bryant, S. H. (2005) *Nucleic Acids Res* **33**, D192-196

58. Viola, R., Nyvall, P., and Pedersén, M. (2001) *Proc Biol Sci* **268**, 1417 - 1422

59. Cavalier-Smith, T. (1999) *J Euk Microbiol* **46**(4), 347-366

60. Cavalier-Smith, T. (1982) *Biol J Linn Soc* **17**, 289-306

61. Bhattacharya, D., and Medlin, L. (1998) *Plant Physiol* **116**, 9-15

62. Cavalier-Smith, T. (2004) *Proc Biol Sci* **271**(1545), 1251-1262

63. Gillott, M. A., and Gibbs, S. P. (1980) *J Phycol* **16**(4), 558-568

64. Embley, T. M., and Martin, W. (2006) *Nature* **440**(7084), 623-630

65. Blennow, A., Nielsen, T. H., Baunsgaard, L., Mikkelsen, R., and Engelsen, S. B. (2002) *Trends Plant Sci* **7**(10), 445-450

66. Smith, A. M., Zeeman, S. C., and Smith, S. M. (2005) *Ann Rev Plant Biol* **56**(1), 73-98

67. Zeeman, S. C., Smith, S. M., and Smith, A. M. (2007) *Biochem J* **401**(1), 13-28

68. Fordham-Skelton, A. P., Chilley, P., Lumbreras, V., Reignoux, S., Fenton, T. R., Dahm, C. C., Pages, M., and Gatehouse, J. A. (2002) *Plant J* **29**(6), 705-715

69. Polekhina, G., Gupta, A., Michell, B. J., van Denderen, B., Murthy, S., Feil, S. C., Jennings, I. G., Campbell, D. J., Witters, L. A., Parker, M. W., Kemp, B. E., and Stapleton, D. (2003) *Curr Biol* **13**(10), 867-871

70. Polekhina, G., Gupta, A., van Denderen, B. J., Feil, S. C., Kemp, B. E., Stapleton, D., and Parker, M. W. (2005) *Structure* **13**(10), 1453-1462

71. Sokolov, L. N., Dominguez-Solis, J. R., Allary, A. L., Buchanan, B. B., and Luan, S. (2006) *Proc Natl Acad Sci U S A* **103**(25), 9732-9737

72. Chung, J.K., Sekiya, F., Kang, H.S., Lee, C., Han, J.S., Kin, S.R., Bae, Y.S., Morris, A.J., and Rhee, S.G. (1997) *J Biol Chem* **272**(25), 15980-15985

73. Maehama, T., and Dixon, J. E. (1998) *J Biol Chem* **273**(22), 13375-13378

74. Guo, S., Stolz, L.E., Lemrow, S.M., and York, J.D. (1999) *J Biol Chem* **274**(19), 12990-12995

75. Hughes, W.E., Cooke, F.T., and Parker, P.J. (2000) *Biochem J* **350**, 337-352

76. Taylor, G.S., Maehama, T., and Dixon, J.E. (2000) *Proc Natl Acad Sci U S A* **97**(16), 8910-8915

77. Robinson, F.L., and Dixon, J.E. (2006) *Trends Cell Biol* **16**(8), 403-412

78. Roach, P. J. (2002) *Curr Mol Med* **2**(2), 101-120

79. Preiss, J., Yung, S. G., and Baecker, P. A. (1983) *Mol Cell Biochem* **57**(1), 61-80

80. Vikso-Nielsen, A., Hao-Jie Chen, P., Larsson, H., Blennow, A., and Moller, B. L. (2002) *Carbohydr Res* **337**(4), 327-333

81. Ritte, G., Lloyd, J. R., Eckermann, N., Rottmann, A., Kossmann, J., and Steup, M. (2002) *Proc Natl Acad Sci U S A* **99**(10), 7166-7171

82. Baunsgaard, L., Lutken, H., Mikkelsen, R., Glaring, M. A., Pham, T. T., and Blennow, A. (2005) *Plant J* **41**(4), 595-605

83. Ritte, G., Heydenreich, M., Mahlow, S., Haebel, S., Kotting, O., and Steup, M. (2006) *FEBS Lett* **580**(20), 4872-4876

84. Yu, T. S., Kofler, H., Hausler, R. E., Hille, D., Flugge, U. I., Zeeman, S. C., Smith, A. M., Kossmann, J., Lloyd, J., Ritte, G., Steup, M., Lue, W. L., Chen, J., and Weber, A. (2001) *Plant Cell* **13**(8), 1907-1918

85. Schnabel, R., and Seitelberger, F. (1968) *Pathol Eur* **3**(2), 218-226

86. Weber, A. P. M., Linka, M., and Bhattacharya, D. (2006) *Eukaryot Cell* **5**(3), 609-612

87. Adl, S.M., Simpson, A.G., Farmer, M.A., Andersen, R.A., Anderson, O.R., Barta, J.R., Bowser, S.S., Brugerolle, G., Fensome, R.A., Fredericq, S., James T.Y., Karpov S., Kugrens P., Krug J., Lane C.E., Lewis L.A., Lodge J., Lynn D.H., Mann D.G., McCourt R.M., Mendoza L., Moestrup O., Mozley-Standridge S.E., Nerad T.A., Shearer C.A., Smirnov A.V., Spiegel F.W., Taylor M.F. (2007) *J Eukaryot Microbiol* **52**(5), 399-451

The text of Appendix A is a reprint of the material as it appears in *Journal of Cell Biology*, 2007, Vol. 178, No. 3, 477-488. Matthew S. Gentry, Robert H. Dowen III, Carolyn A. Worby, Seema Mattoo, Joseph R. Ecker, and Jack E. Dixon. The dissertation author was a major contributing researcher and second author of this paper.

APPENDIX B

Human DNA methylomes at single-base resolution reveal widespread cell-specific epigenetic signatures

ABSTRACT

DNA cytosine methylation is a central epigenetic modification that plays essential roles in cellular processes including genome regulation, development and disease. Despite this, no comprehensive assessment of the precise localization of this modification has been achieved in a mammalian genome. Here we present the first genome-wide, single-base resolution maps of methylated cytosines in human embryonic stem cells and fetal fibroblasts, along with comparative analysis of mRNA and small RNA components of the transcriptome, several histone modifications, and sites of DNA-protein interaction for several key regulatory factors. Widespread differences were identified in the composition and patterning of cytosine methylation between the two genomes. Nearly one-quarter of all methylation identified in embryonic stem cells was present in a non-CG context, suggesting that embryonic stem cells and fibroblasts may utilize different methylation mechanisms to affect gene regulation. Methylation in non-CG contexts showed non-random patterns of localization at functional genomic sequences, including enrichment in gene bodies and depletion in protein binding sites and enhancers. Non-CG methylation disappeared upon induced differentiation of the embryonic stem cells, and was restored in induced pluripotent stem cells. We identified hundreds of differentially methylated regions proximal to genes involved in pluripotency and differentiation, and widespread reduced methylation levels within the fibroblast cells associated with lower transcriptional activity. These reference epigenomes will provide a foundation

upon which future studies can explore the extent of variability of this key epigenetic modification in human disease and development.

INTRODUCTION

Thirty-four years have passed since Riggs, Holliday and Pugh first proposed that cytosine DNA methylation in eukaryotic genomes could act as a stably inherited modification with the potential to affect gene regulation and cellular differentiation (1,2). In the intervening period, intense research effort has expanded our understanding of diverse aspects of DNA methylation in higher eukaryotic organisms. These include elucidation of molecular pathways required for establishing and maintaining DNA methylation, cell-type specific variation in methylation patterns, and the involvement of methylation in multifarious cellular processes such as gene regulation, DNA-protein interactions, cellular differentiation, suppression of transposable elements, embryogenesis, X-inactivation, genomic imprinting and tumorigenesis (3-9). DNA methylation, together with covalent modification of histones, is thought to alter chromatin density and accessibility of the DNA to cellular machinery, thereby modulating the transcriptional potential of the underlying DNA sequence (10). However, while chromatin immunoprecipitation has enabled genome-wide characterization of the location of a number of important histone modifications, no thorough identification of the precise sites of DNA methylation throughout a mammalian genome has yet been reported.

Genome-wide studies of mammalian DNA methylation have previously been conducted, however they have been limited by low resolution of detection (11), sequence-specific bias, or complexity reduction approaches that analyze only a very small fraction of the genome (12-14). In order to improve our understanding of the genome-wide patterns of DNA methylation and its variability between human cell types we have performed massively parallel bisulfite sequencing of unselected DNA fragments. This has produced unbiased, comprehensive maps of the sites of cytosine

DNA methylation at single-base resolution throughout the majority of the human genome in both embryonic stem cells and fibroblasts. Furthermore, we have profiled several important histone modifications, protein-DNA interaction sites of regulatory factors, and the mRNA and small RNA components of the transcriptome to better understand how changes in DNA methylation patterns and histone modifications may affect readout of the proximal genetic information.

EXPERIMENTAL PROCEDURES

Cell Culture

IMR90 cells were obtained from ATCC and cultured under recommended conditions, during which replicate 1 and 2 cells underwent 4 and 5 passages, respectively. H1 and H9 cells were grown in 10cm^2 dishes (approximately 1 x 10^7 cells / dish) in feeder free conditions on Matrigel (BD Biosciences, San Jose, CA) using quality controlled mTeSR1 media for several passages as described previously (15,16), with/without 200 ng/µl BMP4 for 6 days (RND systems, Minneapolis, MN). The cells used for H1 replicate 1 and 2 cells were passage 25 and 27, including the 9 and 5 passages in mTeSR1 media, respectively. H9 cells were passage 42 including several passages in mTeSR1. IMR90 iPS cells were passage 65, with 33 passages in mTeSR1, and prior to cell harvest aliquots of cells were assayed for Oct4 expression by flow cytometry as described previously (15,16). Cells were submitted to the WiCell Cytogenetics Laboratory to confirm normal karyotype.

MethylC-Seq Library Generation

Five µg of genomic DNA was extracted from frozen cell pellets using the DNeasy Mini Kit (Qiagen, Valencia, CA) and spiked with 25 ng unmethylated cI857 *Sam7* Lambda DNA (Promega, Madison, WI). The DNA was fragmented by sonication to 50-500 bp with a Bioruptor (Diagenode, Sparta, NJ), followed by end repair with a nucleotide triphosphate mix free of dCTP. Cytosine-methylated adapters provided by Illumina (Illumina, San Diego, CA) were ligated to the sonicated DNA as per manufacturer's instructions for genomic DNA library construction. Adapter-ligated DNA of 140-210 bp was isolated by 2% agarose gel electrophoresis, and sodium bisulfite conversion was performed on it using the MethylEasy *Xceed* kit (Human

Genetic Signatures, NSW, Australia) as per manufacturer's instructions. One third of the bisulfite-converted, adapter-ligated DNA molecules were enriched by 4 cycles of PCR with the following reaction composition: 2.5 U of uracil-insensitive *PfuTurboC$_x$* Hotstart DNA polymerase (Stratagene), 5 µl 10X *PfuTurbo* reaction buffer, 25 µM dNTPs, 1 µl Primer 1.1, 1 µl Primer 2.1 (50 µl final). The thermocycling parameters were: 95˚C 2 min, 98˚C 30 sec, then 4 cycles of 98˚C 15 sec, 60˚C 30 sec and 72˚C 4 min, ending with one 72˚C 10 min step. The reaction products were purified using the MinElute PCR purification kit (Qiagen, Valencia, CA) then separated by 2% agarose gel electrophoresis and the amplified product was purified from the gel using the MinElute gel purification kit (Qiagen, Valencia, CA). Up to three separate PCR reactions were performed on subsets of the adapter-ligated, bisulfite-converted DNA, yielding up to three independent libraries from the same biological sample. We obtained the final sequence coverage by sequencing all libraries for a sample separately, thus reducing the incidence of "clonal" reads which share the same alignment position and likely originate from the same template molecule in each PCR. Quantitative PCR was used to measure the concentration of viable sequencing template molecules in the library prior to sequencing. The sodium bisulfite non-conversion rate was calculated as the percentage of cytosines sequenced at cytosine reference positions in the Lambda genome.

Small RNA Library Generation

RNA fractions enriched for small RNAs were isolated from cell pellets treated with RNAlater (Life Technologies, Carlsbad, CA) using the mirVana miRNA isolation kit (Life Technologies, Carlsbad, CA) and treated with DNaseI (Qiagen, Valencia, CA) for 30 min at room temperature. Following ethanol precipitation, the small RNAs were separated by electrophoresis on a 15% TBE-urea gel and RNA molecules between approximately 10 and 50 nt were excised and eluted from the gel fragments. Following ethanol precipitation, smRNA-Seq libraries were produced using the Small RNA Sample Prep v1.5 kit (Illumina, San Diego, CA) as per manufacturer's instructions.

Directional RNA-Seq Library Generation

Total RNA was isolated from cell pellets treated with RNAlater using the mirVana miRNA isolation kit and treated with DNaseI (Qiagen, Valencia, CA) for 30 min at room temperature. Following ethanol precipitation, biotinylated LNA oligonucleotide rRNA probes complementary to the 5S, 5.8S, 12S, 18S and 28S ribosomal RNAs were used to deplete the rRNA from 20 µg of total RNA in two sequential RiboMinus reactions (Life Technologies, Carlsbad, CA) as per manufacturer's instructions. Two hundred ng of the remaining RNA was fragmented by metal hydrolysis in 1X fragmentation buffer (Life Technologies, Carlsbad, CA) for 10 min at 94˚C, the reaction was stopped by addition of 2 µl fragmentation stop solution (Life Technologies, Carlsbad, CA). Fragmented RNA was treated with 5 U Antarctic phosphatase (New England Biolabs, Ipswich, MA) for 40 min at 37˚C in the presence of 40 U RNaseOut (Life Technologies, Carlsbad, CA), followed by phosphatase heat inactivation at 65˚C for 5 min. Phosphorylation was performed by addition of 10 U PNK (New England Biolabs, Ipswich, MA), 1 mM ATP, and 20 U RNaseOut and incubation at 37˚C for 1 h. The RNA was purified using 66 µl SPRI beads (Agencourt, Beverly, MA) and eluted in 11 µl 10 mM Tris buffer pH 8.0. One µl of 1:10 diluted adenylated 3' RNA adapter oligonucleotide (5'-UCGUAUGCCGUCUUCUGCUUGidT-3') was added to the phosphorylated RNA and incubated at 70˚C for 2 min followed by placement on ice. The 3’ RNA adapter ligation reaction was performed by addition of 2 µl 10x T4 RNA ligase 2 truncated ligation buffer, 1.6 µl 100 mM $MgCl_2$, 20 U RNaseOut and 300 U T4 RNA ligase 2 truncated (New England Biolabs, Ipswich, MA) and incubation at 22˚C for 1 h. Ligation of the 5’ RNA adapter was performed by addition to the 3’ adapter ligated reaction of 1 µl 1:1 diluted, heat denatured (70˚C 2 min) 5’ RNA adapter oligonucleotide (5’-GUUCAGAGUUCUACAGUCCGACGAUC-3’), 1 µl 10 mM ATP, and 10 U T4 RNA ligase (Promega, Madison, WI), and incubation at 20˚C for 1 h. The RNA was purified using 66 µl SPRI beads (Agencourt, Beverly, MA) and eluted in 10 µl 10 mM Tris buffer pH 8.0. To the RNA ligation products, 2 µl 1:5 diluted RT primer (5’-CAAGCAGAAGACGGCATACGA-3’) was added and heat denatured (70˚C 2 min), followed by incubation on ice. Added to the denatured RNA/primer solution was 4 µl 5x first strand buffer, 1 µl 12.5 mM dNTPs, 2 µl 100 mM DTT, and 40 U RNaseOut, followed by incubation at 48˚C for 1 min. To this, 200 U Superscript II reverse transcriptase (Life Technologies, Carlsbad, CA) was added, followed by incubation at 44˚C for 1 h. The RT reaction was used in a PCR enrichment containing 0.25 µM GEX1 (5’-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3’) and 0.25 µM GEX2 (5’-CAAGCAGAAGACGGCATACGA-3’) primers, 0.25 mM dNTPs, 1x Phusion polymerase buffer and 4 U Phusion hot-start high fidelity DNA polymerase (New England Biolabs, Cambridge, MA) in a 100 µl reaction using the following thermocycling parameters: 98˚C 30 sec, then 15 cycles of 98˚C 10 sec, 60˚C 30 sec and 72˚C 15 sec, ending with one 72˚C 10 min step. The PCR products were purified in two steps, first by purification using 180 µl SPRI beads and elution in 30 µl 10 mM Tris buffer pH 8.0, followed by purification with 39 µl SPRI beads and elution in 10 µl 10 mM Tris buffer pH 8.0. All oligonucleotides were obtained from

Illumina (San Diego, CA). Quantitative PCR was used to measure the concentration of viable sequencing template molecules in the library prior to sequencing.

Chromatin Immunoprecipitation and ChIP-Seq Library Generartion

Chromatin immunoprecipitation (ChIP) for SOX2 (R&D Systems, #AF2018; 5µg) and NANOG (R&D Systems, #AF1997, 5µg) was performed as recently described (Hawkins *et al*., submitted). ChIP for OCT4 (Santa Cruz, #sc8626, 2µg; Santa Cruz, #sc9081, 2µg; R&D Systems, #AF17566, 2µg), p300 (Santa Cruz, #sc585, 5µg), KLF4 (Abcam, #ab21949, 10µg) TAFIIp250/TAF1 (Santa Cruz, #sc735, 5µg) were carried out as previously described with 500µg chromatin and 2-10µg antibody (17,18). ChIP libraries for sequencing were prepared following standard protocols from Illumina (San Diego, CA) with the following minor modifications. Following linker ligation, libraries were run on an 8% acrylamide gel and size selected for 175 - 250bp. This was repeated following PCR amplification. After each size selection, acrylamide was shredded and incubated with 300ul EB buffer (Qiagen, Valencia, CA) overnight at 4ºC or 50 ºC for 20 mins with shaking. DNA was eluted using Nanosep MF filter tubes (Pall, East Hills, NY).

High-throughput Sequencing

MethylC-Seq and RNA-Seq libraries were sequenced using the Illumina Genome Analyzer II (GA II) as per manufacturer's instructions. Sequencing of MethylC-Seq libraries was performed up to 87 cycles to yield longer sequences that are more amenable for unambiguous mapping to the human genome reference sequence. Image analysis and base calling were performed with the standard Illumina pipeline (Firecrest v1.3.4 and Bustard v1.3.4), performing automated matrix and phasing calculations on the PhiX control that was run in the eighth lane of each flowcell.

