A TALE OF THREE PHYTOPATHOGENS: IMPACT OF TRANSPOSABLE ELEMENTS ON GENOME EVOLUTION

by

KOMAL VADNAGARA

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN BIOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2010

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Ellen J. Pritham for providing me the opportunity to work in the lab. She has been a source of support, encouragement and guidance. And I can't thank her enough because this has been a great learning experience for me. I would also like to thank Dr. Cedric Feschotte for the ideas and inputs regarding my research projects. I am thankful to Dr. Andre Pires daSilva for kindly serving on my committee. I would like to thank the Biology Department for endowing me with graduate teaching fellowship through my graduate school years. This teaching experience has been extremely enriching and a valuable one in my life. I would like to thank my labmates: Assiatu Barrie, Jainy Thomas, Megha Bajaj, Claudia Marquez, An La, Jill North, Sarah Schaack and Cheng Sun who have been extremely supportive and helpful with putting together my thesis presentation. I am extremely grateful to Cheng Sun and Qi Wang for providing all the computer assistance right from writing and modifying scripts to any troubleshooting. I would also like to thank all my friends who have just been there when I needed them the most: Amrita Naidu, Aditya Verma, Shweta Panchal, Athena Jagdish, Mahima Varma, Donna Kirkland, Eldon Prince and Marta Galvan.

I thank my parents, Mahesh and Mohini Vadnagara and my brother, Bhavyesh, without whom I would not be anything. This thesis would not be possible without their

constant love, emotional and financial support throughout. I am lucky to have such a beautiful family who believe in me and push me to pursue my dreams. I also dedicate this thesis to my grandfather, Mansukhlal Vadnagara, who has been an inspiration for me from childhood.

April 9, 2010

ABSTRACT

A TALE OF THREE PHYTOPATHOGENS: IMPACT OF TRANSPOSABLE ELEMENTS ON GENOME EVOLUTION

Komal Vadnagara, M.S.

The University of Texas at Arlington, 2010

Supervising Professor: Ellen J. Pritham

The genus *Phytophthora* harbors some notorious plant pathogens like *Phytophthora infestans* (causal of Irish potato famine), *Phytophthora sojae* (soybean rot agent), and *Phytophthora ramorum* (responsible for sudden oak death) that have significant economic, ecological and environmental impact. These phytopathogens exhibit remarkable phenotypic instability and vary tremendously in genome size from 65 Mb (*P. ramorum*) to 240 Mb (*P. infestans*). Complete draft genome sequences revealed that a substantial portion of their genome is occupied by highly repetitive DNA. This extreme genome plasticity is due to an infestation of repetitive virus-like genomic parasites called transposable elements (TEs). TEs are sometimes called jumping genes due to their capacity to move from one place to another in the genome. TEs are usually perceived as potent mutagens and the result of their proliferation in

genome is usually detrimental, although occasionally they can contribute to the evolution of the host in a variety of ways. One such mechanism is transduplication, whereby TEs capture host gene fragments, that is known to give rise to novel genes in plants. Pathogens are in a constant arms race due to their reliance on the host to reproduce and persist and the negative fitness that they impart. Therefore, it was hypothesized that the plastic *P. infestans* genome allows for a rapid response to the ever-changing environment imposed by this evolutionary arms race. To this end, we have employed bioinformatics tools (RepeatScout, RepeatMasker, BLAST tools) to identify different superfamilies of TEs and assess their distribution across three *Phytophthora* species. Much to our surprise, we found 21 TE families carrying host genes accounting for 2.4% of the *P. infestans* genome. Overall, we observe a strong preference of TEs to capture genes that are involved in epigenetic regulation and critical in plant pathogenesis cycle. We report on the detailed structure of these transduplicates and their capacity to encode a functional transposase. Our results show capture of whole cellular genes by TEs and the existence of transcript evidence for the genes captured. This observed pattern of transduplication is different from what is known in plants and other species, where the capture involves gene fragments that are usually pseudogenized. Moreover, detailed analysis of the captured genes show retention of introns confirming that the transduplication events occurred at a DNA level. Cross species and molecular phylogenetic analyses further reveal that a few capture events might have predated the split of *P. infestans* from *P. sojae* and *P. ramorum*. Hereby,

we present an in-depth analysis of various transduplication events and the impact they had in shaping the evolutionary trajectory of these phytopathogens.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 The genus *Phytophthora*: insights into evolutionary and life history

One of the current hypotheses for the eukaryotic tree of life proposes divisions of eukaryotic diversity into five large 'supergroups': Unikonts, Rhizaria, Excavates, Plantae and Chromalveolates (Keeling et al. 2005). Chromalveolates consist of many fascinating protists of environmental, medical and economical importance. Stramenopile, a group within Chromalveolates, includes diverse members like diatoms, algae and oomycetes that have key ecological niches (Figure 1.1) (Parfrey et al. 2006). Oomycetes or water moulds are fungus-like eukaryotic microorganisms that puzzled evolutionary biologists. Because of their morphological resemblance to true fungi, oomycetes remained misclassified as "fungi" for a long time (Govers 2001). However, with the availability of molecular phylogenetic data, it was revealed that oomycetes belong to the Stramenopile group with their close relatives being diatoms and golden-brown algae. Also, the study further demonstrated that oomycetes evolved independently of the true fungi, thus putting an age long debate to rest (Sogin and Silberman 1998).

Despite being fungus look-alike, oomycetes differ from true fungi in their genetics, physiology and biochemistry. For example: Fungi are mostly haploid while oomycetes are diploid. A unique biological feature of oomycetes is that their cell walls

are made up of β-1,3-glucan polymers and cellulose; unlike fungal cell walls that are mainly composed of chitin (Erwin et al. 1983). Oomycetes contain a diverse group of pathogens of insects, fish, vertebrates and plants (Kamoun 2003).

The most extensively studied oomycetes genus is *Phytophthora* (greek for "plant destroyer"), a term coined by Anton de Bary in 1861 (Large 1940). With more than 80 species, each with a varied host range, *Phytophthora* are by far the most important parasites of dicots (Tyler 2007; Judelson 2007).

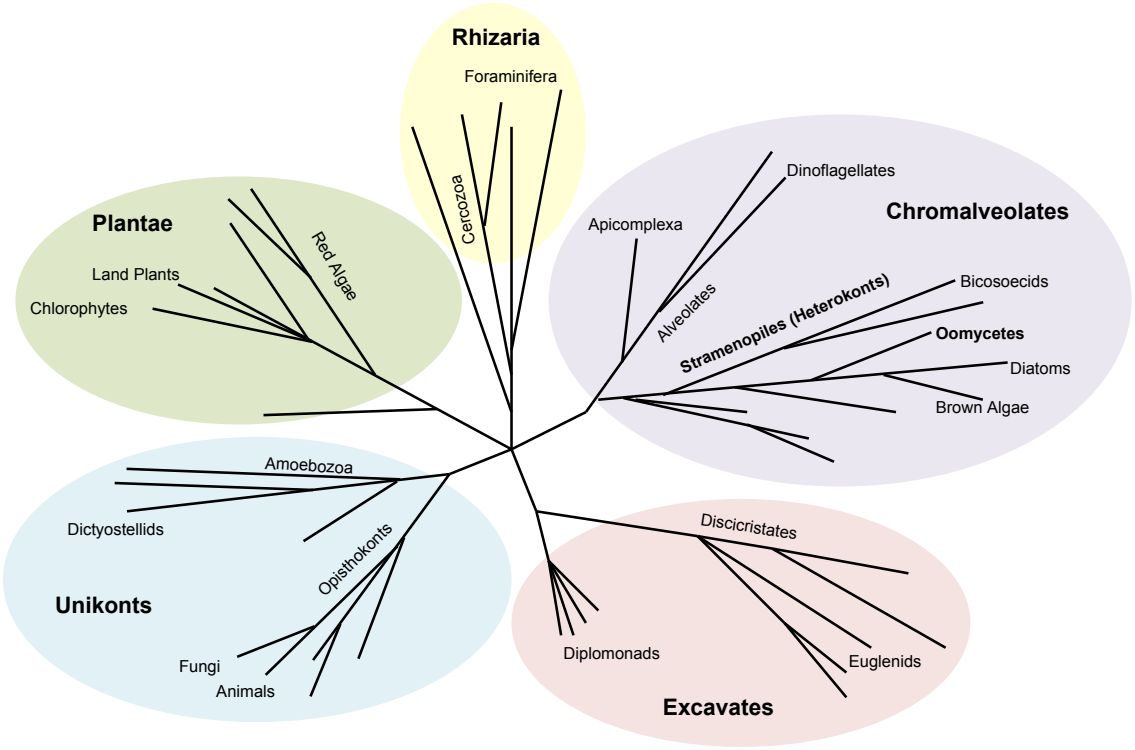


Figure 1.1: The proposed eukaryotic tree of life depicting five supergroups: Unikonts, Plantae, Rhizaria, Chromalveolates and Excavates. (Redrawn from Govers and Gijzen 2006; Keeling et al. 2005).

1.1.1 Life cycle of Phytophthora species

Phytophthora species are the most aggressive pathogens of plants and their lifecycle is broadly depicted in Figure 1.2. Typically, these species are classified as homothallic or heterothallic. The homothallic species are self-fertile and can be selfed, whereas heterothallic species are self-sterile (require two mating types: A1 and A2 to produce offspring) (Tyler 2007).

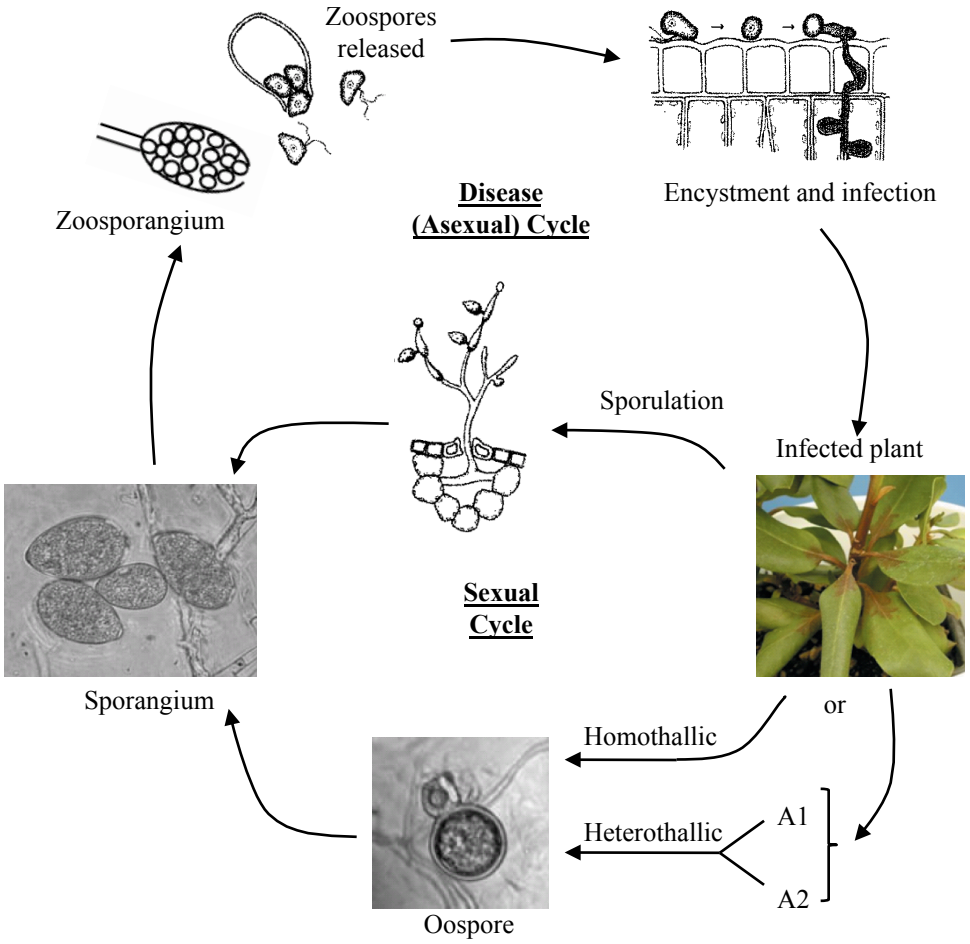


Figure 1.2: Life cycle of *Phytophthora* species. (Redrawn from Tyler 2007; Judelson 1997; and Grunwald et al. 2008).

Unlike fungi, *Phytophthora* species reproduce sexually and meiosis primarily takes place in the oogonium (female organ) and antheridium (male organ), with the latter transferring a haploid nucleus to the oogonium that goes on to differentiate into an oospore (Tyler 2007). Oospores are known to withstand harsh and unfavorable environment. They produce sporangium (short lived) that release swimming zoospores under favorable conditions (high moisture). On encountering plant surfaces, zoospores aid infection by forming adhesive cysts and hyphae to penetrate the host (Judelson 1997).