Validation of Bisulfite Sequencing Results

Primers were designed to amplify a limited number of specific regions of the genome following bisulfite conversion. Genomic DNA was isolated from H1, BMP-treated H1, H9, IMR90 and IMR90 iPS cells, fragmented by sonication, and 1 µg of genomic DNA from each sample was bisulfite converted according to the procedures described above. For each cell type, approximately one tenth of the converted sample was used in 3 distinct PCR reactions (MasterTaq Kit, 5 Prime, Gaithersburg, MD), each containing a different pair of primers designed to amplify a distinct genomic region (Table B.2). Amplified products were separated by gel electrophoresis, gel purified, and cloned using the Zero Blunt TOPO PCR cloning kit (Life Technologies, Carlsbad, CA). Sanger sequencing of multiple clones for each cell type and amplicon was performed to identify the methylation status of cytosines within each region.

Data Analysis

Processing and Alignment of MethylC-Seq Read Sequences

Read sequences produced by the Illumina pipeline in FastQ format were first pre-processed in three steps. Firstly, reads were trimmed to before the first occurrence of a low quality base (PHRED score ≤ 2). Secondly, as a subset of reads contained all or part of the 3' adapter oligonucleotide sequence, every read was searched for the adapter sequence, and if detected the read was trimmed to the preceding base. If the full adapter sequence was not detected, iterative searching of the *k* 3’ terminal bases of the read for the *k* 5’ bases of the adapter was performed, and if detected the read was trimmed to the preceding base. Thirdly, any cytosine base in a read was replaced with thymine. Following pre-processing, reads were sequentially aligned using the Bowtie algorithm (v0.9.9.1; Langmead *et al*. (19)) to two computationally converted NCBI BUILD 36/HG18 reference sequences, the first in which cytosines were replaced with thymines, and the second in which guanines were replaced with adenines. The 48,502 bp Lambda genome was included in the reference sequence as an extra chromosome so that reads originating form the unmethylated control DNA could be aligned. As all cytosines in the reads were replaced with thymines, the methylation status of a particular genomic sequence has no bearing on its ability to map to the reference. Sequences originating from the Watson strand of the genome aligned to the cytosine-free reference sequence, whereas sequences originating from the Crick strand (complement) of the genome aligned to the guanine-free reference sequence after reverse complementation. The following parameters were used in the Bowtie alignment process: --solexa-quals -e 140 -l 20 -n 0 -k 10 --best --nomaqround. For each read, up to 10 of the most highly scoring alignment positions in the reference sequences were returned, tolerating a maximum sum quality score of 140 at mismatch positions. All results of aligning a read to both the Watson and Crick converted genome sequences were combined, and if more than one alignment position existed for a read it was categorized as ambiguously aligned and disregarded. For each cell line, the reads from two biological replicates were pooled to provide greater coverage for identification of the methylcytosines that are presented in this study. Additionally, parallel analysis was performed on each biological replicate to analyse the variability of DNA methylation.

Whole lanes of aligned read sequences were combined in a manner based on the experimental setup. As up to three independent libraries from each biological replicate were sequenced, we first removed reads that shared the same 5’ alignment position within each library, referred to as “clonal” reads, leaving the read at that position that had the highest sum quality score. Subsequently, the reads from all libraries of a particular sample were combined. All unambiguous, or “unique”, read alignments were then subjected to post-processing, which consisted of 3 steps. Firstly, if a read contained more than 3 mismatches compared to the reference sequence, it was trimmed to the base preceding the fourth mismatch. Secondly, the cytosines that were originally removed from the read sequences prior to alignment were incorporated back into the aligned reads. Thirdly, to remove reads that were likely not bisulfite converted, reads that contained more than 3 cytosines in a non-CG context were discarded. Finally, the number of calls for each base at every reference sequence position and on each strand was calculated. Read number for each replicate before and after removal of clonal reads and post-processing is detailed in Table B.1.

Identification of Methylated Cytosines

At each reference cytosine the binomial distribution was used to identify whether at least s subset of the genomes within the sample were methylated, using a 0.01 FDR corrected P-value. Each context of methylation was considered independently: CG, CHG or CHH (where H = A, C or T). We identified methylcytosines while keeping the number of false positives methylcytosine calls below 1% of the total number of methylcytosines we identified. The probability *p* in the binomial distribution B(n,*p*) was estimated from the number of cytosine bases sequenced in reference cytosine positions in the unmethylated Lambda genome (referred to as the error rate: non-conversion plus sequencing error frequency). The bisulfite conversion rates for all samples were over 99%, and the error rates were as follows: H1 replicate 1, 0.007; H1 replicate 2, 0.004; H1 combined replicates, 0.0050; IMR90 replicate 1, 0.002; IMR90 replicate 2, 0.003; IMR90 combined replicates, 0.0024. We interrogated the sequenced bases at each reference cytosine position one at a time, where read depth refers to the number of reads covering that position. For each position, the number of trials (n) in the binomial distribution was the read depth. For each possible value of n we calculated the number of cytosines sequenced (k) at which the probability of sequencing k cytosines out of n trials with an error rate of *p* was less than the value M, where M * (number of unmethylated cytosines) < 0.01 * (number of methylated cytosines). In this way, we established the minimum threshold number of cytosines sequenced at each reference cytosine position at which the position could be called as methylated, so that out of all methylcytosines identified no more than 1% would be due to the error rate.

Correction of DNA Methylation Context Calls Proximal to SNPs

As the cell lines studied have distinct genotypes compared to the Human reference sequence, the sequencing data downstream of every site of non-CG methylation was interrogated to determine whether the cytosine in the H1 and IMR90 cell lines was truly in the non-CG context. If the consensus call at the base downstream (+1) of a non-CG methylcytosine was a guanine, the methylcytosine context was corrected to mCG. Furthermore, the context of any methylcytosine that had been identified on the opposite strand to the +1 guanine was subsequently corrected to mCG. At positions where +1 bases were potentially heterozygous for a

SNP, two conditional tests were performed on the surrounding sequence to test for any evidence that the site represented a CG dinucleotide. Firstly, when there was sequence coverage on the opposite strand, if the +1 position displayed at least 20% guanine and on the opposite strand displayed at least 20% cytosine, the methylcytosine context was corrected to mCG. Furthermore, a methylcytosine was added on the opposite strand at this site if the base calls at the position passed the binomial test to the same significance threshold as used in the initial methylcytosine calling. Secondly, if the strand opposite the +1 position had no sequence coverage and the +1 position displayed a similar number of guanine base calls as the cytosine calls at the methylcytosine, the methylation context was corrected to mCG.

Identification of Differentially Methylated Cytosines

For each cell type the DNA methylation data is comprised of the combination of MethylC-Seq performed on two biological replicates of different passage number. To compare the mCG overlap between the two biological replicates for H1 and IMR90 cells, the mCGs from the binomial distribution analysis from each replicate were selected and the read coverage for each replicate was determined at each position. To compare only mCG that possess similar sequence read coverage, a ratio of the coverage between replicates was calculated and only positions with a depth ratio between 0.8 and 1.2 were considered for the Venn diagram analysis. The mCHGs and mCHHs for the H1 biological replicates were compared in an identical fashion.

A two-tailed Fisher’s Exact Test was used to identify cytosines that are differentially methylated between the H1 and IMR90 cell types. Only mCGs determined using the binomial distribution analysis in at least one cell type and those mCG covered by at least 3 reads in at least one cell type were considered for testing.

P-value thresholds were selected such that the number of false positives is less than 5% of all mCG positions called as significantly different (5% FDR). A total of 6,023,738 mCG where identified as more highly methylated in H1 cells (*p*-value < 0.007433) and 124,161 mCG where identified as more highly methylated in IMR90 cells (*p*-value < 0.000153).

mCHG and mCHH Enriched Genes

Density of methylated or all occurrences of CHG and CHH in 10Kb regions throughout the genome was determined. The hypergeometric distribution was used to determine the enrichment of methylated occurrences in comparison to the total number of sites in a given window, taking into account the total number of methylated and total occurrences across the whole chromosome. Windows with a over-representation P-value less then 1e-20 were considered and Ref Seq whose TSS is within 10Kb from the center of each window were selected.

Genome Annotation

Genomic regions were defined based on NCBI BUILD 36/HG18 coordinates downloaded from UCSC web site. Promoters are arbitrarily defined as regions 2Kb upstream the TSS for each Ref Seq transcript. According to the UCSC annotation many Ref Seq transcripts can be associated with a given gene, and they can have the same or alternative TSS. Gene bodies are defined as the transcribed regions, from the start to the end of transcription sites for each Ref Seq. In case of genomic regions with strand information, those on the reverse strand were reversed. Consequently, mean methylation profiles over all the occurrences of a genomic region on the genome are oriented from 5' to 3'.

mC and mC/C Methylation Profiles

Genomic regions were divided in 20 uniformly sized bins. In particular, for genomic regions in genes, the 20 bins span from the 5' to the 3' end. Rather, for genomic regions centered at annotated genomic elements or obtained by ChIP-Seq experiments, an arbitrarily sized window was centered at the center of each genomic element or ChIP-Seq peak, as indicated in the figures or figure legends. All occurrences of genomic regions were checked for having sufficient coverage in H1 and IMR90 methylomes. Regions with more than one quarter insufficiently covered (less than a total of 3 reads in both strands) are masked. For regions centered at annotated features the same criteria were applied to check the coverage in the central 10% of the region. Masked genomic regions were not used in the determination of the mean profile.

Absolute (mC) methylation content was determined for each bin based on the number of calls of a given methylation type (mCG, mCHG or mCHH) divided by the bin size. For the symmetric mCG, sites where methylation is observed in at least one strand were counted, while for mCHG and mCHH this measure is determined as the sum of methylation calls of a given type on both strands. Relative methylation content (mC/C) was determined as the absolute methylation content divided by the total number of sites of the same type on the genome independently from their methylation level. In particular, for mCG the total number of CG sites was determined only for one strand, as there is a correspondent number if the same sites on the opposite strand. Rather, for mCHG and mCHH, the total number of CHG or CHH occurrences on the genome was determined. The total number of sites was again normalized by the bin size. Analysis of NCBI BUILD 36/HG18 genome reference sequence was performed using R and Bioconductor tools and annotation libraries (www.r-project.org, www.bioconductor.org; Gentleman *et al*. (20)).

Identification of Differentially Methylated Regions (DMRs)

A sliding window approach was used to find regions of the genome enriched for sites of higher levels of DNA methylation in IMR90 relative to H1, as identified by Fisher's Exact Test. A window size of 1 kb was used, progressing 100 bp per iteration. When a 1 kb window containing at least 4 differential mCG was identified, the region was extended in 1 kb increments until a 1 kb increment was reached that contained less than 4 differential mCG. After extension termination, a region containing at least 10 differential mCG and at least 2 kb in length were reported as a DMRs.

Identification of Partially Methylated Domains (PMDs)

A sliding window approach was used to find regions of the genome in IMR90 that were partially methylated, based on the measurements of the level of methylation at each mCG. A window size of 10 kb was used, progressing 10 kb per iteration. When a 10 kb window was identified that contained at least 10 mCG, each covered by at least 5 MethylC-Seq reads, for which the average methylation level of these mCG was less than 70%, the region was extended in 10 kb increments. Extension was terminated when a 10 kb increment was reached that had an average methylation level of greater than 70% or less than 10 mCG, and the region was reported as a PMD.

Mapping smRNA-Seq Reads

smRNA sequence reads in FastQ format were produced by the Illumina analysis pipeline. smRNA-Seq reads that contained at least 5 bases of the 3’ adapter sequence were selected and this adapter sequence removed, retaining the trimmed reads that were from 16 to 37 nt in length. These processed reads in FastQ format were aligned to the human reference genome (NCBI BUILD 36/HG18) with the Bowtie alignment algorithm using the following parameters: --solexa-quals -e 1 -l 20 -n 0 -a -m 1000 --best --nomaqround. Consequently, any read that aligned with no mismatches to the and to no more than 1000 locations in the NCBI BUILD 36/HG18 reference genome sequence was retained for downstream analysis.

Identification of smRNA Clusters

A sliding window approach was used to find regions of the genome in that displayed dense clusters of smRNAs. A window size of 1 kb was used, progressing 200 bp per iteration. When a 1 kb window was identified that contained more than 10 non-redundant smRNA reads the region was extended in 500 bp increments until a 500 bp increment was reached that contained less than 3 non-redundant smRNA reads. After extension termination, a region containing at least 100 smRNA and at least 3 kb in length were reported as a smRNA cluster.

Mapping RNA-Seq Reads

Read sequences produced by the Illumina analysis pipeline were aligned with the ELAND algorithm to the NCBI BUILD 36/HG18 reference sequence and a set of splice junction sequences generated from known splice junctions in the UCSC Known Genes. Reads that aligned to multiple positions were discarded. Reads per kilobase of transcript per million reads (RPKM) were calculated with the CASAVA software package.

Mapping and Enrichment Analysis of ChIP-Seq Reads

Following sequencing cluster imaging, base calling and mapping were conducted using the Illumina pipeline. Clonal reads were removed from the total mapped tags, retaining only the monoclonal unique tags that mapped to one location in the genome, where each sequence is represented once. Regions of tag enrichment were identified as recently described (Hawkins *et al.*, submitted).

Data Visualization in the AnnoJ Browser

MethylC-Seq, RNA-Seq, ChIP-Seq and smRNA-Seq sequencing reads and positions of methylcytosines with respect to the NCBI BUILD 36/HG18 reference sequence, gene models and functional genomic elements were visualized in the AnnoJ 2.0 browser, as described previously (21). The data mentioned above can be viewed in the AnnoJ browser at: http://neomorph.salk.edu/human_methylome.

RESULTS

Single-Base Resolution Maps of DNA Methylation for Two Human Cell Lines

The first single-base DNA methylomes of a higher eukaryote, the flowering plant *Arabidopsis thaliana*, were produced using a method referred to as MethylC-Seq (21) or BS-Seq (22). In this method, genomic DNA is treated with sodium bisulfite (BS) to convert cytosine, but not methylcytosine, to uracil, and subsequent high-throughput sequencing. We performed MethylC-Seq for two human cell lines, H1 human embryonic stem cells (23) and IMR90 fetal lung fibroblasts (24). We sequenced reads of up to 87 bases in length on the Illumina Genome Analyzer II to generate 1.16 and 1.18 billion reads, respectively, that aligned uniquely to the human reference sequence (NCBI build 36/HG18). The total sequence yield was 87.5 and 91.0 gigabases (Gb), with an average read depth of 14.2X and 14.8X per strand of the 3.08 Gb human genome sequence, for H1 and IMR90 respectively (Figure B.1.A). The MethylC-Seq data covers most of the human genome in each cell type; over 86% of both strands of the 3.08 Gb human reference sequence are covered by at least one sequence read (Figure B.1.B), accounting for 94% of the cytosines in the genome.

At each cytosine in the genome the binomial distribution was used to identify whether methylation is detectable at a significant level, at a 1% false discovery rate (see Experimental Procedures). We detected approximately 62 million and 45 million methylcytosines in H1 and IMR90 cells, respectively (Figure B.2.A), comprising 5.83% and 4.25% of the cytosines with sequence coverage. Full browsing of the entire dataset at single base resolution can be performed at http://neomorph.salk.edu/human_methylome using the AnnoJ browser (www.annoj.org). Of the methylcytosines detected in the IMR90 genome, 99.98%

were in the CG context, and the total number of mCG sites is very similar in both cell types. In the H1 stem cells we detected abundant DNA methylation in non-CG contexts (mCHG and mCHH, where H = A, C or T), comprising almost 25% of all cytosines at which DNA methylation is identified, and accounting for most of the difference in total methylcytosine number between the cell types (Figure B.2.A). The prevailing assumption is that mammalian DNA methylation is located almost exclusively in the CG context, however a handful of studies have previously detected non-CG methylation in human cells, and in particular in embryonic stem cells (25,26). Bisulfite-PCR, cloning and sequencing of selected loci displaying H1 non-CG methylation in several human cell lines revealed that a second embryonic stem cell line, H9 (23), displayed mCHG and mCHH at conserved positions, confirming that non-CG methylation is likely a general feature of human ES cells (Figure B.3, Table B.2). In addition, like IMR90 cells, BMP4-induced H1 cells lost non-CG methylation at several loci examined while methylation in the CG context was maintained, indicating that the pervasive non-CG methylation is lost upon differentiation. Furthermore, analysis of these loci in IMR90 induced pluripotent stem (iPS) cells revealed restored non-CG methylation (Figure B.3). Overall this demonstrates that the CHG and CHH methylation identified in H1 and absent in IMR90 are not simply due to genetic differences between the two cell types, but rather the presence of non-CG methylation is characteristic of an embryonic stem cell state. For each cell type, two biological replicates were performed with cells of different passage number (see Experimental Procedures), and comparison of the methylcytosines identified independently in each replicate revealed a high concordance of cytosine methylation status between replicates (Figure B.4). Exemplifying both cell-specific differential methylation and the presence of non-CG methylation is the *OCT4* gene. In the H1 genome, *OCT4* contains non-CG methylation and has a 5’ domain that is largely free of mCG, while the gene is more extensively CG methylated in IMR90 (Figure B.2.B), with a concomitant ~50-fold reduction in *OCT4* transcript levels (data not shown). The absence of mCHG and mCHH methylation in IMR90 coincided with significantly lower transcript abundance of the *de novo* DNA methyltransferases (DNMTs) *DNMT3A* and *DNMT3B* and the associated *DNMT3L* in IMR90 cells (Figure B.5), which is supported by a previous study of DNA methylation in ES cells and somatic cells (25) and by the determined target sequence specificity of these DNMTs (27,28).

Multiple reads covering each methylcytosine can be used as a readout of the fraction of the sequences within the sample that are methylated at that site (22), here referred to as the methylation level of a specific cytosine. We surveyed all methylated sites that were covered by more than 10 MethylC-Seq reads to profile the methylation level of each of these cytosines in the genome. In the H1 genome we observed that 77% of mCG sites were 80-100% methylated, whereas 85% of mCHG and mCHH sites were only 10-40% methylated (Figure B.2.C). This profile closely resembles the mCG and mCHH methylation pattern in the *Arabidopsis* genome (21), and indicates that at sites of non-CG methylation only a fraction of the surveyed genomes in the sample are methylated. Notably, only 56% of mCG sites in IMR90 were highly methylated (80-100%), with the remainder distributed over a wide range of methylation levels (Figure B.2.C), indicating that although the total number of mCG sites in H1 and IMR90 is similar, in general the IMR90 mCG sites were typically less frequently methylated. This is supported by measurement of how frequently a cytosine was observed in BS modified DNA in all instances at which a CG site was sequenced: 82.7% and 67.7% of all sequenced CG sites were methylated in H1 and IMR90, respectively.