1.2 The tale of three phytopathogens

Recent advances in the field have facilitated discovery and isolation of many new species. Currently, *Phytophthora* consists of more than 80 species of destructive parasites (Tyler 2007). However, the three most celebrated and well-studied species of the genus are *Phytophthora ramorum*, *Phytophthora sojae* and *Phytophthora infestans.*

1.2.1 Phytophthora ramorum

Phytophthora ramorum is a devastating pathogen that causes 2 kinds of diseases: sudden oak death and ramorum blight on trees and woody ornamentals. With a broad host range, spanning over 109 plant species, *P. ramorum* is responsible for the rapid deforestation in coastal California and Oregon (Figure 1.3) (Grunwald et. at 2008).

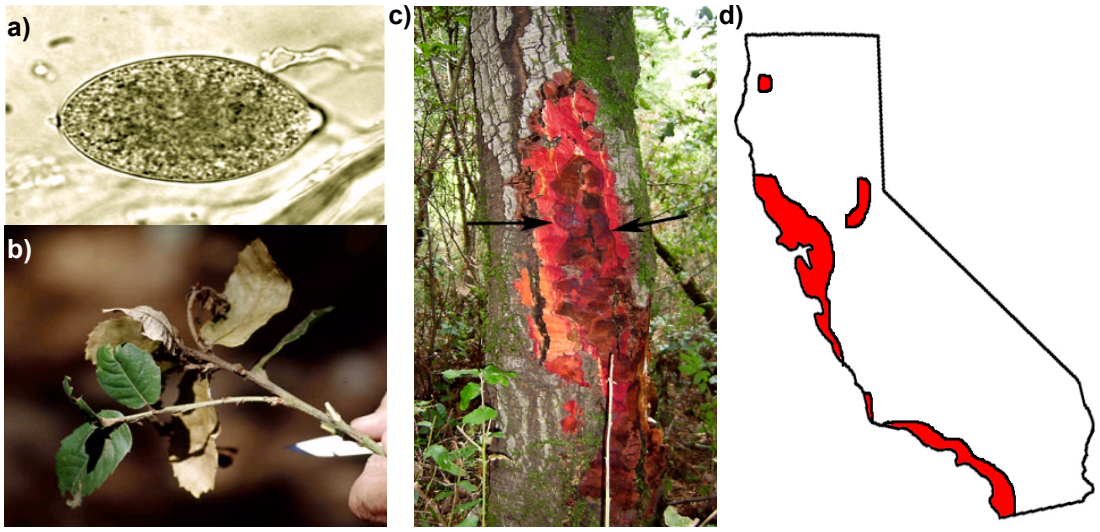


Figure 1.3: Sudden oak death agent, *Phytophthora ramorum.* a) Sporangia (photo: Matteo Garbelotto, UC Berkeley), b) Leaf death due to ramorum blight (Credit: Joseph O'Brien, USDA Forest Service), c) Bleeding cankers on Oak tree, d) Geographic distribution of sudden oak death in CA, USA (Images credit: Jennifer L. Parke and D. Schmidt, http://cisr.ucr.edu/sudden_oak_death.html).

However, its geographic location is not restricted to North America; *P. ramorum* is a major concern in European nurseries and gardens as well. It is currently managed by the eradication of infected nurseries, forests and quarantine in many areas around the globe. *P. ramorum* is easily distinguishable from other *Phytophthora* species as it forms large chlamydospores (thick-walled spores that can survive harsh, unfavorable environment). It is an obligate heterothallic parasite (self sterile, requiring opposite mating types to form oospores). The economic losses due to this pathogen are estimated to be in millions of dollars in US alone due to loss of many ornamental crops and nurseries with recreational, cultural and ecological value (Grunwald et al. 2008).

1.2.2 Phytophthora sojae

The second phytopathogen of interest is *Phytophthora sojae*, first described in 1950s (Hildebrand 1959). It is an obligate homothallic (self-fertile) parasite. Compared to *P. ramorum*, *P. sojae* has a much narrower host range with its primary host being only soybean. However, some wildflowers have been reported to be susceptible to infection by this parasite. *P. sojae* causes stem and root rot in soybeans, an economically important plant (Figure 1.4.). The economic losses are estimated to be in billions of dollars for the agricultural industry worldwide (Tyler 2007).

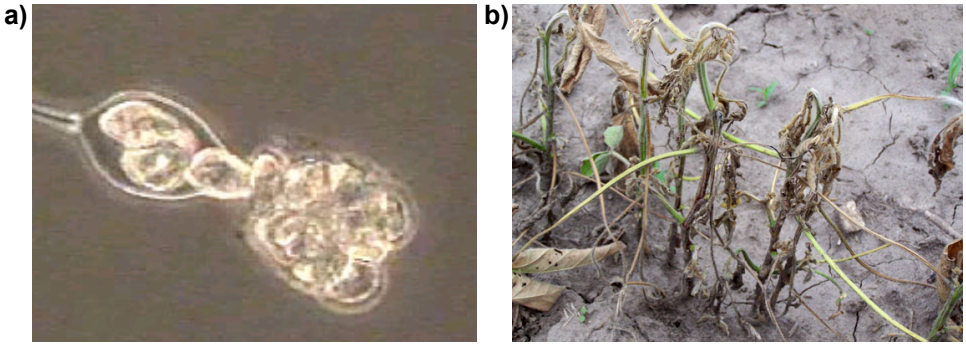


Figure 1.4: *Phytophthora sojae.* a) Sporangium releasing zoospores (photo: Edward Braun, Dept. of Plant Pathology, Iowa State University), b) Soybean stem and root rot (photo: Carl A. Bradley, University of Illinois at Urbana-Champaign).

1.2.3 Phytophthora infestans

Arguably, the most notable pathogen of the genus is *Phytophthora infestans*. It was causal of Irish potato epidemic in 1845-46 with severe aftermath like famine, hunger related deaths and large-scale immigration that left tremendous impact on human history (Reader 2009). *P. infestans* is an obligate heterothallic (self sterile, divided into two mating types: A1 and A2) parasite (Brasier 1992). This devastating

pathogen causes late blight of potato and tomato and to date poses a great threat to food crops worldwide (Figure 1.5). Potato is the fourth largest food crop, and the economic losses due to late blight are estimated to be at $6.7 billion (Haverkort et al. 2008).

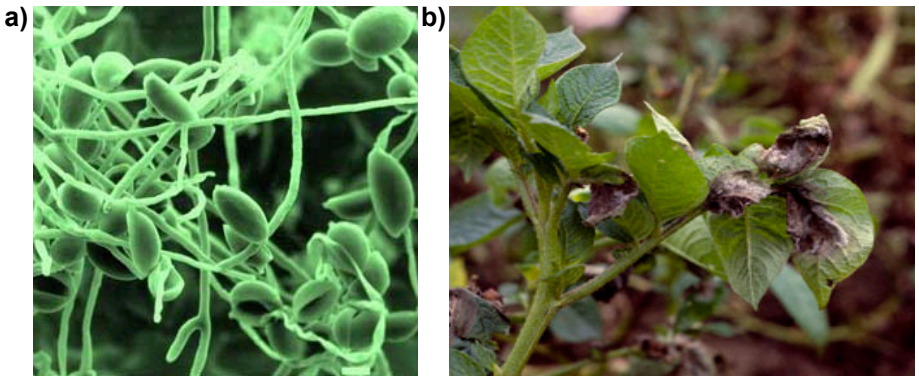


Figure 1.5: Late potato blight agent, *Phytophthora infestans.* a) An electron microscope depicting the growth of *P. infestans* on the surface of a potato leaf (Photo: http://www.afbini.gov.uk/index/news/news-releases/news-releases-archive-2008.htm?newsid=15062), b) Infected potato plant (Photo: W.E. Fry, Cornell University).

1.2.4 Phylogeny of Phytophthora species

The genus wide phylogeny of *Phytophthora* was constructed by Blair et al. 2008 using seven nuclear loci about 8700 nucleotides. Within this robust phylogenetic tree, *P. ramorum* falls into clade 8c, the second most basal group, while *P. sojae* and *P. infestans* fall into clade 7b and 1c respectively (Figure 1.6). These three species of interest represent different phylogenetic clades and are genetically distant from one another.

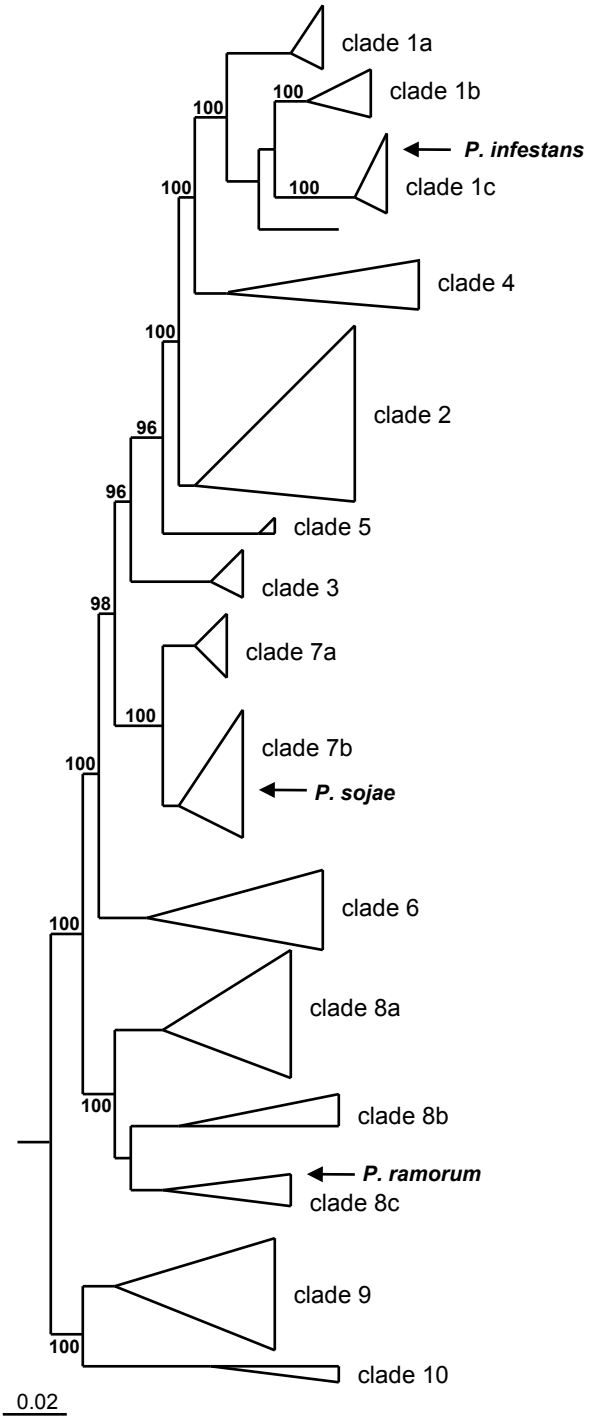


Figure 1.6: Phylogeny for the genus *Phytophthora.* (Redrawn from Blair et al. 2008; Grunwald et al. 2008). The numbers on nodes represent Bayesian posterior probabilities.

1.3 *Phytophthora* enters genomics era

Due to their economic, ecological and environmental impact in the recent times, three *Phytophthora* species: *P. ramorum, P. sojae* and *P. infestans* sequencing projects were undertaken using whole genome shotgun (WGS) approach. Owing to the unusually complex, diploid genomes, sequencing *Phytophthora* species became a mammoth task. Year 2004 marked the completion of first two *Phytophthora* genomes, followed by *P. infestans* (Tyler et al. 2006; Govers and Gijzen 2006; Haas et al. 2009). There is a high level of conserved synteny, with a core set of ~9500 orthologues genes being present in all three genomes (Table 1.1). With this wealth of data being available, *Phytophthora* research has now entered an exciting era with the required tools available to geneticists and genomicists to explore the genomes of these pathogenic oomycetes, to unravel the mechanisms of pathogenesis and develop strategies to protect plants worldwide.

1.3.1 Genome size variation among Phytophthora spp.

The genome sizes within *Phytophthora* species vary tremendously, from *P. ramorum* (~65 Mb) to *P. sojae* (~95 Mb) and the largest being *P. infestans* (~235 Mb) (Judelson 2007). There has been no evidence of whole genome or segmental duplications to explain the remarkable genome size variation across *Phytophthora* spp. However, it has been noted that a substantial portion of these pathogens' genome is comprised of repetitive DNA. Moreover, with increase in genome size, there is an increase in repetitive content (Table 1.1). The analysis of the repetitive repertoire

demonstrated that *Phytophthora* genomes are colonized by diverse populations of many virus-like repetitive particles called transposable elements (TEs).