A global-scale view of DNA methylation levels revealed that the density of DNA methylation shows large variations throughout each chromosome (Figure B.2.D). Sub-telomeric regions of the chromosomes frequently showed higher DNA methylation density (Figure B.2.D, Figure B.6), which was previously reported as important for control of telomere length and recombination (29,30). The smoothed profile of DNA methylation density in 100 kb windows shows that on the chromosomal level the density profile of mCG in H1 and IMR90 is similar. The density profiles of mCHG and mCHH reveal that non-CG methylation is present throughout the entire chromosome, and notably that changes in density of the non-CG methylation are distinct from that of mCG in a number of regions.

Pervasive Non-CG DNA Methylation in Embryonic Stem Cells

In order to further characterize the abundant non-CG methylation throughout the H1 genome, we first compared the average density of methylation in each sequence context relative to the underlying density of all potential sites of methylation in each context (henceforth referred to as the relative methylation density), throughout various genomic features (Figure B.7.A, Figure B.8). As expected, we observed a correlation in the density of mCG and the distance from the transcriptional start site (TSS), with mCG density increasing in the 5’ UTR to a similar level in exons, introns and the 3’ UTR as to 2 kb upstream of the TSS (Figure B.7.A). In agreement with previous studies, we generally observed lower relative densities of methylation at CG islands and TSS, however a subset of these regions did not display this depletion (Figure B.9; Deng *et al*. (13), Meissner *et al*. (14), Brunner *et al*. (31)). mCHG and mCHH methylation densities also decrease significantly toward the TSS and return to the same level as 2 kb upstream at the end of the 5’ UTR, however within exons, introns and 3’ UTRs the non-CG methylation densities are twice as high. Intriguingly, the mCHH density is approximately 15-20% higher in exons than within introns and the 3’ UTR. To determine whether there was a relationship between gene activity and non-CG methylation level within the gene body we performed strand-specific RNA-Seq (21). We observed a positive correlation between gene expression and gene body mCHG ($R = 0.60$) or mCHH ($R = 0.58$) density (Figure B.7.B), with highly expressed genes containing approximately 3-fold higher non-CG methylation density than non-expressed genes (Figure B.10.A).

We identified 447 and 226 genes that are proximal to genomic regions highly enriched for mCHG and mCHH, respectively, with 180 genes in common. An example of non-CG methylation enrichment in such a gene, *Splicing Factor 1*, is shown in Figure B.7.C. Analysis of gene ontology terms for each set revealed significant enrichment for genes involved in RNA processing, RNA splicing, and RNA metabolic processes ($P < 2 \times 10^{-11}$, Figure B.10.B). Unexpectedly, we found a significant enrichment of non-CG methylation on the anti-sense strand of gene bodies, for both mCHG and mCHH enriched sets of genes ($P < 0.1$ and $P < 0.001$, respectively, Figure B.7.D). The anti-sense strand serves as the template for RNA polymerization, and further investigation will be required to determine if there are functional repercussions of this non-CG methylation strand bias. We also observed that genes in H1 had significantly more RNA originating from introns than in IMR90, relative to the total number of sequenced reads in each sample, and this discrepancy in intronic read abundance was significantly enhanced in the mCHG and mCHH enriched genes ($P < 0.001$, Figure B.7.E).

In the *Arabidopsis* genome, the methylation state of a cytosine in the CG and CHG contexts is highly correlated with the methylation of the cytosine on the opposite strand within the symmetrical site (21,22). While we observed that 99% of mCG sites from the human cell lines were methylated on both strands, surprisingly mCHG was highly asymmetrical, with 98% of mCHG sites being methylated on only one strand. This raises an interesting question as to how these sites of DNA methylation are consistently methylated in a considerable fraction of the genomes without two hemi-methylated CHG sites as templates for faithful propagation of the methylation state (Figure B.2.C). It is not yet known whether continual, but indiscriminate, *de novo* methyltransferase activity preferentially methylates particular CHG sites after replication, or if a persistent targeting signal is present that drives CHG methylation.

We analyzed the genome sequence proximal to sites of non-CG methylation to determine whether enrichment of particular local sequences were evident, as previously reported in the *Arabidopsis* DNA methylomes (21,22). Whereas no local sequence enrichment is observed for mCG sites, a preference for the TA dinucleotide upstream of non-CG methylation was observed (Figure B.7.F). Furthermore, the base following a non-CG methylcytosine is most commonly an A, with a T also observed relatively frequently, a sequence preference observed in previous *in vitro* studies of the mammalian DNMT3 methyltransferases (27,28). These local sequence enrichments were not evident when all cytosines were analyzed, regardless of their methylation status, and the level of methylation at a non-CG methylation site did not appear to influence the local sequence enrichment (Figure B.11).

To determine whether there was any preference for the distance between adjacent sites of DNA methylation in the human genome, we analyzed the relative distance between methylcytosines in each context within 50 nucleotides in introns. We focused on introns because these are genomic regions enriched in non-CG methylation, but unlike exons, are not constrained by protein coding selective pressures (Figure B.7.G and Figure B.7.H). Analyses for random genomic sequences and exons are presented in Figure B.12 together with mCG spacing patterns. For methylcytosines in all contexts, a periodicity of 8 bases is evident (Figure B.7.G, Figure B.7.H, Figure B.12), but interestingly a strong tendency is observed for two pairs of 8-base separated mCHG sites spaced with 13 bases between them. A 10 base periodicity is also evident for mCHH sites, corresponding to a single turn of the DNA helix, as previously observed in the *Arabidopsis* genome (22), indicating that the molecular mechanisms governing *de novo* methylation at CHH sites may be common between the plant and animal kingdoms. A structural study of the mammalian *de novo* methyltransferase DNMT3A and its partner protein DNMT3L found that 2 copies of each form a heterotetramer that contains two active sites separated by the length of 8-10 nucleotides in a DNA helix (32,33). The consistent 8-10 nucleotide spacing we observe in the human genome suggests that DNMT3A may be responsible for catalysing the methylation at non-CG sites. Notably, the mCHG and mCHH relative spacing patterns are distinct, suggesting that this sub-categorization of the non-CG methylation is appropriate, and that distinct pathways may be responsible for the deposition of mCHG and mCHH in the human genome.

DNA Methylation is Depleted at DNA-Protein Interaction Sites

Numerous past studies have documented that DNA methylation can alter the ability of some DNA binding proteins to interact with their target sequences (34-38). In order to further investigate this relationship we used ChIP-Seq (39) to identify sites of protein-DNA interaction in H1 cells for a set of proteins important for gene expression in the pluripotent state, namely NANOG, SOX2, KLF4, and OCT4, as well as proteins involved in the transcription initiation complex and in enhancers (TAF1 and p300, respectively; data not shown). In general, we observed a decrease in the profile of relative methylation density toward the site of interaction, particularly in the non-CG context, independently from proximity to the TSS (Figure B.13.A and Figure B.14). The IMR90 genome shows lower average density of methylation at H1 SOX2 and p300 interaction sites, but has similar CG methylation densities for the H1 NANOG and OCT4 and interaction sites, even though the genes encoding these proteins are transcribed at a very low level in IMR90 relative to H1 (47 - 50 fold less mRNA), and are not considered to be functional in fibroblasts. This suggests that these genomic regions are generally maintained in a less methylated state in multiple cell types regardless of the occupancy of DNA binding proteins, though this does not exclude the possibility that other DNA binding proteins are still present at these sites.

We next analyzed the patterns of DNA methylation in sets of enhancers either unique to each cell type or shared. ChIP-Seq was utilized to detect the location of enhancers throughout the H1 and IMR90 genomes, defined as regions of simultaneous enrichment of the histone modifications H3K4me1 and H3K27ac (Heintzman *et al*. (40), Figure B.13.B). We examined the average relative DNA methylation density in each context and found that at IMR90-specific enhancers the CG methylation density was lower than flanking regions of the genome, whereas at these same genomic locations in the H1 genome we observed an increase in mCG density and there was no change in non-CG methylation density (Figure B.13.B). In contrast, at H1-specific enhancers there was no change in mCG density in either the H1 or IMR90 genome, but non-CG methylation density decreased approximately 3-fold at the enhancer sites, relative to the density 5 kb up- and downstream. The set of enhancer sites present in both H1 and IMR90 cells showed both of these cell-specific patterns: lower mCG density in IMR90 and lower non-CG methylation density in H1.

The specific depletion of DNA methylation at active enhancers in each cell type indicates maintenance of these elements in an unmethylated state, potentially preventing interference in the process of protein-DNA interactions at these sites. Consistent with this, a recent study of DNA methylation in 12 Mb of the genome of human T-cells identified some lineage specific differentially methylated regions coincident with enhancers, and *in vitro* methylation-dependent inhibition of the enhancer activity (41). Notably, H1 cells have depleted non-CG methylation but not mCG, in contrast to the mCG depletion at IMR90 enhancers, possibly indicating cell-type specific utilization of different categories of DNA methylation.

Widespread Cell-Specific Patterns of DNA Methylation

The paradigm of DNA methylation controlling aspects of cellular differentiation necessitates that patterns of methylation vary in different cell types. Numerous studies have previously documented differences in DNA methylation between cell types and disease states (7,8,10,42). With comprehensive maps of DNA methylation throughout the genomes of the two distinct cell types, we next characterized changes in DNA methylation evident between the H1 and IMR90 DNA methylomes, and explored how these changes may relate to the distinctiveness of these cells.

The Pearson correlation coefficient of the mCG methylation state between H1 and IMR90 was calculated for 20 equally sized windows flanking or within various genomic features (Figure B.15.A). This provided a measure of the conservation of methylation states in a given location within or around a genomic feature between H1 and IMR90, and is distinct from a comparison of the average relative density of DNA methylation, as presented above (Figure B.13). An increased and high mCG correlation level was observed in correspondence to genomic regions expected to display a more constitutive epigenetic state, such as CG islands and TSS. We also observed a greater correlation at translational start sites and splice junctions. Gene promoters displayed an increase in correlation as the distance from the TSS decreased. Surprisingly, we observed that the correlation in introns is highest toward the 5’ exon-intron junction and decreased throughout the length of the introns. Notably, at the sites of protein-DNA interaction surveyed in Figure B.13.A, we observed a decrease in the correlation of methylation compared to the flanking 1.5 kb of the genome, except for KLF4 (data not shown). This decrease was most pronounced at the predicted site of protein-DNA interaction, indicating that even though the mCG depletion is a general feature at protein binding sites, when a pairwise comparison of the methylation status at each site between H1 and IMR90 is performed a significant decrease in the conservation of methylation is observed.

Surprisingly, we found that a large proportion of the IMR90 genome displayed lower levels of CG methylation than H1 (Figure B.2.C). Contiguous regions with an average methylation level less than 70% were identified (mean length = 153 kb), which we termed partially methylated domains (PMDs; Figure B.16). The PMDs comprise a large proportion of every autosome (average = 38.4%), and 80% of the IMR90 X chromosome (Figure B.17), consistent with the lower levels of DNA methylation reported in the inactive X chromosome (43). As IMR90 cells are derived from a female (XX), it is anticipated that simultaneous sequencing of BS-converted genomic DNA from both the inactive and the active X chromosomes will manifest as partial methylation throughout the majority of the X chromosome. However, the widespread prevalence of PMDs on the autosomes was unexpected. To explore whether the decrease in DNA methylation in IMR90 autosomes was a stochastic event on both parental chromosomes, or whether it may be indicative of differential methylation on each allele, we analyzed the ratio of methylated to unmethylated CG sites within individual MethylC-Seq reads (Figure B.15.B). Considering only reads with at least 2 CG sites, 73% and 57% of the reads were fully methylated at CG sites, in H1 and IMR90 respectively, again indicating the lower overall levels of CG methylation in IMR90. In IMR90 chromosome X PMDs a similar pattern was observed in the methylation status of the H1 MethylC-Seq reads, while in IMR90 28% of reads were completely unmethylated and only 37% of reads were fully methylated (Figure B.15.B). A very similar pattern was observed at IMR90 PMDs throughout the autosomes, raising the intriguing possibility that large tracts of the autosomes could be affected by a mode of repression similar to X-chromosome inactivation in these differentiated cells.

Upon inspection of 5,644 genes with a TSS located in or within 10 kb of a PMD, we found a strong enrichment for these genes to be less expressed in IMR90 ($P = 2 \times 10^{-47}$, Fisher's Exact Test). Specifically, of all of the genes that are more highly expressed in H1 (H1 transcript abundance at least 3 fold higher than IMR90), 42% were located within PMDs, compared to only 13% of all more highly expressed genes in IMR90 cells being located in PMDs (Figure B.15.C and Figure B.16). We observed that many of the partially methylated and down-regulated genes in IMR90 displayed lower proximal H3K4me3 and H3K36me3 modifications, and higher proximal H3K27me3 levels (Figure B.18, Hawkins *et al.*, submitted). While in IMR90 cells we observed a positive correlation between the mean gene body mCG methylation level and gene expression, no such relationship was discernible in H1 cells (Figure B.15.D). Consequently, the positive correlation between gene expression and gene body methylation recently reported (12) could be re-interpreted as a depletion of methylation in repressed genes in differentiated cells.

Epigenetic Regulation of Genes and Endogenous Retroviruses

A sliding window approach was used to identify differentially methylated regions (DMRs) enriched for cytosines where IMR90 is more highly methylated than H1 (5% FDR, Fisher's Exact Test), exemplified in Figure B.19. We identified 491 DMRs, and in a window spanning 20 kb upstream to 20 kb downstream of each DMR we surveyed mCG density, mRNAs, smRNAs, H3K4me3, H3K36me3, H3K27me3 (for histone modifications see Hawkins *et al.*, submitted), genes, and repetitive elements (Figure B.20.A, data not shown). The DMRs were associated with 139 genes more highly expressed in H1 and 113 up-regulated in IMR90. More than half of these genes were associated with DMRs located within 2 kb upstream of the TSS or the 5' UTR, which include factors previously defined as playing a role in embryonic stem cell function (International Stem Cell Initiative (44); Figure B.21).

Complete linkage hierarchical clustering of these data revealed two broad categories of transcriptional activity, histone modifications and DNA methylation proximal to the DMRs (Figure B.20.A). Group 1 DMRs are associated with high proximal H3K4me3, H3K36me3, and transcriptional activity relative to IMR90, and are unmarked by H3K27me3 in both cell types. Closer inspection revealed that a subset of the group 1 DMRs are located at dense clusters of small RNAs that map to annotated Human Endogenous Retroviruses (HERVs; Villesen *et al*. (45)). In the H1 transcriptome 1,184 dense clusters of small RNAs were identified (see Experimental Procedures), 85% of which were within 1 kb of an annotated HERV. Notably, these clusters were effectively absent in IMR90. Of the 61 H1 small RNA clusters coincident with DMRs, 93% were located in HERVs, and nearly all displayed proximal H3K4me3 and downstream H3K36me3, indicative of being actively transcribed (Figure B.20.B).

This accounts for 35% of the 164 HERVs that are coincident with smRNAs and show the active marks H3K4me3 and H3K36me3 (Figure B.22). We observed abundant expression downstream of the HERVs, with a number of transcripts displaying similarity to endonuclease reverse transcriptases. The group 2 DMRs are associated with gene-rich sequences that are more highly expressed in IMR90 cells and generally display a depletion of LINEs in the flanking sequence, with concomitant H3K27me3 modification and less DNA methylation, as observed in many IMR90 PMDs. Furthermore, group 2 regions in H1 frequently display both H3K4me3 and H3K27me3 modifications, characteristic of the bivalent state that are thought to instil a suppressed but poised transcriptional status (46,47). Many of these regions show markedly less H3K27me3 in IMR90 as well as more DNA methylation, suggesting that prior repression may be relieved, and defining a set of genes potentially regulated by DNA methylation and involved in the developmental transition from a pluripotent to differentiated state.

DISCUSSION

We found extensive differences between the DNA methylomes of two human cell types, revealing the highly dynamic nature of this epigenetic modification. With the rapidly decreasing cost of DNA sequencing it will be feasible to expand the high-resolution approach presented in this study to a multitude of methylomes. We have demonstrated that analysis of DNA methylation at genomic positions distal to promoters and CG islands is not only justified, but also essential for understanding its full regulatory capabilities within a genome. Single-base resolution analysis allows unprecedented insight into the establishment and maintenance of DNA methylation. Study of potential fine-scale interactions of methylation with protein binding events has been largely unexplored, but is now possible with this detailed knowledge of the precise positions of methylation. To this regard, we confirmed that depletion of DNA methylation is a common feature at protein binding sites but that the extent of this pattern can be specific for a given protein and even independent of its expression.

The genomic context of the DNA methylation is resolved, here revealing abundant methylation in the non-CG context, which is typically overlooked in alternative methodologies. Profiling of enhancers and different patterning of CG and non-CG methylation in gene bodies and their different correlation with gene expression suggest possible alternative roles for DNA methylation in these two contexts. The exclusivity of non-CG methylation in stem cells, likely maintained by continual *de novo* methyltransferase activity and not observed in somatic tissues, suggests that it may possibly play a key role in the origin and maintenance of this pluripotent state. Essential future studies will need to explore the prevalence of non-CG methylation in diverse cell types, including the temporal variation throughout differentiation, and its potential reestablishment in induced pluripotent states.

Considering the recent discovery of the 6th base, 5-hydroxymethylcytosine (48,49), which can be detected by bisulfite sequencing, it is exciting to speculate whether particular functional sequences of the genome may be marked by this modification to affect a unique mode of regulation. Finally, future profiling of DNA methylomes throughout the temporal, positional and developmental space of the human organism will be key to unravelling the full variability and functionality of this heritable modification.