Table 1.1: Data obtained from genome sequencing projects of the three *Phytophthora* species.

	P. ramorum	***P. sojae***	***P. infestans***
Genome size	65 Mb	95 Mb	240 Mb
Chromosomes †	-	10-13	8-10
Coverage (fold)	5.6	7.9	7.6
G+C content	53.9%	54.4%	51.0%
Collinear blocks	37 Mb	52 Mb	85 Mb
Repeat (%)	28%	39%	74%
Number of genes	14,451	16,988	17,797
Intergenic space within collinear blocks	270-1551 bp	307-2319 bp	224-3070 bp
Intergenic space outside collinear blocks	566-4351 bp	753-5896 bp	664-19144 bp
Phytophthora orthologues	12,136	12,427	11,893
Phytophthora core orthologues*	9664	9550	9583

† (Kamoun 2003). Table redrawn from Haas et al. 2009. * These core orthologues contain atleast one orthologues gene from each of the three *Phytophthora* spp.

1.4 Dynamics of Transposable Elements

TEs are sometimes called jumping genes due to their ability to move from one place to another in the genome. They are classified into two broad group based on their replicative strategies: class 1 or retrotransposons move via RNA intermediate and class 2 or DNA transposons move via DNA intermediate. Retrotransposons are further divided into sub-classes (LTR, non-LTR and DIRS) based on their mode of integration in the genome. Likewise, DNA transposons are further classified into 3 categories: classic cut n paste, rolling circle (or *Helitrons)* and self-replicating (*Mavericks*/*Polintons)* (Figure 1.7) (Feschotte and Pritham 2007; Pritham 2009).

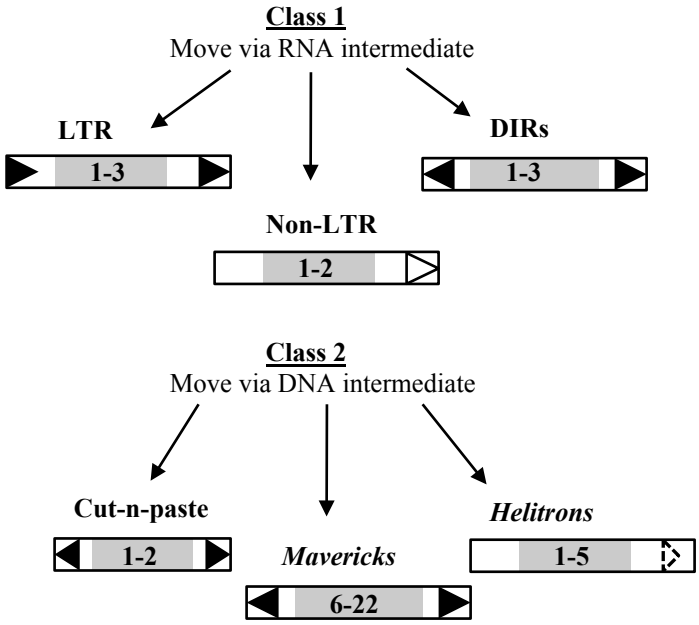


Figure 1.7: Structure of various transposable elements. Black Inverted arrows indicate terminal inverted repeats (TIRs). Black arrows in the same orientation indicate long terminal repeats (LTRs). Dashed arrow indicates palindrome-like sequences. White arrow indicates poly a tail. The numbered, grey boxes indicate the number of ORFs and proteins encoded by the autonomous transposons (Redrawn from Pritham 2009).

1.4.1 Transposable elements mediated genome expansion of Phytophthora infestans.

Close inspection of the *Phytophthora* genome sequences by us and others revealed that they are chalk full of diverse populations of TEs. Both class 1 and class 2 TEs make up a significant portion of these pathogens' genomes (Figure 1.8). However, the explosive spread of class 1, LTR / *Gypsy* elements, underlies the expansion of *P. infestans* genome. Moreover, two families of *Gypsy* elements, *Gypsy Pi-1* and *Albatross* account for about ~ 33 % of the genome (Haas et al. 2009). The comparative genome wide analysis of the three *Phytophthora* species revealed very unusual bimodal genome architecture: the "core genome" consists of highly conserved gene blocks that are

interrupted by dynamic repeats, the “plastic genome” (Gijzen 2009). More notably, it was observed that the intergenic distance increases with the increasing genome size (Table 1.1). These expanded intergenic regions are the result of direct accumulation and proliferation of transposable elements (Haas et al. 2009).

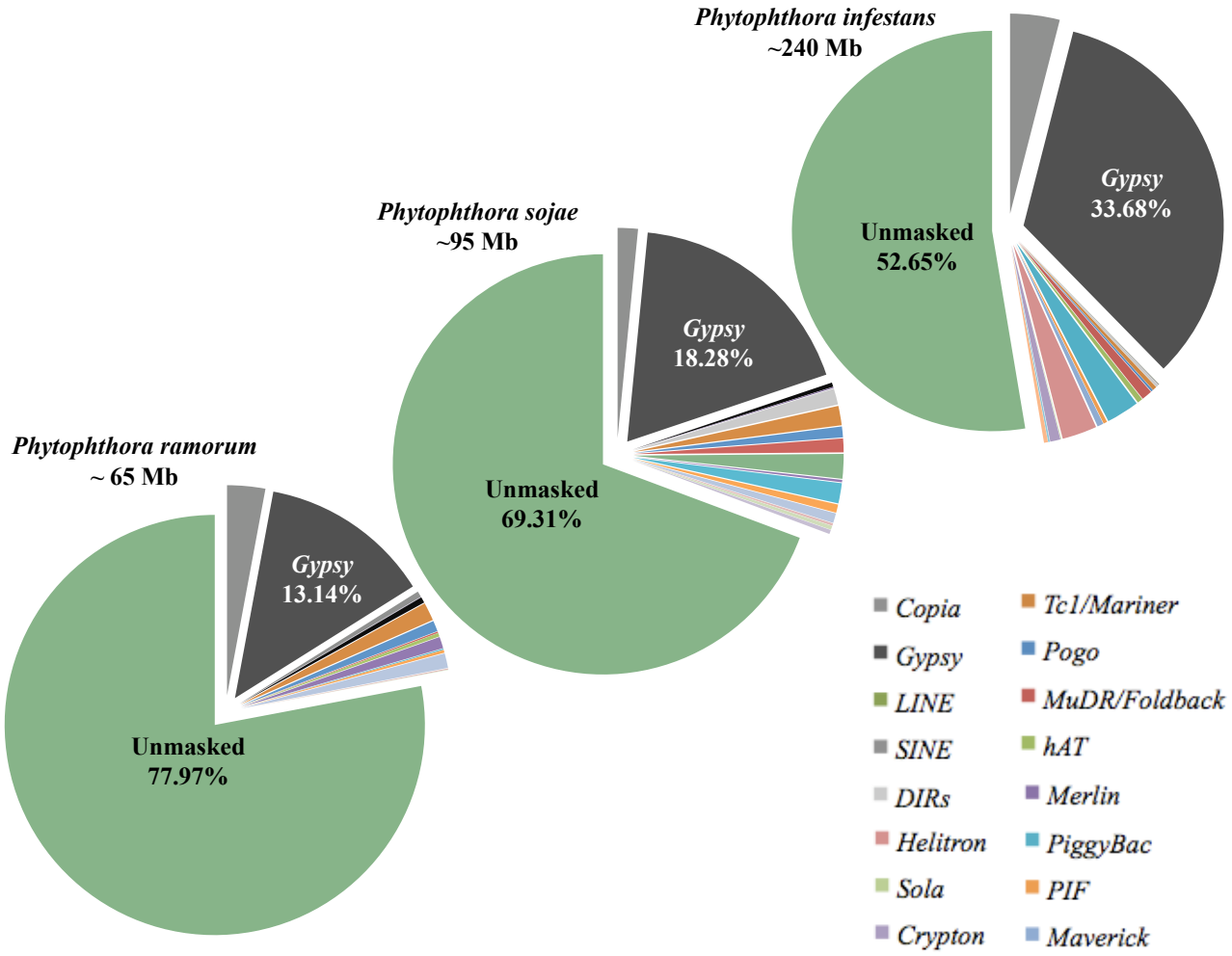


Figure 1.8: Genome expansion mediated by explosive spread of class 1, *Gypsy* retroelements, in *P. infestans*. (Vadnagara K, Pritham EJ, unpublished).

1.4.2 Genomic impact of transposable elements.

With the accumulation of data from various genome-sequencing projects, it has been very well established that TEs make up a substantial fraction of many eukaryotic genomes (Feschotte and Pritham 2007). Often perceived as "junk or selfish DNA", TEs can contribute to the evolution of the host in a myriad of ways. One of the most powerful tools to produce genetic variability is recombination. The sequence similarity between TEs can promote recombination between unrelated fragments of DNA leading to deletion or duplication events. These events are generally detrimental; however, sometimes they do have positive impact on the host, for example, the evolution of human glycophorin gene family was the result of several unequal recombination events between *Alu* elements (Class 1, non LTR) (Makalowski 2000). Another interesting facet of TEs is that by inserting into the proximity of genes, TEs can modulate the expression of adjacent genes both at transcriptional and post-transcriptional level (Feschotte 2008). The most classic example is the *Ac/Ds* transposons first described by Barbara McClintock in 1950s (McClintock 1950). These elements in maize were influencing the expression of the neighboring genes thereby leading to mosaic pattern of corn kernels. With the recent advances in science, there is substantial evidence demonstrating a whole slew of DNA binding proteins and transcription factors being derived from molecular domestication of many transposon proteins (Feschotte 2008). In recent times, one mechanism that has garnered a lot of curiosity is transduplication, a process whereby host genes are captured by transposons.

1.4.3 "Selfish DNA" lends a helping hand: the birth of new genes

Transduplication is a highly intriguing mechanism considering the propensity of TEs to mobilize the captured gene fragments leading to diversification and expansion of host genes. There is mounting evidence suggesting this process occurs on a massive scale and is extremely rampant in many plant genomes (Jiang et al. 2004; Holligan et al. 2006; Feschotte and Pritham 2009). For example, Jiang and co-workers reported >3000 *MULEs* (cut-n-paste DNA transposon) that had captured more than 1000 cellular genes in rice. Interestingly, some of these elements were shown to be evolving under purifying selection echoing a functional constraint (Hanada et al. 2009). It has been speculated that this remarkable ability of TEs to capture and re-shuffle gene fragments from various loci might represent a unique mechanism for the evolution of new chimeric genes in plants. However, the mechanism of abduction of host genes remains to be elucidated. Nonetheless, it was observed that the captured genic regions retained introns, suggestive that the transposition does not involve an RNA intermediate. Besides *MULEs*, there has been documentation of transduction by other transposable elements such as *CACTA*, *Helitrons* and the retrotransposon, *L1* (Kawasaki and Nitasaka 2004; Moran et al. 1999, Lal and Hannah 2005). With more examples being brought to attention, it appears that transduplication might be an intrinsic ability of transposons. In summary, the multifaceted personality of TEs has helped shape the evolutionary trajectory of their hosts from time to time.

1.5 Implications and aims of this study

Pathogens are in a constant arms race with their host due to their reliance on the host to reproduce and persist and the negative fitness impact that they impart. *Phytophthora* species demonstrate exquisite genetic flexibility that enables them to rapidly adapt to their ever-changing environment (Tyler 2007). It is hypothesized that the strategy that ensures this adaptive success is the bipartite genome organization. The subsets of genes critical to plant infection are located in the (repeat rich) "plastic" genome that is undergoing rapid evolutionary changes. This plasticity is directly attributed to proliferation and persistence of TEs. Hence, these phytopathogens seem to be able to outwit their plant hosts and develop resistance to host defenses (Tyler 2007) as a by-product of TE amplification. Therefore, comprehensive analysis of transposons will foster our understanding and help develop novel strategies to combat these insidious parasites. I will combine two approaches to analyze the diverse populations of transposable elements. First approach is *de novo* identification of repeats and second is homology-based method detailed in chapter 2.

This study aims in assessing the genomic impact of the explosive amplification of TEs, as well as the consequences of the competition between TE families and its effect on the biology of transposons and the host. What remains unknown is how a genome can tolerate such massive scale amplification of TEs? Do TEs have intricate targeting mechanisms that allow the avoidance of transcriptionally active regions to ensure safety? How does a TE family successfully out-compete other TE families

during the same time period? To decipher some of these questions, the goals of this study include:

1. Undertaking a genome-wide analysis of the diverse populations of transposable element (TEs) in three *Phytophthora* species: *P. ramorum, P. sojae* and *P. infestans*.

2. Conducting an in-depth analysis of the transduplicates and its evolutionary impact on the late potato blight agent, *P. infestans*.