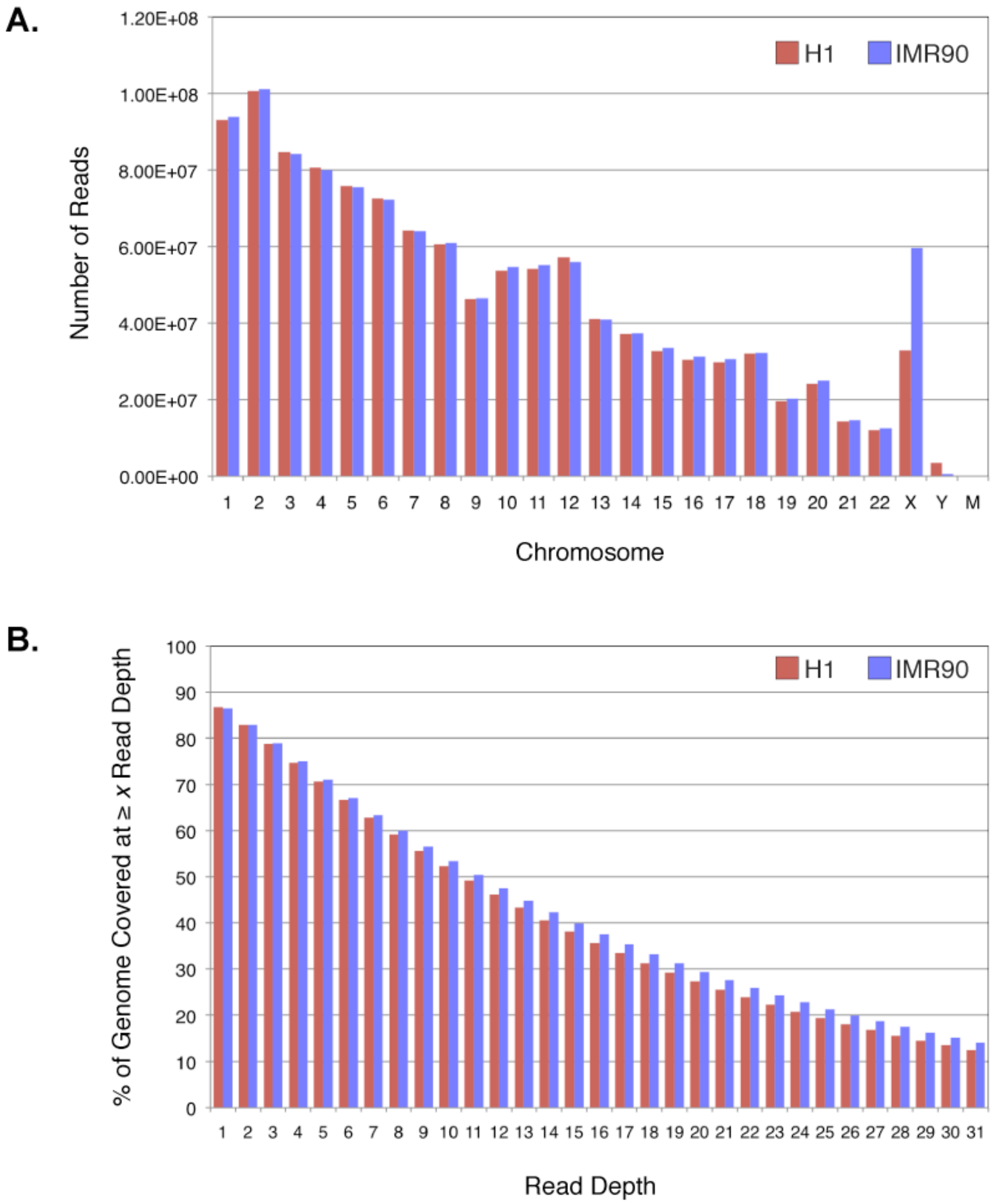


Figure B.1 Uniquely mapped reads and coverage for MethylC-Seq.

(A) The number of uniquely mapped MethylC-Seq reads for each chromosome for H1 and IMR90 cells.
(B) The percent of the H1 and IMR90 genomes that are covered by a differing minimum number of MethylC-Seq reads.

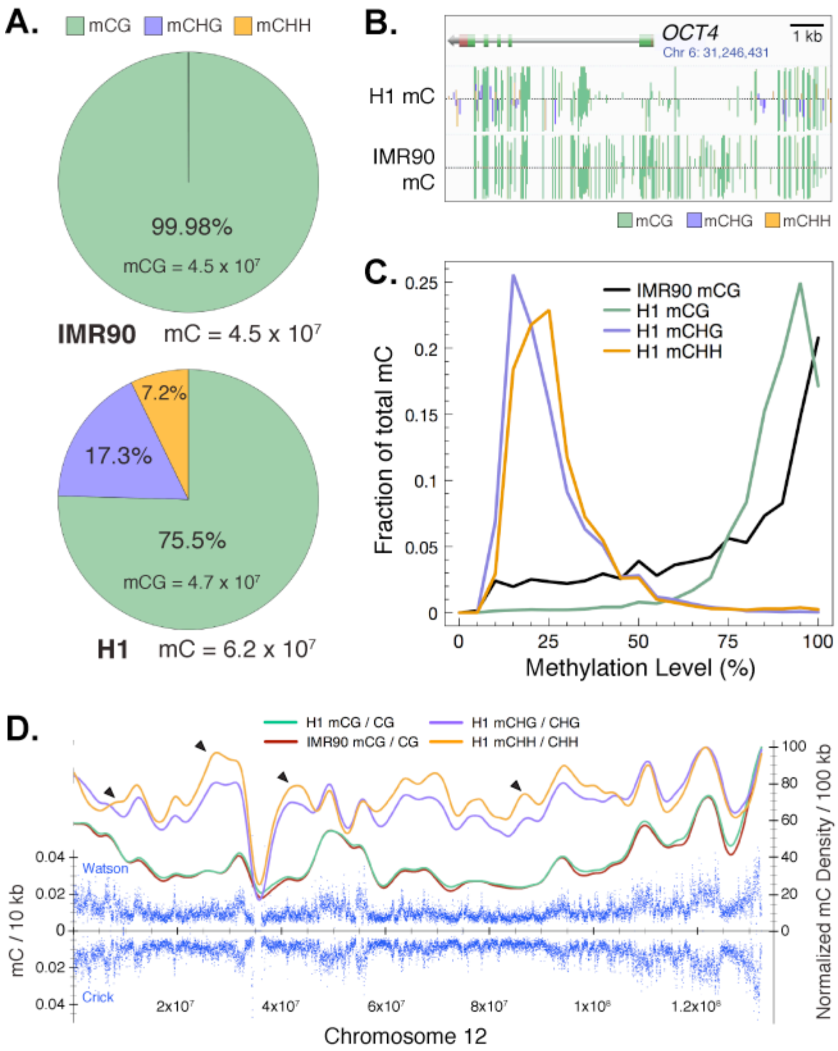


Figure B.2 Global trends of Human DNA methylomes.

(A) The percent of methylcytosines identified for H1 and IMR90 cells in each sequence context (CG, CHG, CHH, where H = A, C, or T).
(B) AnnoJ data browser representation of the germ-line specific gene *OCT4*, which is demethylated in H1 cells but methylated in IMR90 cells.
(C) Distribution of the methylation level in each sequence context. The y-axis indicates the fraction of the total methylcytosines that display each methylation level (x-axis), where methylation level was determined as the fraction of reads at a reference cytosine containing cytosines following bisulfite conversion, and requiring more than 10 reads total.
(D) The density of methylcytosines identified in chromosome 12. Blue dots indicate the density of all methylcytosines H1 cells in 10 kb windows. Smoothed lines represent the density of methylcytosines in each context in H1 and IMR90 cells. Black triangles indicate various regions in which trends in CG and non-CG methylation density varies.

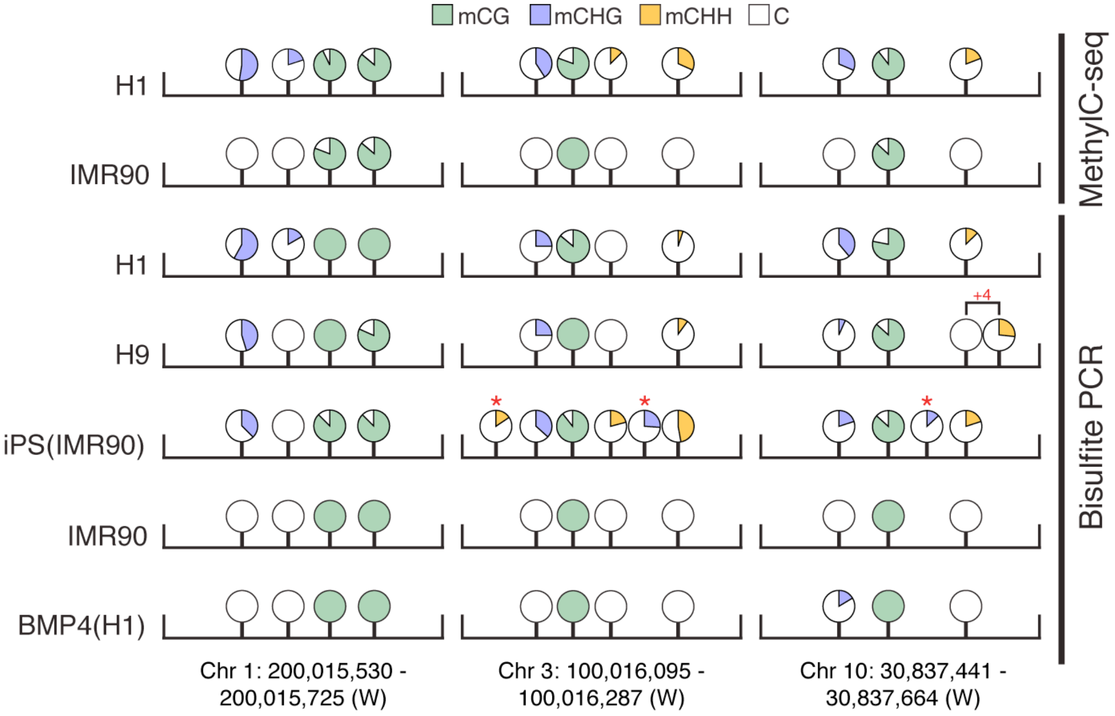


Figure B.3 Bisulfite-PCR validation of non-CG DNA methylation at selected loci within H1, H9, iPS (IMR90) stem cells or differentiated BMP4-treated H1 and IMR90 cells.

Bisulfite converted genomic DNA was amplified by PCR at the indicated genomic regions, cloned and analysed by Sanger sequencing. The sequence context of the DNA methylation site is displayed according to the key and the percent methylation at each position is represented by the fill of each circle (see Table B.2 for values). Non-CG methylated positions indicated by an asterisk are unique to that cell type and “+4” indicates a mCHH that is shifted 4 bases downstream in H9 cells.

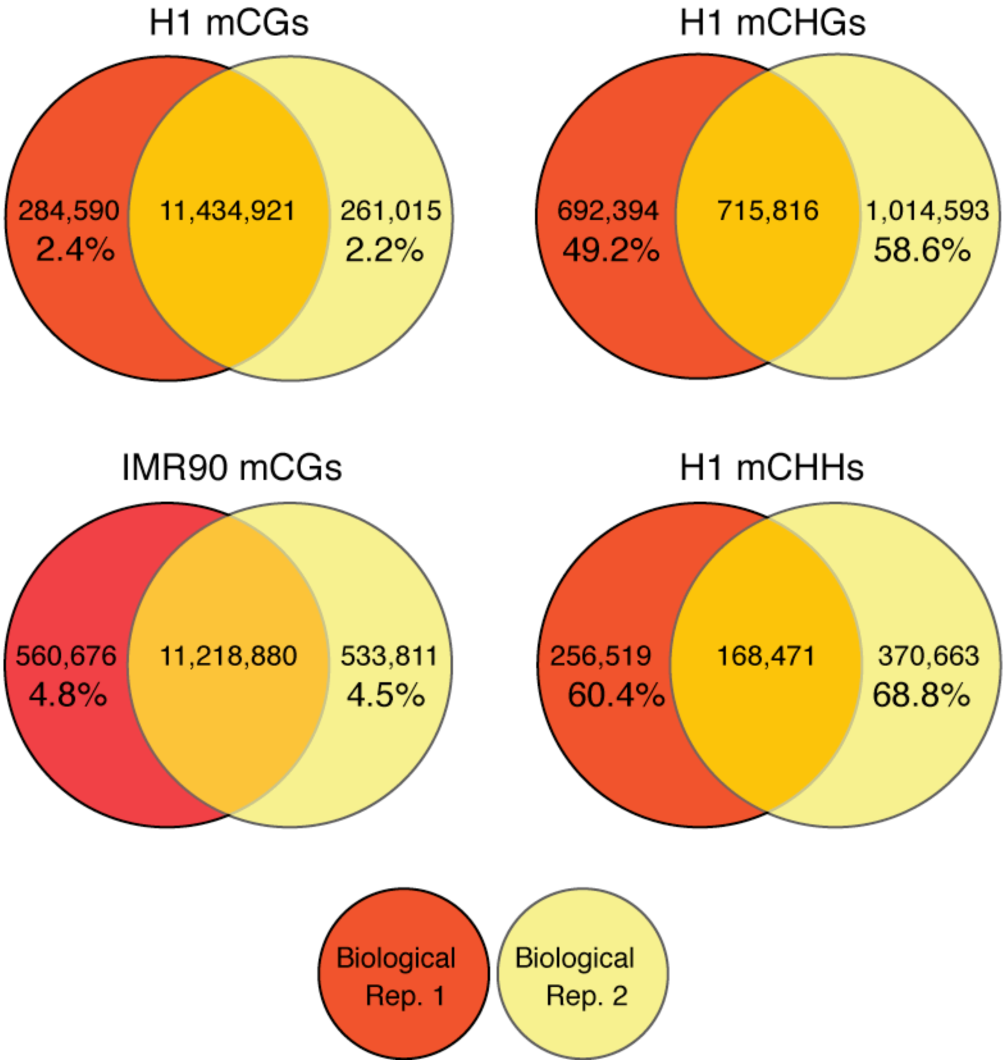


Figure B.4 Venn diagrams representing the overlap in methylcytosines between the two biological replicates of H1 and IMR90 cell types.

Methylcytosines with similar sequencing depth were compared and classified as unique to biological replicate 1 (red), unique to replicate 2 (yellow) or common to both replicates (orange). The number of methylcytosines in each category is listed, as well as the percent methylcytosines unique with each biological replicate.

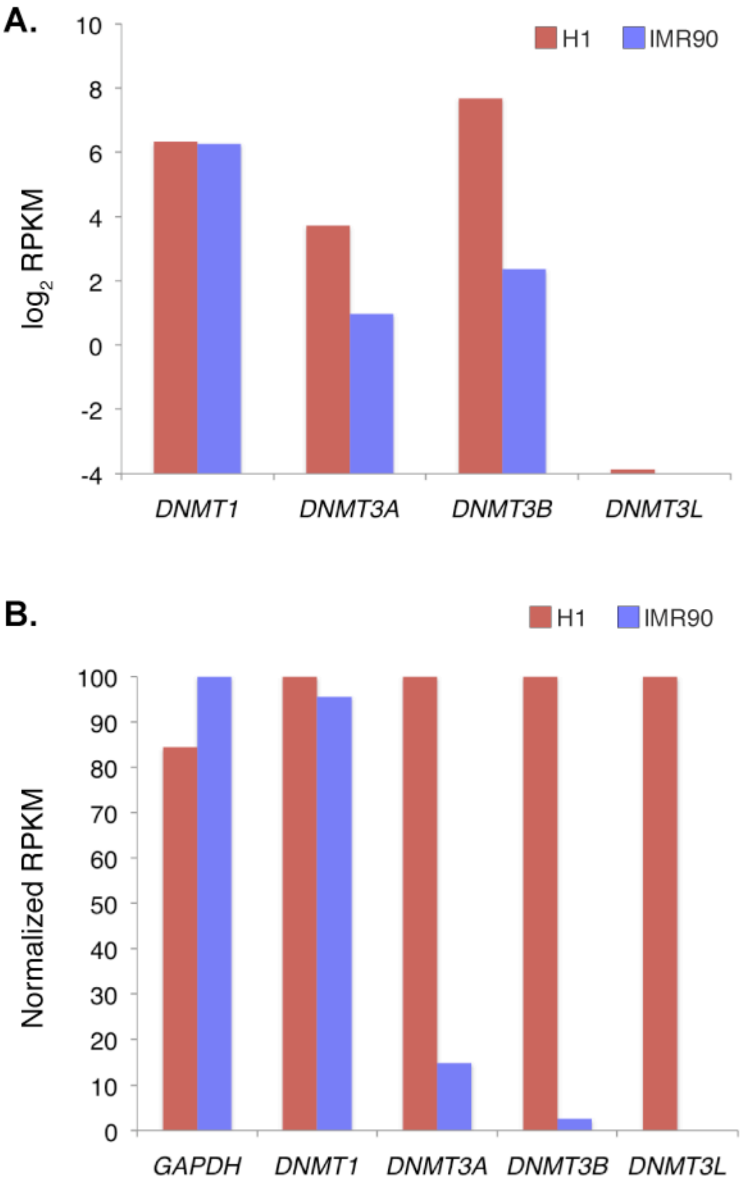


Figure B.5 Differential expression of *DNMT* genes in H1 and IMR90 cells.

(A) $\log_2$RPKM measurements of transcript abundance for *DNMT1*, *DNMT3A*, *DNMT3B*, and *DNMT3L* from RNA-Seq.
(B) Maximum normalized RPKM measurements of transcript abundance for *GAPDH*, *DNMT1*, *DNMT3A*, *DNMT3B*, and *DNMT3L* from RNA-Seq.