CHAPTER 2

TRANSPOSABLE ELEMENTS MEDIATED CAPTURE, DIVERSIFICATION AND EXPANSION OF GENE FAMILIES IN *PHYTOPHTHORA INFESTANS*.

2.1 Introduction

With the accumulation of data from various genome-sequencing projects, it has been very well established that transposable elements (TEs or transposons), mobile genetic entities, make up a significant fraction of many eukaryotic genomes. For instance, ~50% of human and ~84% of maize genome is made up of transposon-derived DNA (Lander et al. 2001; Schnable et al. 2009). TEs are broadly classified into two main groups based on their replicative strategies. Class 1 or Retrotransposons move via RNA intermediate. These elements are further divided into sub-classes: LTR (*Ty1/Copia* and *Ty3/Gypsy*), non-LTR (*LINE* and *SINE*) and tyrosine recombinase (*DIRs*) based on their integration mechanisms. Class 2 or DNA transposons move via DNA intermediate. These elements are subsequently divided into 3 categories: classic cut n paste, rolling circle (*Helitrons*) and self-replicating (*Mavericks*). The cut and paste transposons are represented by 15 superfamilies classified based upon unique structural features like the size and sequence of target site duplications (TSDs) and terminal inverted repeats (TIRs) and the transposase enzyme that is encoded by autonomous copies (Feschotte and Pritham 2007; Bao et al. 2009; Goodwin et al. 2003). The *Helitrons* insert into a dinucleotide TA, have specific 5' and 3' termini and encode

a putative Rep/Helicase protein (Feschotte and Wessler 2001). While the *Mavericks* are large TEs that engender a five bp TSD and have large terminal inverted repeats. *Maverick* elements encode anywhere from six to 20 open reading frames (ORFs) (Pritham et al. 2007). The ability of TEs to replicate themselves and propagate in the genome allows them to shape the genome architecture and lead to genetic variation among species.

Usually perceived as "selfish or parasitic" DNA, TEs can contribute to the evolution of their host in a multitude of ways. One of the most intriguing examples of recent times is the remarkable ability of TEs to incorporate gene fragments and mobilize them, in a process termed transduplication. This phenomenon is extremely prevalent in many plant genomes (Jiang et al. 2004; Holligan et al. 2006). For example, there are over 3000 *PackMULEs*, *MULE* (cut and paste DNA TE) elements 'packed' with host gene fragments, documented in rice (Hanada et al. 2009). In most cases, the captured gene fragments by TEs have been shown to evolve like pseudogenes. For instance, the *Helitron* insertion in ba1-ref in maize is laden with several pseudogenes (Gupta et al. 2005). However, in the case of rice *PackMULEs*, the transduplication mechanism is speculated to be a novel mechanism giving birth to new genes (Jiang et al. 2004).

The genus *Phytophthora* harbors some notorious plant pathogens like *Phytophthora infestans*, a veteran on the plant pathology scene. Most notably known as the causal agent of the Irish potato famine in 1845-46, *P. infestans* is arguably an insidious parasite that poses great threat to tomato and potato crops every season

worldwide (Reader 2009). This pathogen exhibits exquisite genetic flexibility allowing it to rapidly adapt to its ever-changing environment. It has been speculated that the key to this amazing adaptive success lies in the plasticity of its genome that enable it to quickly adapt to host defenses. The plasticity of a genome is often times attributable to the intense activity of activity and proliferation of TEs (Haas et al. 2009).

To try to understand the role of TEs in generating plasticity in the genome we undertook a detailed overview of the extent of transposon diversity in this phytopathogen. Our analysis brings to light an unprecedented level of transduplication of whole genes mediated by DNA transposons representing four different superfamilies: *MULEs*, *Helentrons*, *PIF* and *PiggyBac*. In addition to finding cases of TEs carrying functional host genes, our results show the abduction of host genes has led to the diversification and expansion of various cellular genes. In addition, we also report on the existence of transcript evidence for the captured genes and assess the potential impact to both the biology of the TE and on the host parasite.

2.2 Material and Methods

2.2.1 Homology-based methods to examine the diversity of transposable elements

A series of tblastn searches (from November 2008-January 2009) were conducted to detect the presence of proteins of known class 1 and class 2 TE superfamilies against *P. infestans* WGS data deposited at NCBI (http://www.ncbi.nlm.nih.gov). The resulting hits of these searches were then tested to identify structural characteristics unique to each superfamily of TEs by careful examination of (up to 5 kb) flanking sequences. These structural characteristics include

terminal inverted repeats (TIRs), long terminal repeats (LTR), poly A tails and target site duplications (TSDs). Thereafter, a series of blastn searches were conducted by fusing ~50 bp upstream and downstream region flanking the transposon to identify any potential paralogous (within genome) empty sites. An empty site was annotated when another region is identified in the same genome that lacks the insertion yet contains the unduplicated target site.

2.2.2 Computational data mining of transduplicates

The *P. infestans* genome was downloaded from NCBI through GenBank accession number AATU01000000. A program called, RepeatScout version 1.0.3 was used to generate a consensus repeat library using default settings (seed size of 50 and >10 copies to be considered a repeat) (Price et al. 2005). Tandem repeat and low complexity filters were run on RepeatScout output files to remove low complexity repeats and repeats with tandemly duplicated motifs. Thereafter, Repclass (a classification tool) was run on these RepeatScout generated libraries to automate the classification of repeats (Feschotte et al. 2009). One of the limitations of RepeatScout is that the generated consensus found in the library does not extend all the way to the ends of the repeat. Therefore to validate each repeat, BLAST tools were used to identify structural characteristics such as terminal inverted repeats, long terminal repeats, target site duplications and coding capacity (ORFs, open reading frames) typical of TEs. To identify transduplicates, each family of TE was translated in six frames using an Expasy translate tool (http://ca.expasy.org/tools/dna.html). Open reading frames (ORFs) typical of TEs like transposase, integrase, helicase, replicate,

endonuclease etc were filtered out. Thereafter, PSI-blast search was conducted on any additional atypical ORFs to identify homologies to annotated proteins in the database and gather more insights into potential function of these proteins (Altschul et al. 1997). Furthermore, NCBI Conserved Domain Database (CDD) (Marchler-Bauer et al. 2009) was used to query non-TE ORFs to identify any conserved protein domains (CD). To assess the presence of transcriptional evidence, both protein and nucleotide query for transduplicates were used to conduct blastn and tblastn searches against *P. infestans* EST database at NCBI. Hits with e value of $< 10^{-4}$ were considered significant.