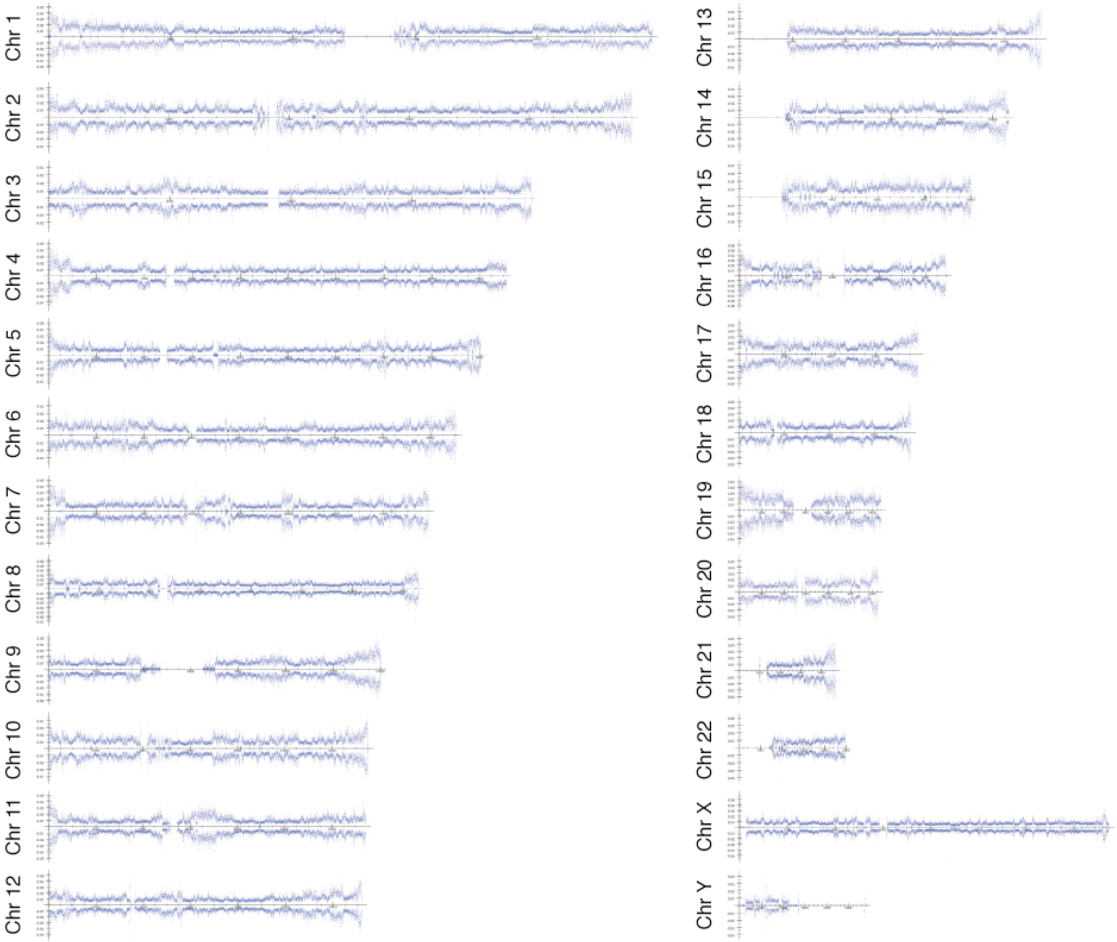
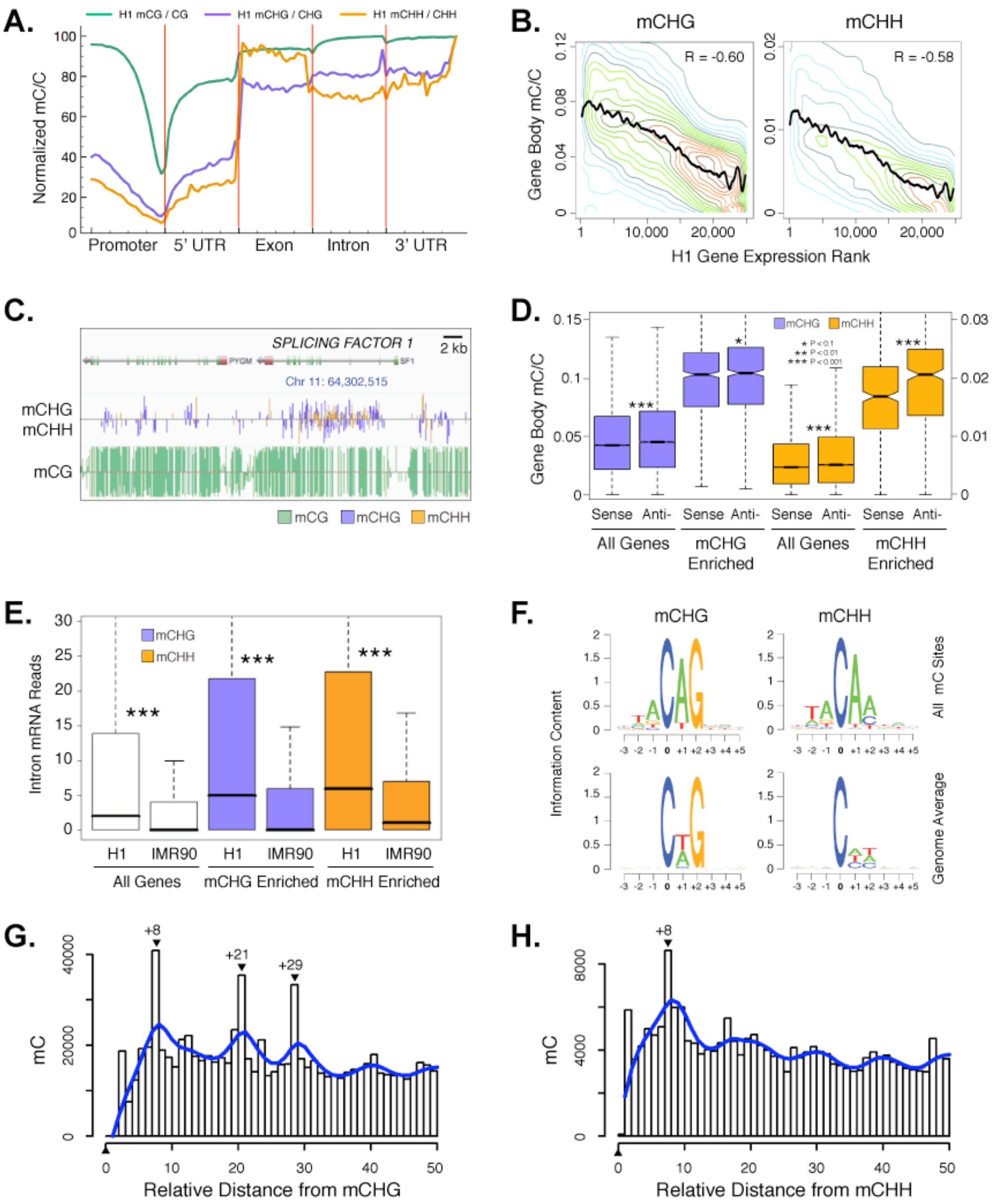


Figure B.6 The density of methylcytosines identified in all chromosomes in H1 cells. Blue dots indicate the density of all methylcytosines in 10 kb windows.

Figure B.7 Non-CG DNA methylation in H1 embryonic stem cells.

(A) The relative methylation density (the ratio of methylcytosines to all potential sites of cytosine methylation) in H1 in each sequence context is represented throughout different gene-associated regions: promoters (2 kb upstream of the transcriptional start site), 5’ UTRs, exons, introns and 3’ UTRs. Each region was divided into 20 equally sized windows and for each sequence context the mean mC/C profile was normalized to the maximum value.
(B) Relative methylation density within gene bodies (y-axis) as a function of gene expression (x-axis), with increasing transcript abundance from right to left. Colored lines represent density of the underlying data points and smoothing with cubic splines is displayed in black.
(C) Graphical representation of sites of CG and non-CG methylation at a non-CG methylation enriched gene, *Splicing Factor 1*.
(D) The average relative methylation densities in each sequence context within gene bodies on the sense or anti-sense strand relative to gene directionality. P-values for differences between sense and antisense densities are indicated.
(E) Number of mRNA intronic reads in all genes or genes associated with non-CG enriched regions, in H1 and IMR90.
(F) Logo plots of the sequences proximal to sites of non-CG DNA methylation in each sequence context in H1 cells. Three bases flanking every site of methylation were analysed to identify local sequence preferences. The information content of each base represents the level of sequence enrichment.
(G) Prevalence of mCHG sites (y-axis) as a function of the number of bases between adjacent mCHG sites (x-axis) based on all non redundant pair-wise distances up to 50 nt in all introns. The blue line represents smoothing with cubic splines.
(H) Prevalence of mCHH sites (y-axis) as a function of the number of bases between adjacent mCHH sites (x-axis) based on all non redundant pair-wise distances up to 50 nt in all introns. The blue line represents smoothing with cubic splines.


A.
H1 mCG / CG
H1 mCHG / CHG
H1 mCHH / CHH
Normalized mC/C
100
80
60
40
20
0
Promoter
5' UTR
Exon
Intron
3' UTR
B.
mCHG
mCHH
R = -0.60
R = -0.58
Gene Body mC/C
0.12
0.08
0.04
0
0.02
0.01
0
1
10,000
20,000
H1 Gene Expression Rank
C.
SPLICING FACTOR 1
2 kb
PYGM
SF1
Chr 11: 64,302,515
mCHG
mCHH
mCG
mCG
mCHG
mCHH
D.
Gene Body mC/C
0.15
0.1
0.05
0
0.03
0.02
0.01
0
mCHG
mCHH
* P < 0.1
** P < 0.01
*** P < 0.001
***
*
***
***
Sense
Anti-
All Genes
mCHG Enriched
mCHH Enriched
E.
Intron mRNA Reads
30
25
20
15
10
5
0
mCHG
mCHH
***
***
***
H1
IMR90
All Genes
mCHG Enriched
mCHH Enriched
F.
mCHG
mCHH
Information Content
2
1.5
1
0.5
0
-3 -2 -1 0 +1 +2 +3 +4 +5
All mC Sites
Genome Average
G.
+8
+21
+29
mC
40000
20000
0
10
20
30
40
50
Relative Distance from mCHG
H.
+8
mC
8000
4000
0
10
20
30
40
50
Relative Distance from mCHH

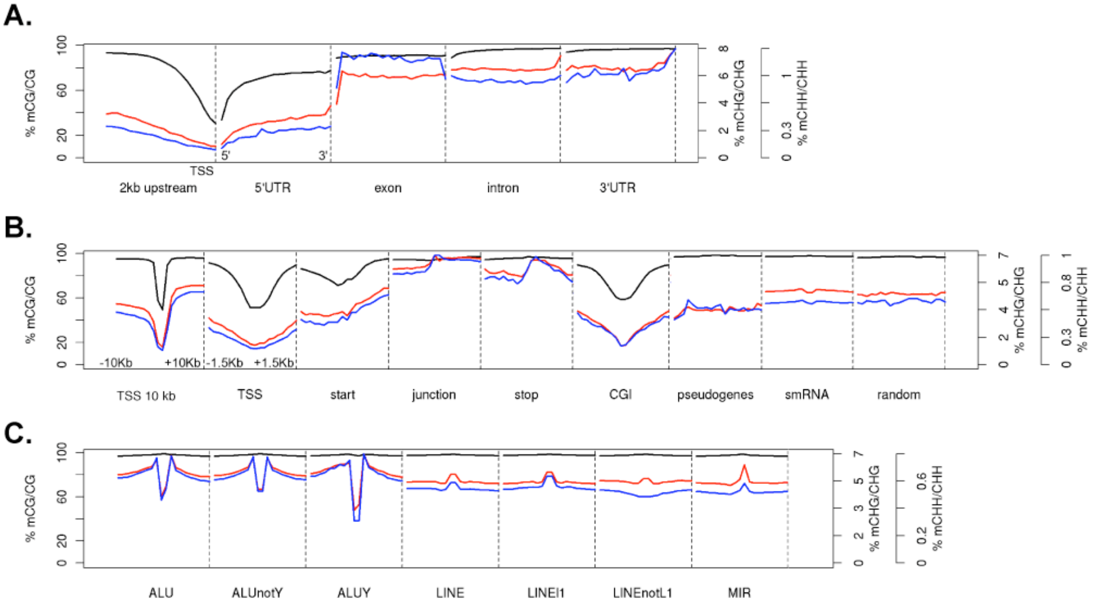


Figure B.8 Mean mC/C profiles over genomic regions.

(A) Gene body regions were divided in 20 bins from the 5’ to 3’ end, and the mean mC/C level within each bin for each methylation type was determined (mCG/CG, black; mCHG/CHG, red; mCHH/CHH, blue).
(B) Gene regulatory elements were analysed as in (A).
(C) Transposable repeated genomic regions were analysed as in (A).

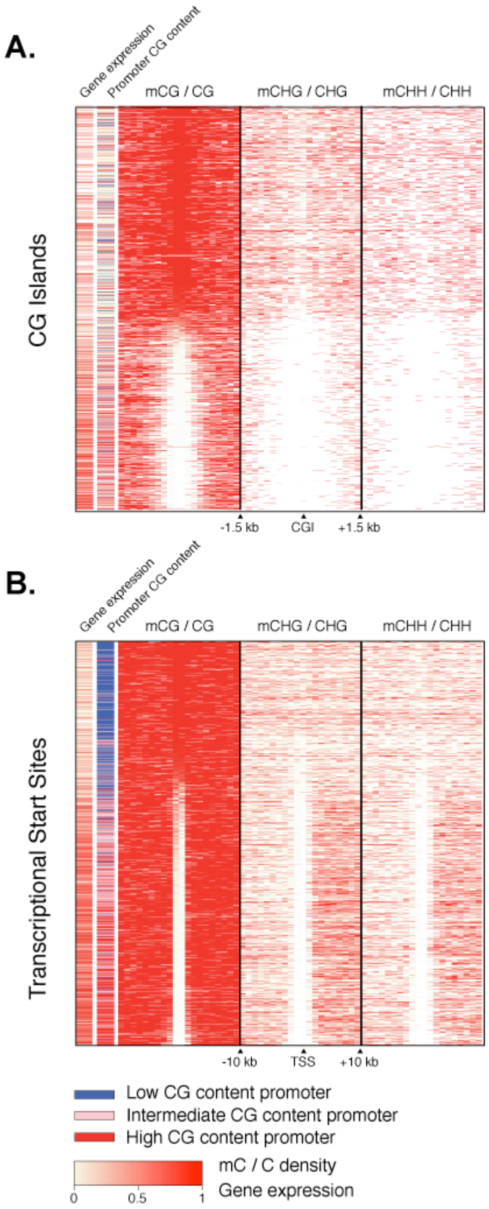


Figure B.9 DNA Methylation at CG islands and transcriptional start sites.

(A) Relative DNA methylation density at CG islands (1.5 kb upstream/downstream) is displayed with downstream gene expression and promoter CG content. Each CG island was assigned to the closet gene whose transcriptional start site is within 10 kb. As expected, low CG content promoters are highly methylated, or close to highly methylated CG islands, and lie close to low expressed genes. However, high CG content promoters are poorly methylated and usually lie close to highly expressed genes.

(B) Transcriptional start sites were analysed as in (A).

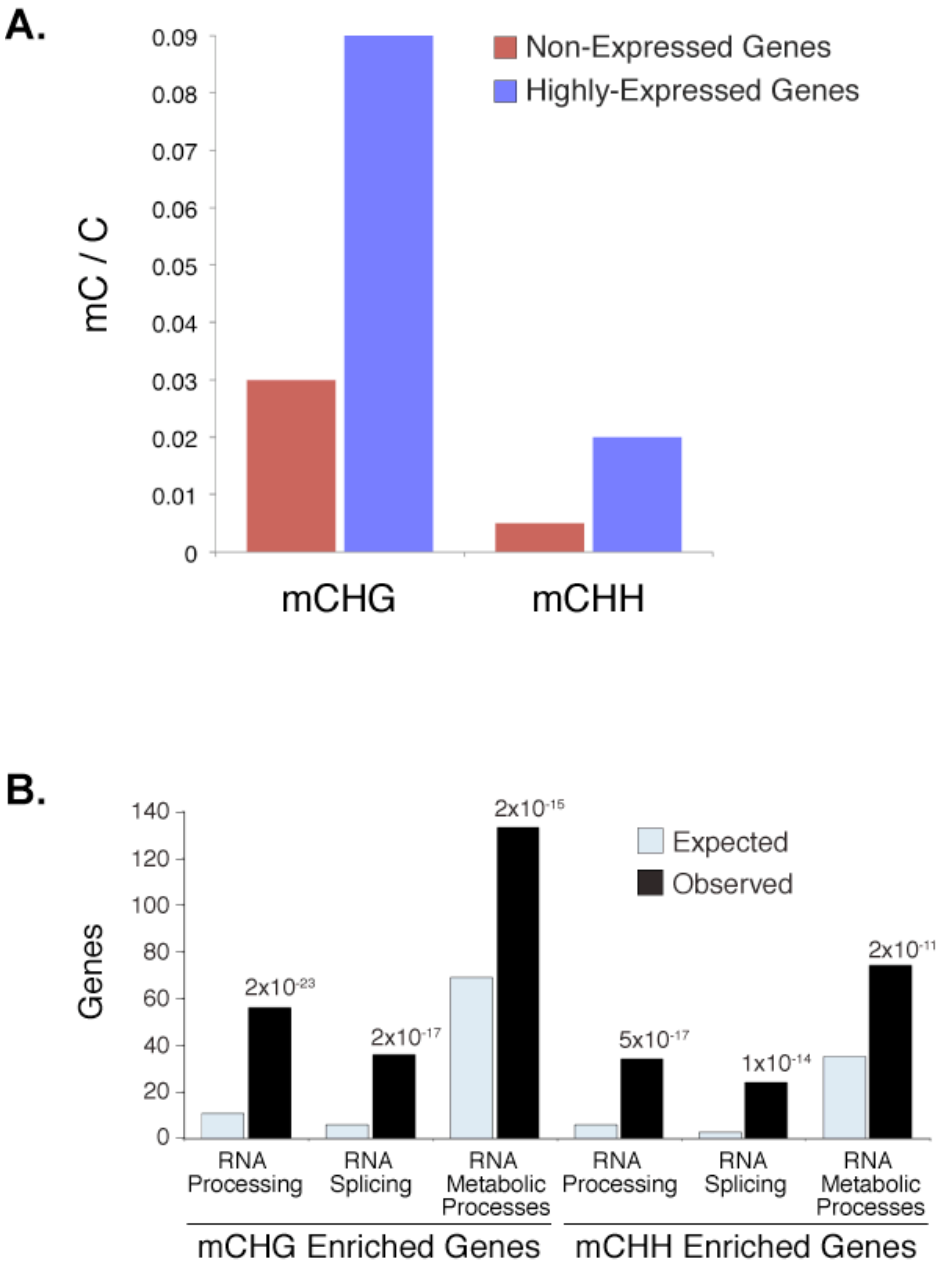


Figure B.10 Functional analysis of non-CG gene body methyaltion reveals a enrichment for highly-expressed genes.

(A) Enrichment of non-CG methylation in non-expressed and highly-expressed genes in H1 cells.
(B) Over-representation of GO terms of genes within 20 kb of genomic regions displaying the highest enrichment of CHG and CHH methylation. The enrichment P-value is shown for each GO term.

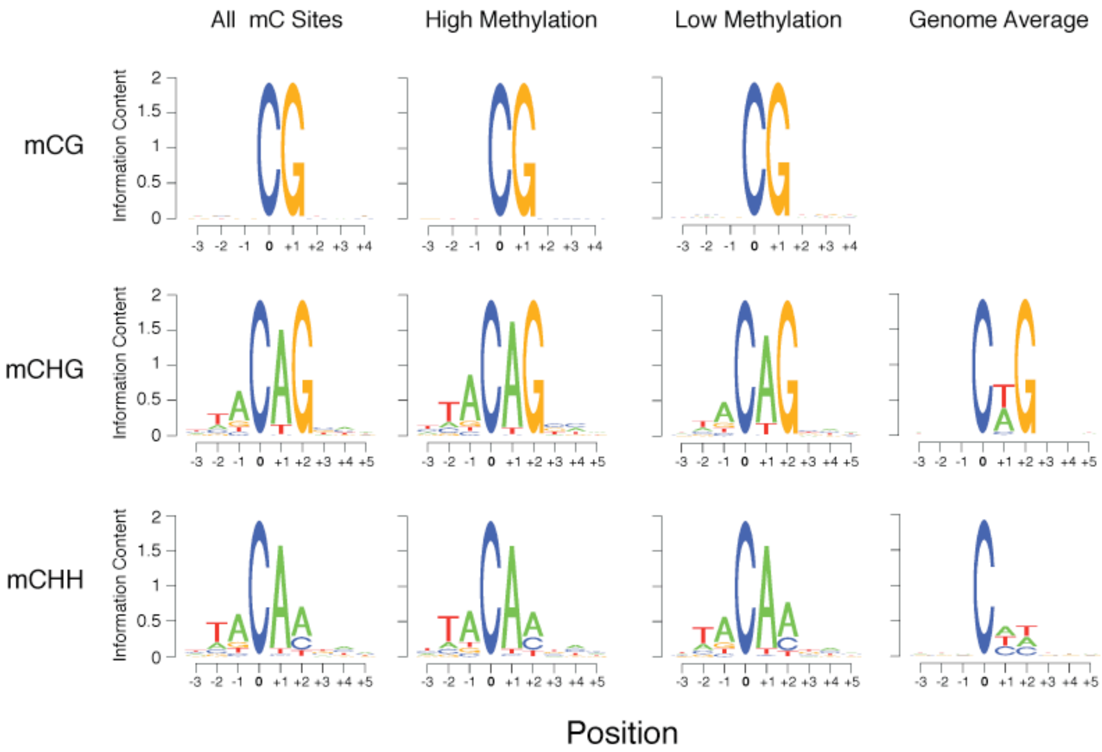


Figure B.11 Logo plots of the sequences proximal to sites of DNA methylation in each sequence context in H1 cells.

Logo plots are presented for all methylcytosines, and methylcytosines that display a high methylation level (CG >75% methylated, non-CG >25% methylated), or low methylation level (CG <75% methylated, non-CG <25% methylated). Three bases flanking every site of methylation were analysed to identify local sequence preferences. The information content of each base represents the level of sequence enrichment.

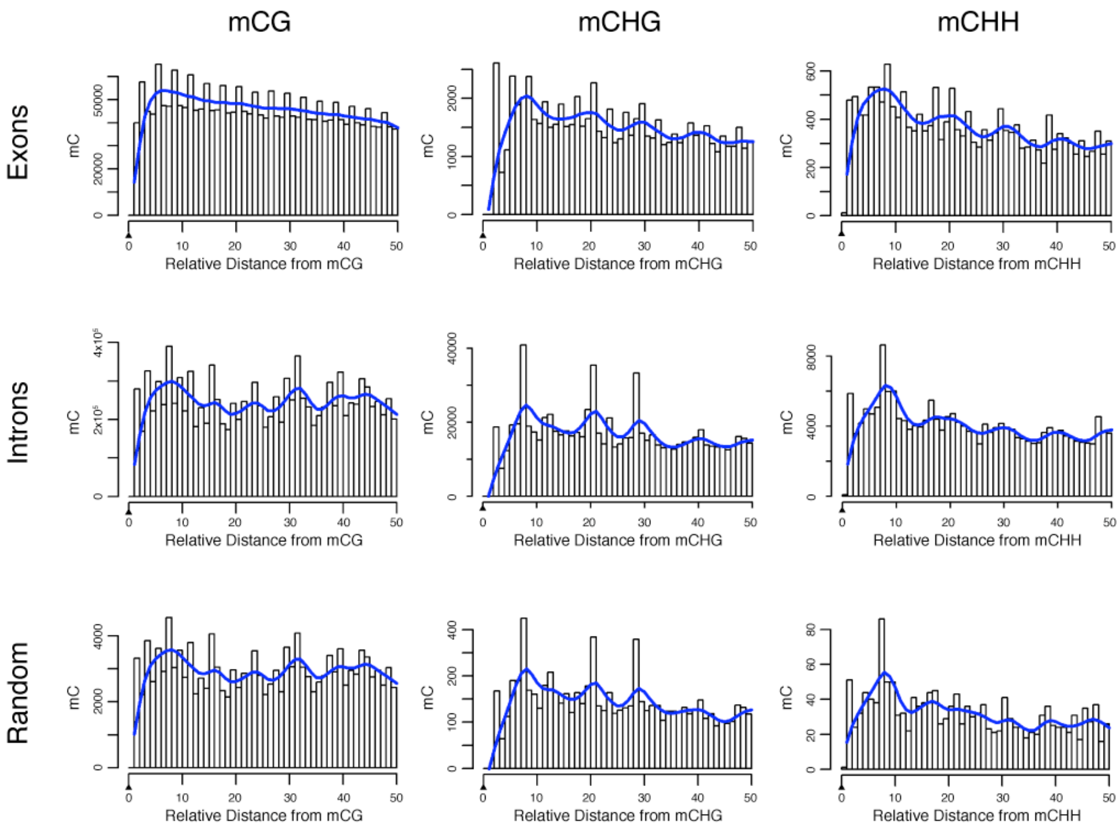


Figure B.12 Spacing of adjacent methylcytosines in different contexts.

Prevalence of mCG/mCHG/mCHH sites (y-axis) as a function of the number of bases between adjacent mCG/mCHG/mCHH sites (x-axis) based on all non redundant pair-wise distances up to 50 nucleotides in exons, introns, and random sequences. The blue lines represent smoothing by cubic splines

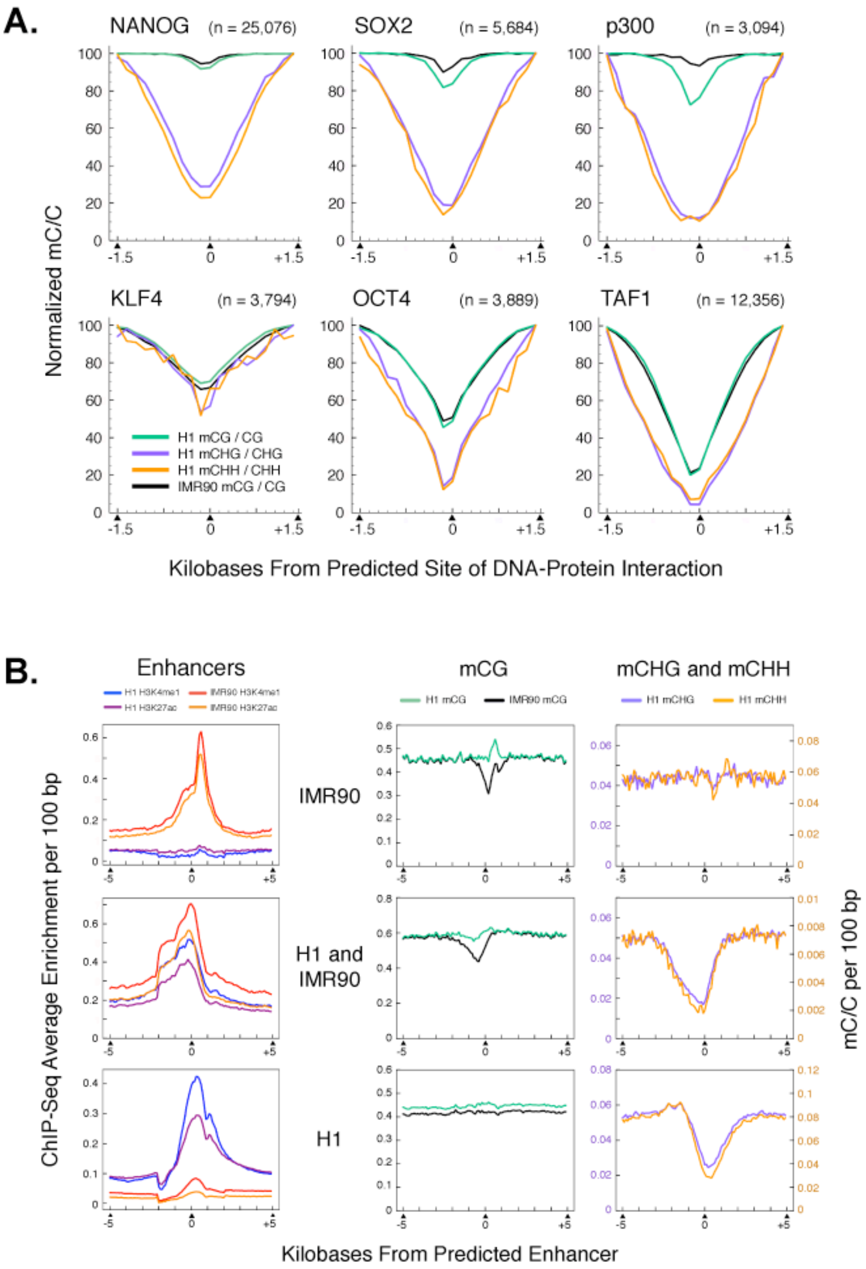


Figure B.13 Density of DNA methylation at sites of DNA-protein interaction.

(A) The average relative DNA methylation densities in each sequence context are shown from 1.5 kb upstream to 1.5 kb downstream of the predicted sites of DNA-protein interaction identified by ChIP-Seq.

(B) Co-localization of H3K4me1 and H3K27ac ChIP-Seq read enrichment indicative of enhancer sites that have been grouped into 3 sets: specific to H1 cells (top), IMR90 cells (bottom), or common to both H1 and IMR90 cells (middle). The average relative DNA methylation densities in each sequence context in 100 bp windows are displayed throughout 5 kb upstream to 5 kb downstream of the sites in each of the sets.

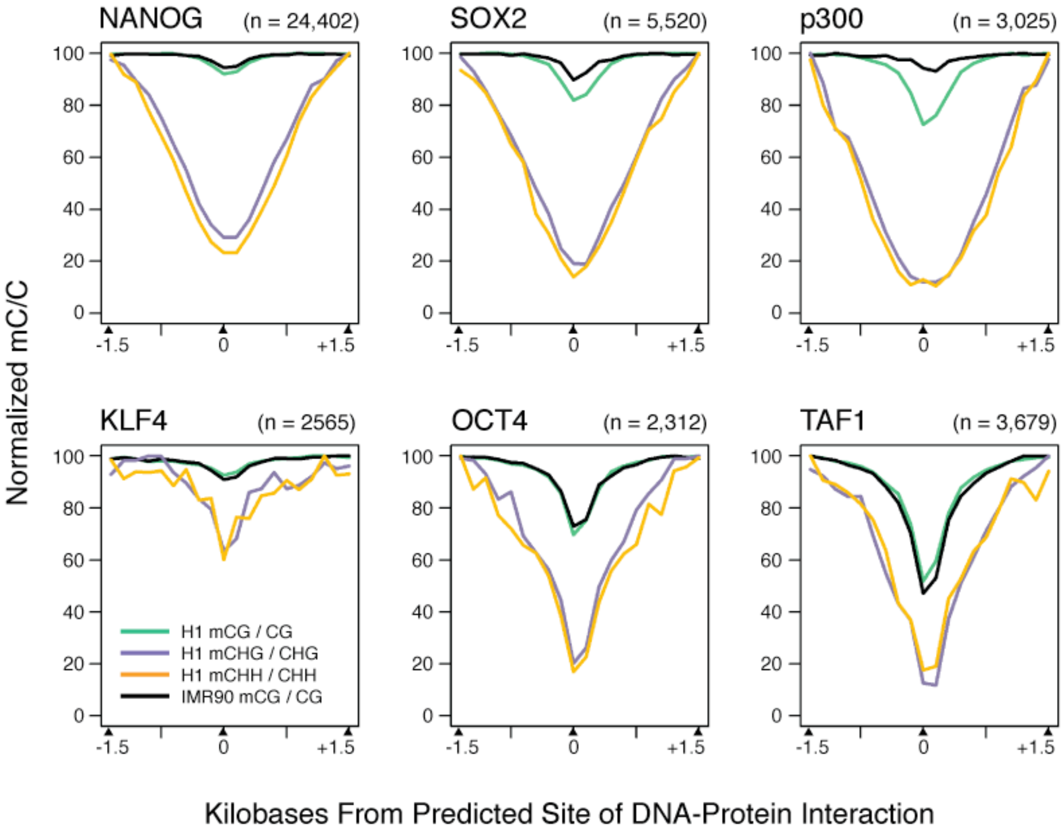
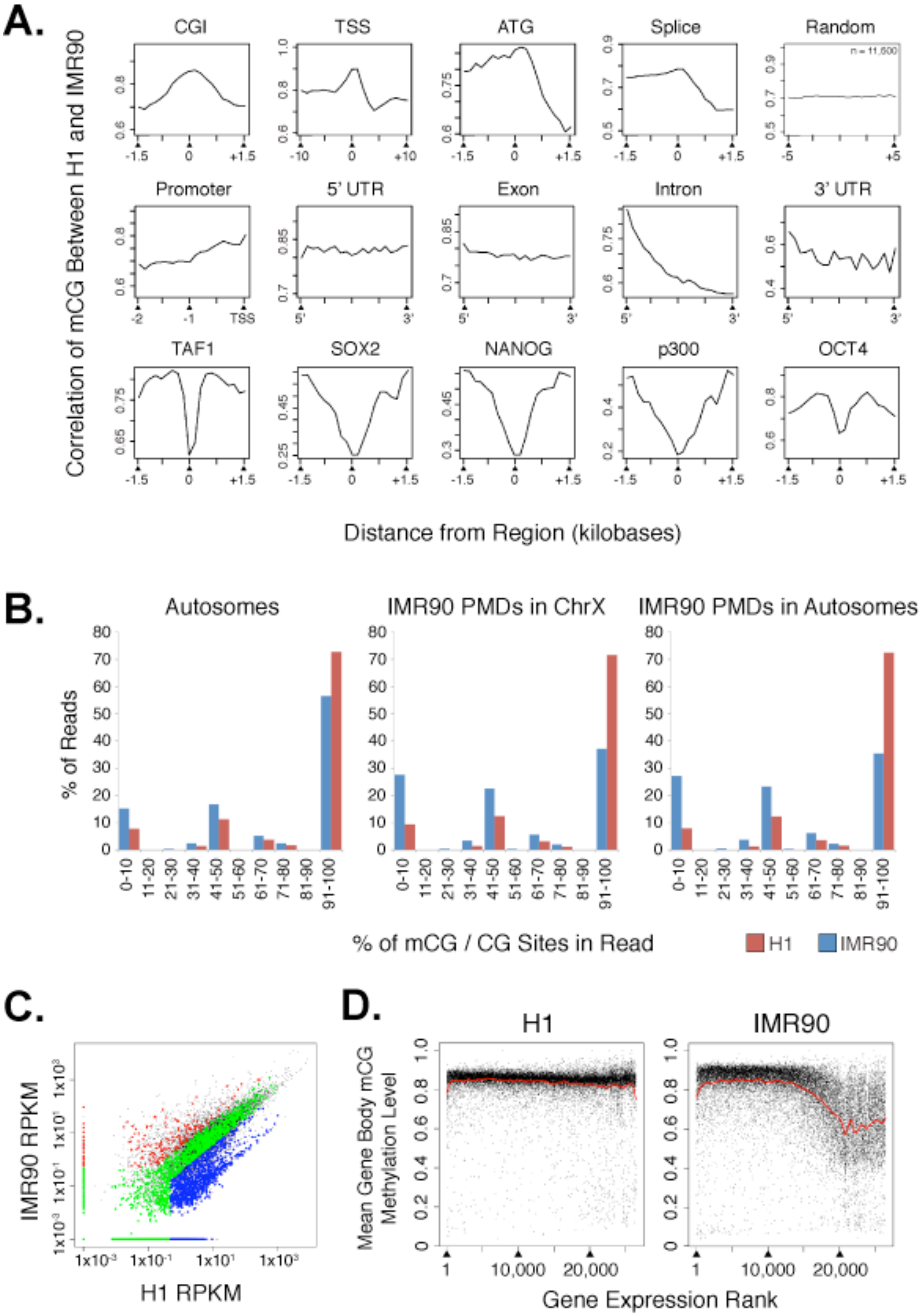


Figure B.14 DNA methylation at sites of DNA-protein interaction.

The average relative DNA methylation densities in each sequence context are shown from 1.5 kb upstream to 1.5 kb downstream of the predicted sites of DNA-protein interaction identified by ChIP-Seq that were at least 1.5 kb from the closest transcriptional start site.

Figure B.15 Cell-type variation in DNA methylation.

(A) The Pearson correlation coefficient of mCG methylation density (y-axis) between H1 and IMR90 at various genomic features. Regions were divided in 20 equally sized bins from 5' to 3' or based on the distance from the localization of the genomic feature as indicated. Pearson correlation was determined in each bin considering all the H1 and IMR90 occurrences of the given genomic region.
(B) For MethylC-Seq reads located within a set of genomic regions, the percentage of CG sites within each read that were methylated was calculated, and the percent of all reads within the regions (y-axis) that were methylated at given percentages (x-axis) is displayed. This is presented for H1 and IMR90 MethylC-Seq reads in autosomes, in IMR90 partially methylated domains on chromosome X, and IMR90 partially methylated domains in autosomes.
(C) Comparison of transcript abundance between H1 and IMR90 of genes with a transcriptional start site located in or within 10 kb of a PMD. Black dots indicate all genes in the genome, blue and red dots indicate PMD genes whose expression is 3-fold higher in H1 or IMR90, respectively, and green indicates PMD genes not differentially expressed.
(D) For each gene in H1 and IMR90, the mean gene body mCG methylation levels were calculated at mCG sites covered by at least 10 reads between both strands for all genes, and plotted against the gene expression rank value, 1 being the most expressed.


A.
Correlation of mCG Between H1 and IMR90
CGI
TSS
ATG
Splice
Random
n = 11,600
Promoter
5' UTR
Exon
Intron
3' UTR
TAF1
SOX2
NANOG
p300
OCT4
-1.5
0
+1.5
-10
+10
-5
+5
-2
-1
TSS
5'
3'
Distance from Region (kilobases)
B.
Autosomes
IMR90 PMDs in ChrX
IMR90 PMDs in Autosomes
% of Reads
0
10
20
30
40
50
60
70
80
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
% of mCG / CG Sites in Read
H1
IMR90
C.
IMR90 RPKM
H1 RPKM
1x10⁻³
1x10⁻¹
1x10¹
1x10³
D.
H1
IMR90
Mean Gene Body mCG Methylation Level
0
0.2
0.4
0.6
0.8
1.0
1
10,000
20,000
Gene Expression Rank

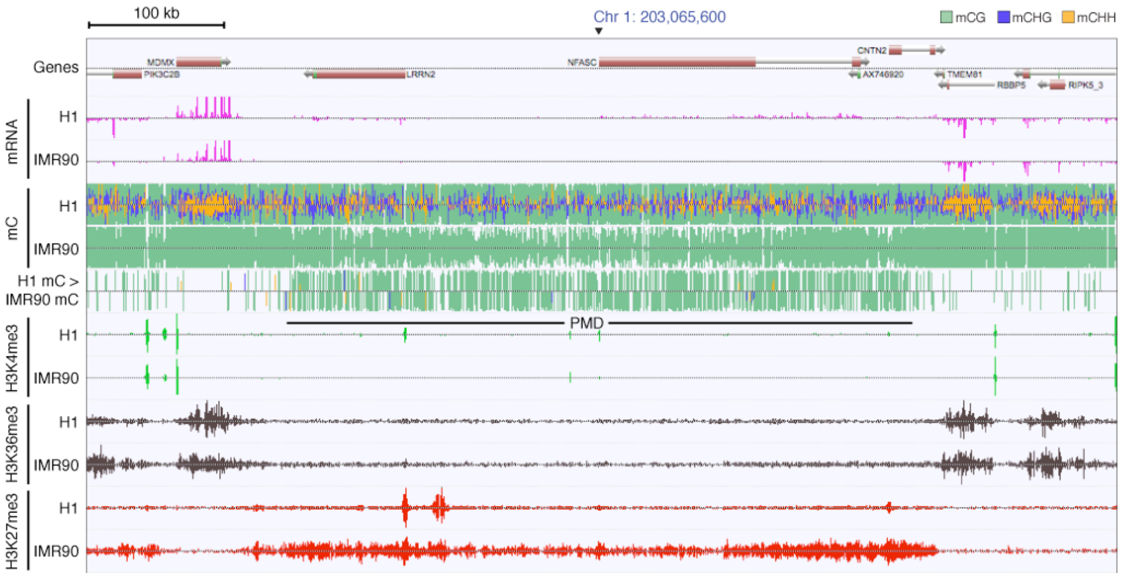


Figure B.16 DNA methylation, mRNA, and histone modifications in H1 and IMR90 at a PMD in IMR90.

For DNA methylation tracks, vertical lines above and below the dotted central line indicate the presence of methylcytosines on the Watson and Crick strands, respectively. The color represents the context of DNA methylation, as indicated, and the vertical height of the line indicates the methylation level of each methylcytosine. The IMR90 < H1 mC track indicates methylcytosines that are significantly more methylated in H1 relative to IMR90 at a 5% FDR (Fisher's Exact Test), and the color represents the context of DNA methylation. Vertical bars in the mRNA and histone modification tracks represent sequence tag enrichment.

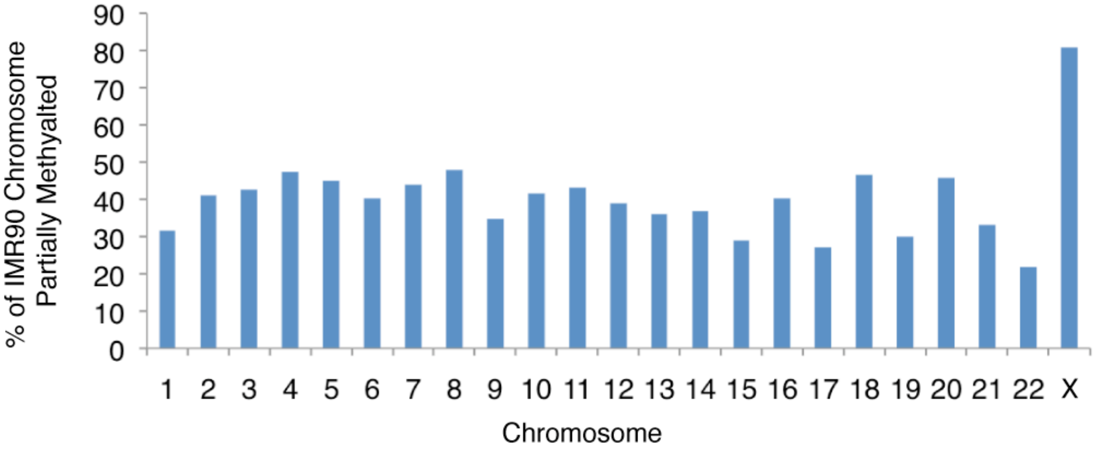


Figure B.17 The percent of each IMR90 chromosome that is within a Partially Methylated Domain.

The fraction of each chromosome that is covered by a PMD in IMR90 was calculated and plotted for each autosome, as well as the X chromosome.

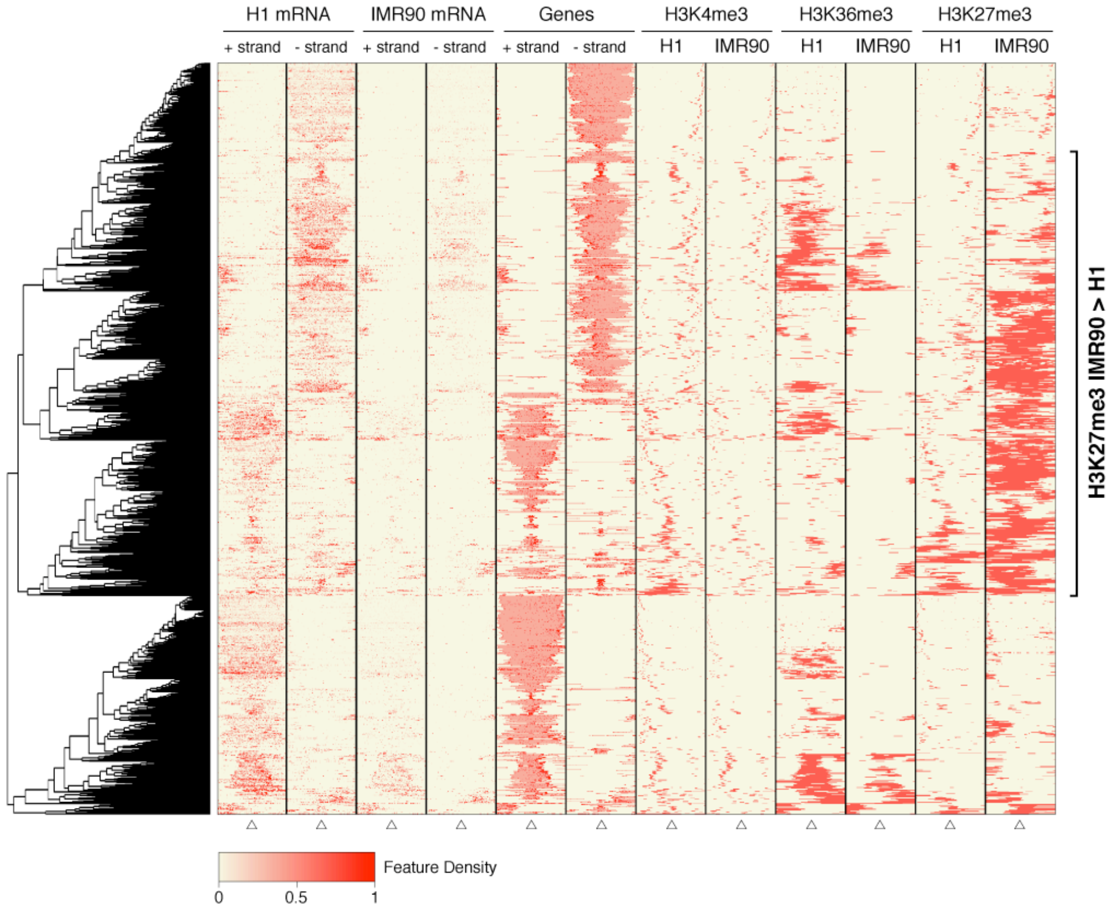


Figure B.18 Transcriptional activity and epigenetic modifications at Partially Methylated Domains.

The density of strand-specific mRNA reads, as well as the presence of domains of H3K4me3, H3K36me3, and H3K27me3 in H1 and IMR90 was profiled 20 kb upstream to 20 kb downstream of each gene located in an IMR90 PMD. Open triangles indicate the central point in each 40 kb window. Also displayed is the presence within the Human reference sequence of genes on each strand, where pink coloring indicates the gene body and dark red boxes represent exons. The complete linkage hierarchical clustering of the regions based on these data is presented.

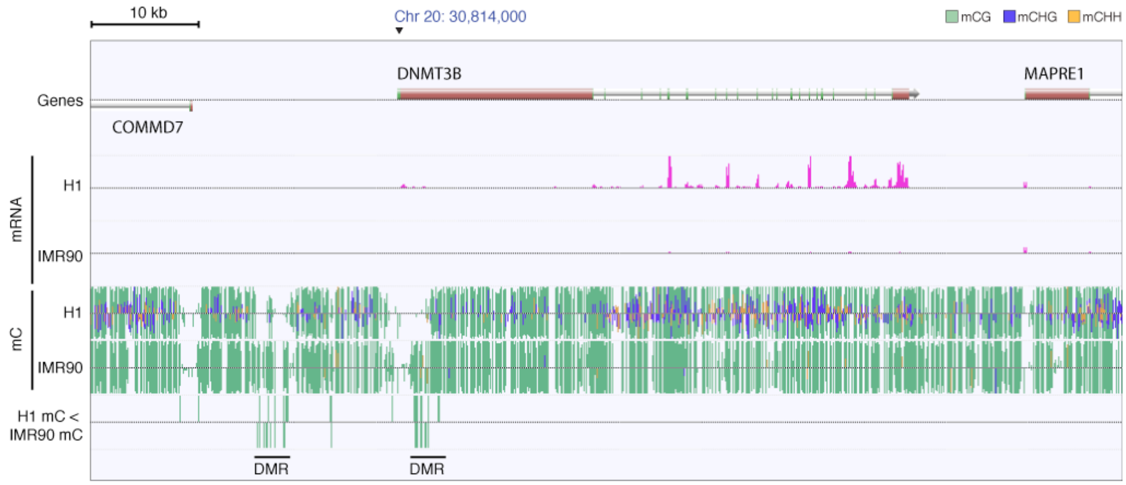


Figure B.19 Differentially methylated regions proximal to *DNMT3B*.

AnnoJ genome browser display of DNA methylation and mRNA at two DMRs upstream of *DNMT3B*. For DNA methylation tracks, vertical lines above and below the dotted central line indicate the presence of methylcytosines on the Watson and Crick strands, respectively. The color represents the context of DNA methylation, as indicated, and the vertical height of the line indicates the methylation level of each methylcytosine. The H1 < IMR90 mC track indicates methylcytosines that are significantly more methylated in IMR90 relative to H1 at a 5% FDR (Fisher's Exact Test), and the color represents the context of DNA methylation.

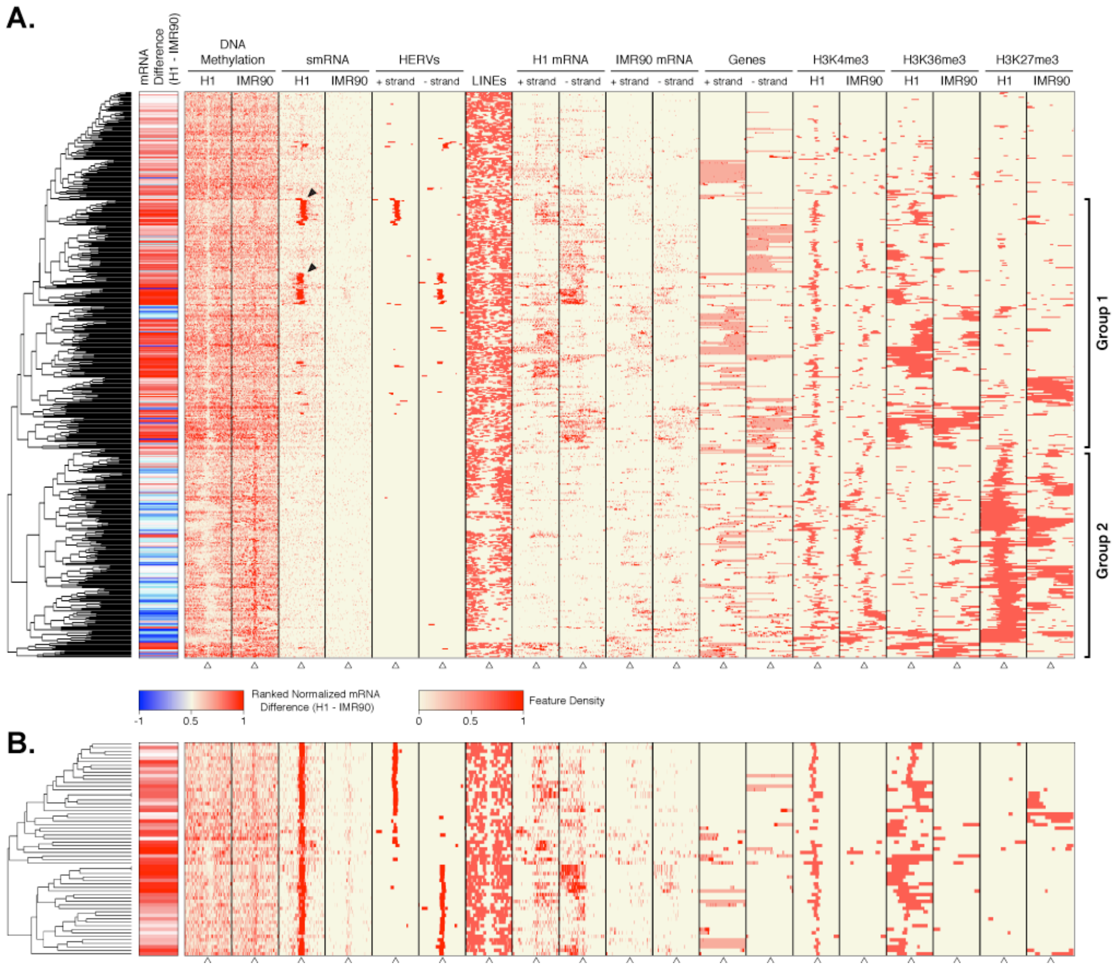


Figure B.20 Clustering of genomic, epigenetic and transcriptional features at differentially methylated regions.

(A) The density of DNA methylation, smRNA reads, strand-specific mRNA reads, and the presence of domains of H3K4me3, H3K36me3 and H3K27me3 in H1 and IMR90 was profiled 20 kb upstream to 20 kb downstream of each of the 491 DMRs where DNA methylation was more prevalent in IMR90 than H1. Open triangles indicate the central point in each window. Concurrently displayed on the side colorbar is the difference between H1 and IMR90 mRNA levels. Also displayed is the presence within the Human reference sequence of predicted Human Endogenous Retroviruses (HERVs), LINEs, and genes on each strand, where pink coloring indicates the gene body and dark red boxes represent exons. Complete linkage hierarchical clustering of the regions based on these data is presented. Black triangles indicate regions enriched for smRNAs that are coincident with HERVs. Group 1 represents DMRs associated with sequences that are more highly expressed in H1 cells, are enriched for H3K4me3 and H3K36me3, and are depleted in H3K27me3. Group 2 contains DMRs that are associated with sequences that are more highly expressed in IMR90 cells and generally enriched for H3K4me3 and H3K27me3.
(B) Clustering was performed for the 61 smRNA clusters that co-localize with DMRs, utilizing the same data as in panel (A).

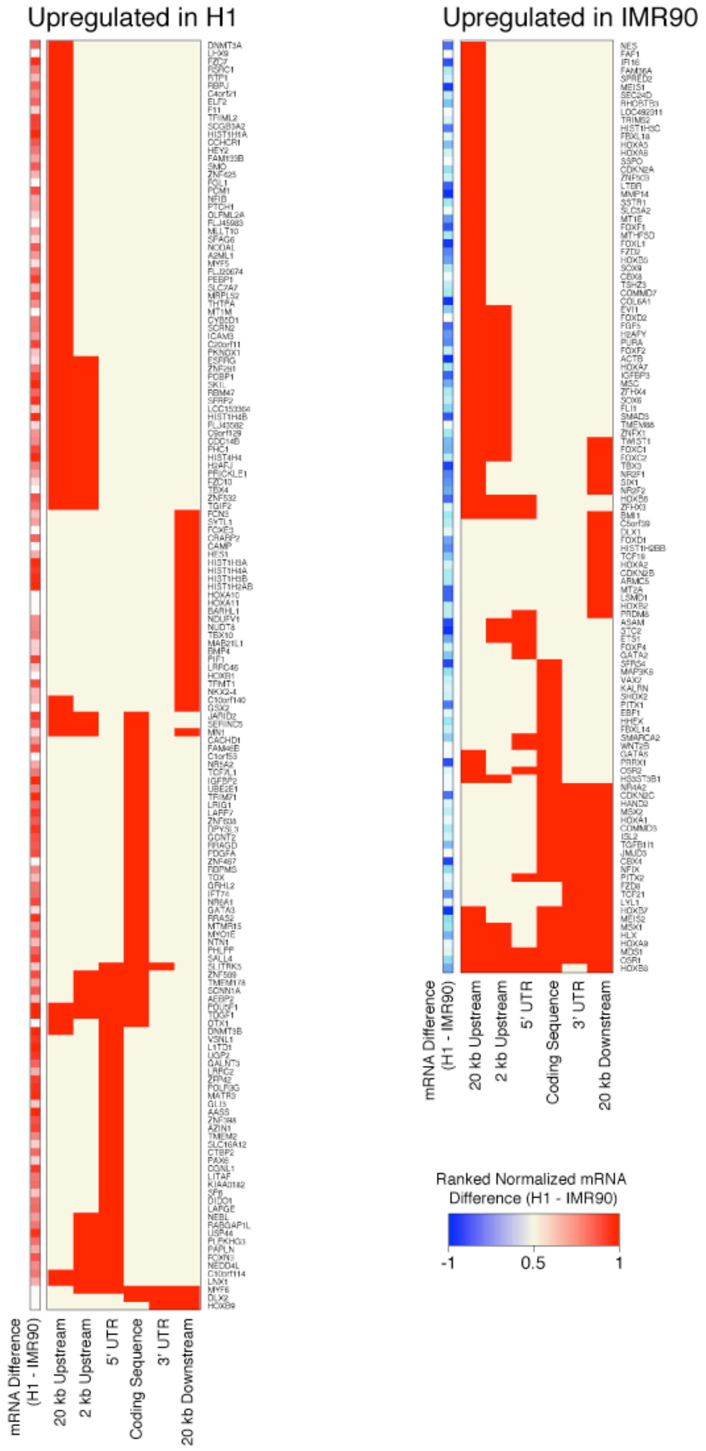


Figure B.21 Genes within 20 kb of IMR90 hypermethylated regions that are more expressed in H1 or IMR90.

Gene regions bearing differential methylation are indicated in red. Side colorbar displays normalized differential expression (red and blue for genes upregulated in H1 and IMR90, respectively).

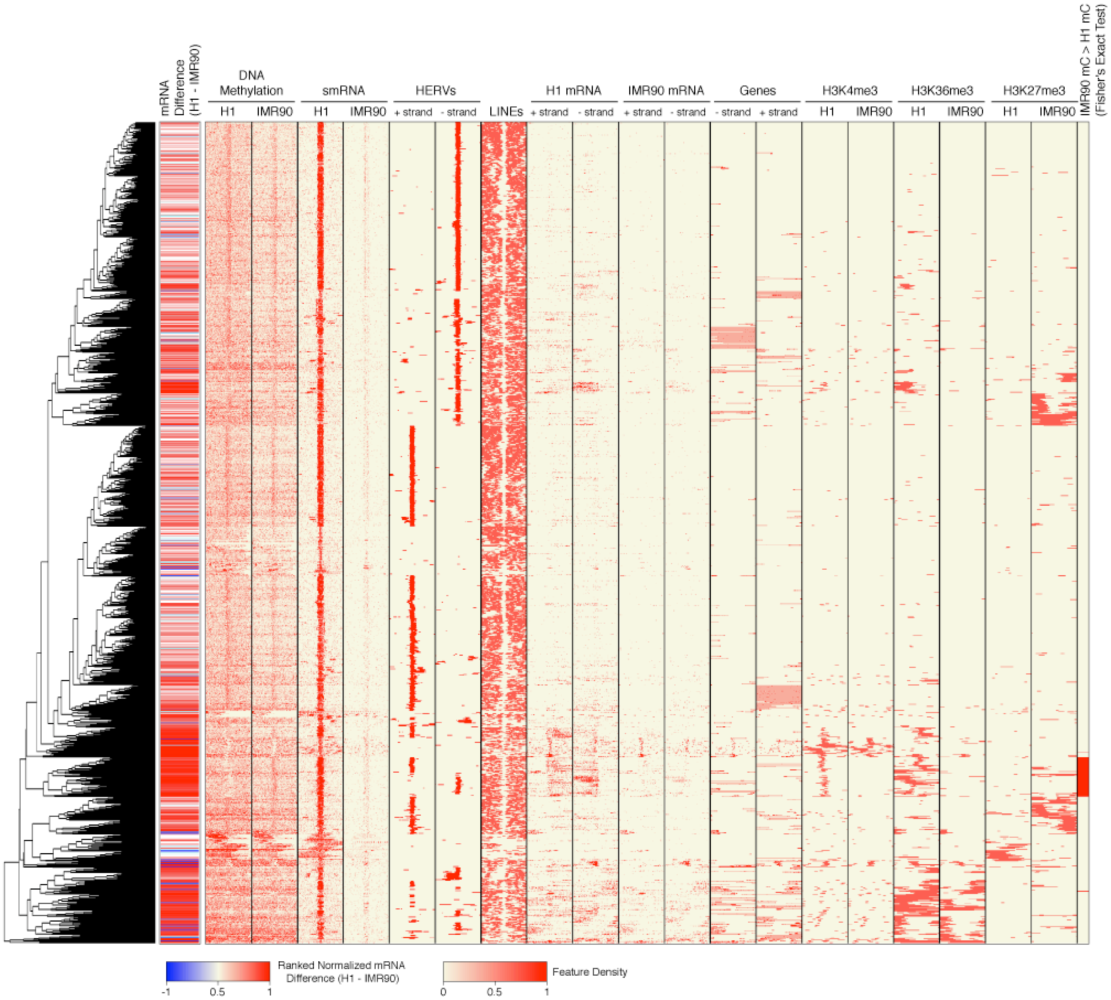


Figure B.22 Transcriptional activity and epigenetic modifications at small RNA clusters.

The density of DNA methylation, smRNA reads, strand-specific mRNA reads, and the presence of domains of H3K4me3, H3K36me3 and H3K27me3 in H1 and IMR90 was profiled 20 kb upstream to 20 kb downstream of each of the small RNA clusters identified in H1. Open triangles indicate the central point in each window. Concurrently displayed on the side colorbar is the difference between H1 and IMR90 mRNA levels. Also displayed is the presence within the Human reference sequence of predicted Human Endogenous Retroviruses (HERVs), LINEs, and genes on each strand, where pink coloring indicates the gene body and dark red boxes represent exons. The complete linkage hierarchical clustering of the regions based on these data is presented. Black triangles indicate regions enriched for smRNAs that are coincident with HERVs.

Table B.1 MethylC-Seq read number for each replicate and library before and after removal of clonal reads and post-processing.

Sample	Library	Mapped reads	Clonal reads removed	Post-processed
H1 replicate 1	A	371,110,537	301,695,580	625,394,739
	B	372,383,553	326,481,661	-
H1 replicate 2	A	324,385,074	286,911,246	529,263,306
	B	267,974,472	241,524,826	-
	C	2,548,189	2,413,936	-
H1 combined	All	1,338,401,825	1,159,027,249	1,154,658,045
IMR90 replicate 1	A	282,861,549	255,089,801	563,354,527
	B	263,215,177	238,240,832	-
	C	74,273,121	70,261,312	-
IMR90 replicate 2	A	278,447,167	245,766,218	620,520,572
	B	326,435,295	282,843,648	-
	C	98,506,925	92,229,458	-
IMR90 combined	All	1,323,739,234	1,184,431,269	1,183,875,099

Table B.2 Details of BS-PCR experiments.

Primer	Chromosome	Coordinates	Primer sequence (5' - 3')
hs_bspcr_1F	1	200015530 - 200015725	GTAATTGGTAGAGAAATGAATTTATTTAG
hs_bspcr_1R	1	-	CTCTTTTCTAAAACCTCTTAAACTTTTATC
hs_bspcr_3F	3	100016095 - 100016287	GGTATAATGTTAGAAAGTGATATATTATGAAAATAAATTG
hs_bspcr_3R	3	-	CTCATATAAATCCATCTACTCCCTCATCAC
hs_bspcr_10F	10	30837441 - 30837664	GAGTGATTTTAATATTTTGATTAAGAGG
hs_bspcr_10R	10	-	CATACAAACCATCAAATCACATTTCCTAC

Bisulfite-PCR Chr. 1	**mC #1**	**mC #2**	**mC #3**	**mC #4**
Coordinate (context)	*200015586 (mCHG)*	*200015620 (mCHG)*	*200015647 (mCG)*	*200015676 (mCG)*
H1 MethylC-seq	10/19	6/30	40/43	52/61
IMR90 MethylC-seq	0/30	0/42	47/58	50/58
H1 BS-PCR	7/12	2/12	12/12	12/12
H9 BS-PCR	5/11	0/11	11/11	9/11
iPS(IMR90) BS-PCR	6/16	0/16	14/16	14/16
IMR90 BS-PCR	0/5	0/5	5/5	5/5
H1(BMP4) BS-PCR	0/10	0/10	10/10	10/10

Bisulfite-PCR Chr. 3	**mC #1**	**mC #2**	**mC #3**	**mC #4**	**mc #5**	**mc #6**
Coordinate (context)	*100016170 (mCHH)*	*100016182 (mCHG)*	*100016191 (mCG)*	*100016201 (mCHH)*	*100016209 (mCHG)*	*100016237 (mCHH)*
H1 MethylC-seq	2/16	9/22	21/26	4/32	1/31	13/41
IMR90 MethylC-seq	0/11	0/17	20/20	0/22	0/24	0/30
H1 BS-PCR	1/20	5/20	17/20	0/20	1/20	1/20
H9 BS-PCR	1/20	5/20	20/20	0/20	1/20	2/20
iPS(IMR90) BS-PCR	3/19	7/19	17/19	4/19	5/19	9/19
IMR90 BS-PCR	0/5	0/5	5/5	0/5	0/5	0/5
H1(BMP4) BS-PCR	0/11	0/11	11/11	0/11	0/11	0/11

Bisulfite-PCR Chr. 10	**mC #1**	**mC #2**	**mC #3**	**mC #4**	**mc #5**
Coordinate (context)	*30837519 (mCHG)*	*30837540 (mCG)*	*30837574 (mCHG)*	*30837605 (mCHH)*	*30837609 (mCHH)*
H1 MethylC-seq	11/35	25/28	0/29	6/31	0/28
IMR90 MethylC-seq	0/32	20/23	0/34	0/46	0/47
H1 BS-PCR	9/23	18/23	0/23	3/23	0/23
H9 BS-PCR	1/15	13/15	1/15	0/15	4/15
iPS(IMR90) BS-PCR	3/15	13/15	2/15	3/15	0/15
IMR90 BS-PCR	0/3	3/3	0/3	0/3	0/3
H1(BMP4) BS-PCR	1/6	6/6	0/6	0/6	0/6

Primers used for bisulfite sequencing validation experiments are displayed in the top table. In the details of the bisulfite PCR sequencing results, the numerator and denominator are the methylated cytosines and total number of sequenced clones, respectively.

REFERENCES

1. Holliday, R., and Pugh, J. E. (1975) *Science* **187**(4173), 226-232

2. Riggs, A. D. (1975) *Cytogenet Cell Genet* **14**(1), 9-25

3. Bestor, T. H. (2000) *Hum Mol Genet* **9**(16), 2395-2402

4. Li, E., Bestor, T. H., and Jaenisch, R. (1992) *Cell* **69**(6), 915-926

5. Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K. D., Carrington, J. C., Doerge, R. W., Colot, V., and Martienssen, R. (2004) *Nature* **430**(6998), 471-476

6. Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999) *Cell* **99**(3), 247-257

7. Reik, W. (2007) *Nature* **447**(7143), 425-432

8. Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009) *Nat Struct Mol Biol* **16**(5), 564-571

9. Weber, M., and Schubeler, D. (2007) *Curr Opin Cell Biol* **19**(3), 273-280

10. Cedar, H., and Bergman, Y. (2009) *Nat Rev Genet* **10**(5), 295-304

11. Rauch, T. A., Wu, X., Zhong, X., Riggs, A. D., and Pfeifer, G. P. (2009) *Proc Natl Acad Sci U S A* **106**(3), 671-678

12. Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., LeProust, E. M., Park, I. H., Xie, B., Daley, G. Q., and Church, G. M. (2009) *Nat Biotechnol* **27**(4), 361-368

13. Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E. M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I. H., Yu, J., Daley, G. Q., Eggan, K., Hochedlinger, K., Thomson, J., Wang, W., Gao, Y., and Zhang, K. (2009) *Nat Biotechnol* **27**(4), 353-360

14. Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008) *Nature* **454**(7205), 766-770

15. Ludwig, T. E., Bergendahl, V., Levenstein, M. E., Yu, J., Probasco, M. D., and Thomson, J. A. (2006) *Nat Methods* **3**(8), 637-646

16. Ludwig, T. E., Levenstein, M. E., Jones, J. M., Berggren, W. T., Mitchen, E. R., Frane, J. L., Crandall, L. J., Daigh, C. A., Conard, K. R., Piekarczyk, M. S., Llanas, R. A., and Thomson, J. A. (2006) *Nat Biotechnol* **24**(2), 185-187

17. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007) *Nat Genet* **39**(3), 311-318

18. Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., and Ren, B. (2007) *Cell* **128**(6), 1231-1245

19. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) *Genome Biol* **10**(3), R25

20. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004) *Genome Biol* **5**(10), R80

21. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008) *Cell* **133**(3), 523-536

22. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008) *Nature* **452**(7184), 215-219

23. Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., and Jones, J. M. (1998) *Science* **282**(5391), 1145-1147

24. Nichols, W. W., Murphy, D. G., Cristofalo, V. J., Toji, L. H., Greene, A. E., and Dwight, S. A. (1977) *Science* **196**(4285), 60-63

25. Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R. (2000) *Proc Natl Acad Sci U S A* **97**(10), 5237-5242

26. Woodcock, D. M., Crowther, P. J., and Diver, W. P. (1987) *Biochem Biophys Res Commun* **145**(2), 888-894

27. Aoki, A., Suetake, I., Miyagawa, J., Fujio, T., Chijiwa, T., Sasaki, H., and Tajima, S. (2001) *Nucleic Acids Res* **29**(17), 3506-3512

28. Gowher, H., and Jeltsch, A. (2001) *J Mol Biol* **309**(5), 1201-1208

29. Gonzalo, S., Jaco, I., Fraga, M. F., Chen, T., Li, E., Esteller, M., and Blasco, M. A. (2006) *Nat Cell Biol* **8**(4), 416-424

30. Steinert, S., Shay, J. W., and Wright, W. E. (2004) *Mol Cell Biol* **24**(10), 4571-4580

31. Brunner, A. L., Johnson, D. S., Kim, S. W., Valouev, A., Reddy, T. E., Neff, N. F., Anton, E., Medina, C., Nguyen, L., Chiao, E., Oyolu, C. B., Schroth, G. P.,

Absher, D. M., Baker, J. C., and Myers, R. M. (2009) *Genome Res* **19**(6), 1044-1056

32. Ferguson-Smith, A. C., and Greally, J. M. (2007) *Nature* **449**(7159), 148-149

33. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007) *Nature* **449**(7159), 248-251

34. Bell, A. C., and Felsenfeld, G. (2000) *Nature* **405**(6785), 482-485

35. Clark, S. J., Harrison, J., and Molloy, P. L. (1997) *Gene* **195**(1), 67-71

36. Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., and Tilghman, S. M. (2000) *Nature* **405**(6785), 486-489

37. Kitazawa, S., Kitazawa, R., and Maeda, S. (1999) *J Biol Chem* **274**(40), 28787-28793

38. Mancini, D. N., Singh, S. M., Archer, T. K., and Rodenhiser, D. I. (1999) *Oncogene* **18**(28), 4108-4119

39. Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007) *Science* **316**(5830), 1497-1502

40. Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009) *Nature* **459**(7243), 108-112

41. Schmidl, C., Klug, M., Boeld, T. J., Andreesen, R., Hoffmann, P., Edinger, M., and Rehli, M. (2009) *Genome Res* **19**(7), 1165-1174

42. Jones, P. A., and Baylin, S. B. (2007) *Cell* **128**(4), 683-692

43. Hellman, A., and Chess, A. (2007) *Science* **315**(5815), 1141-1143

44. Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P. W., Beighton, G., Bello, P. A., Benvenisty, N., Berry, L. S., Bevan, S., Blum, B., Brooking, J., Chen, K. G., Choo, A. B., Churchill, G. A., Corbel, M., Damjanov, I., Draper, J. S., Dvorak, P., Emanuelsson, K., Fleck, R. A., Ford, A., Gertow, K., Gertsenstein, M., Gokhale, P. J., Hamilton, R. S., Hampl, A., Healy, L. E., Hovatta, O., Hyllner, J., Imreh, M. P., Itskovitz-Eldor, J., Jackson, J., Johnson, J. L., Jones, M., Kee, K., King, B. L., Knowles, B. B., Lako, M., Lebrin, F., Mallon, B. S., Manning, D., Mayshar, Y., McKay, R. D., Michalska, A. E., Mikkola, M., Mileikovsky, M., Minger, S. L., Moore, H. D., Mummery, C. L., Nagy, A., Nakatsuji, N., O'Brien, C. M., Oh, S. K., Olsson, C., Otonkoski, T., Park, K. Y., Passier, R., Patel, H., Patel, M., Pedersen, R., Pera, M. F., Piekarczyk, M. S., Pera, R. A., Reubinoff, B. E., Robins, A. J., Rossant, J., Rugg-Gunn, P., Schulz, T. C., Semb, H., Sherrer, E. S., Siemen, H., Stacey,

G. N., Stojkovic, M., Suemori, H., Szatkiewicz, J., Turetsky, T., Tuuri, T., van den Brink, S., Vintersten, K., Vuoristo, S., Ward, D., Weaver, T. A., Young, L. A., and Zhang, W. (2007) *Nat Biotechnol* **25**(7), 803-816

45. Villesen, P., Aagaard, L., Wiuf, C., and Pedersen, F. S. (2004) *Retrovirology* **1**, 32

46. Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jorgensen, H. F., John, R. M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., and Fisher, A. G. (2006) *Nat Cell Biol* **8**(5), 532-538

47. Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006) *Cell* **125**(2), 315-326

48. Kriaucionis, S., and Heintz, N. (2009) *Science* **324**(5929), 929-930

49. Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. (2009) *Science* **324**(5929), 930-935

The text of Appendix B, in full, has been submitted for publication by Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A.Thomson, Bing Ren, and Joseph R. Ecker. The dissertation author was a major contributing researcher and second author of this paper.