2.2.3 Identification of the parental copy of genes captured

To locate the parental gene copy, a series of rigorous tblastn searches were conducted using the corresponding captured gene sequence from transduplicates. For each significant hit with sufficient sequence available, the nucleotide sequences flanking the gene were extracted to determine if the putative gene is a parental copy or is carried by TEs. The extracted flanking sequences were aligned to identify structural characteristics typical of TEs, such as the presence of terminal inverted repeats (TIRs) and target site duplications (TSDs). If no such structural characteristics were identified then these entries were classified as a parental copy of the gene captured.

2.2.4 Copy number estimation and sequence divergence analysis

RepeatMasker is a program that provides positional distribution of the repeats, helps delineate clear boundaries of the repeats and provides an estimation of copy number of the TEs in the genome (A. F. A. Smit, R. Hubley, and P. Green; http://repeatmasker.org). The manually curated library of transduplicates was used to

repeatmask the genome of *P. infestans* with RepeatMasker version 3.1.5 using default settings. Thereafter, a perl script was run that automated the calculation of total bp accounted for each repeat in the genome and to get a copy number estimation using RepeatMasker output. Additionally, the RepeatMasker output provided the percent divergence for each family. Subsequently, this data was used to construct sequence divergence frequency plots using Microsoft Excel.

2.2.5 Phylogenetic analysis

Protein sequences were aligned with ClustalW, ClustalX and alignments were refined manually using GeneDoc version 2.6.02 (Larkin et al. 2007; Nicholas et al. 1997). Bayesian phylogenetic tree were constructed through the program MrBayes version 3.1 (Huelsenbeck and Ronquist 2001), applying a mixed amino acid model with a discrete gamma distribution, with four rate categories and random starting trees. Two independent runs with four Markov chains were run for several million generations until the split frequency was <0.005. Temperate difference between the 'cold' and the 'heated' chain was set to 0.5 to improve the chain swap. The sampling frequency was set to 1000 for each analysis. Maximum-likelihood phylogenies were constructed using PhyML v2.4.4 (Guindon et al. 2003) with 1000 bootstrap replicated and JTT amino acid model. Thereafter, a program called TreeView version 1.6.6 was used to display the respective phylogenies (Page 1996).

2.3 Results

2.3.1 Analysis of repetitive repertoire of P. infestans genome

To this end, we employed combination of homology-based searches using BLAST tools and *de novo* methods (see methods) to analyze repeat driven genome of *P. infestans*. We found representatives of five groups of retrotransposons and 10 DNA transposon superfamilies that constitute approximately 47% of the *P. infestans* genome, which is largely in agreement with the recent publication of the draft genome sequence (Haas et al. 2009). This comprehensive and detailed analysis allowed us to begin to investigate the impact of this recent amplification of TEs in this genome both on the host as well as on the TEs themselves.

2.3.2 Transposable elements landscape of P. infestans

2.3.2.1 Retrotransposon landscape

Retrotransposons occupy approximately 38% of the *P. infestans* genome. Moreover, the explosive spread of *Gypsy* (LTR) is the underlying factor behind the genome expansion. About third of *P. infestans* genome is comprised of LTR elements: *Gypsy* and *Copia*. Both *Gypsy* and *Copia* elements encode a gag and pol protein, however the organization of various domains within these proteins is inherently different. There are two different kinds of *Gypsy* elements present in the genome: one group with a CHROMO domain and one without (Figure 2.1). Overall, LTR elements dramatically vary in size from 4500-15000 bp. Another group of retroelements are non-LTR that encode endonuclease (EN) and reverse transcriptase (RT) and possess a poly A tail on 3' end.

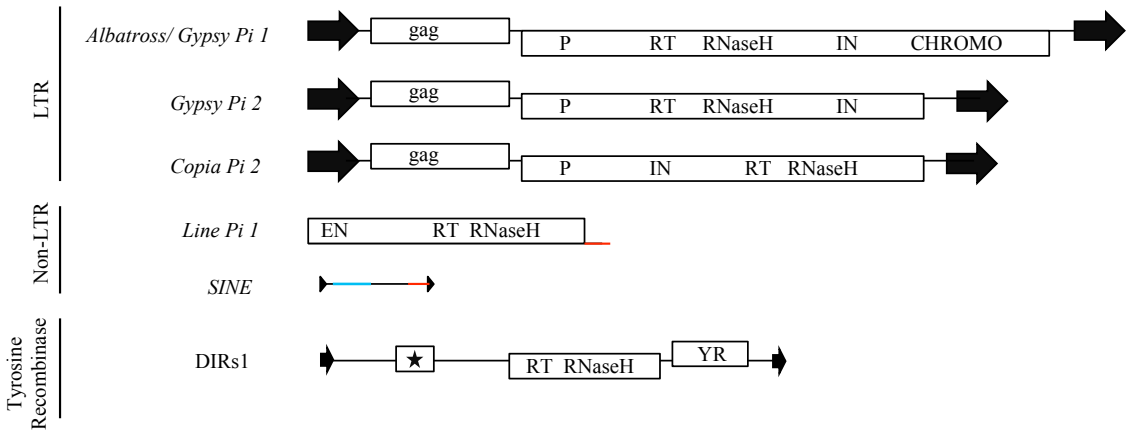


Figure 2.1: Retrotransposon landscape of *P. infestans*. Direct black arrows indicate long terminal repeats (LTR), direct triangles represent target site duplications (TSDs), blue lines = 5' tRNA related region, red lines = poly A tail. EN = Endonuclease, RT = Reverse Transcriptase, IN = Integrase, CHROMO = CHRomtin Organization Modifier domain, P = Protease, star symbol denotes unknown open reading frame.

LINE elements are autonomous elements (encode EN and RT) as opposed to *SINE* elements that do not encode any proteins; hence they are called non-autonomous (Whisson et al. 2005). And the third category of retroelements is *DIRs* that harbor tyrosine recombinase and proteins with with RT and RNaseH domain. Moreover, *P. infestans DIRs* have additional ORF of hypothetical protein with no known function. Together, these ORFs are flanked by terminal inverted repeats (Figure 2.1).

2.3.2.2 DNA transposon landscape

In agreement with previous TE reports, we identify complete TEs presenting nine DNA superfamilies including: *Tc1/mariner/pogo, PiggyBac, Mutator, hAT, Sola, Cryptons, Merlin, Helitrons* and *Mavericks*. Besides distinct structural characteristics, each superfamily encodes the typical proteins necessary for transposition (Figure 2.2).

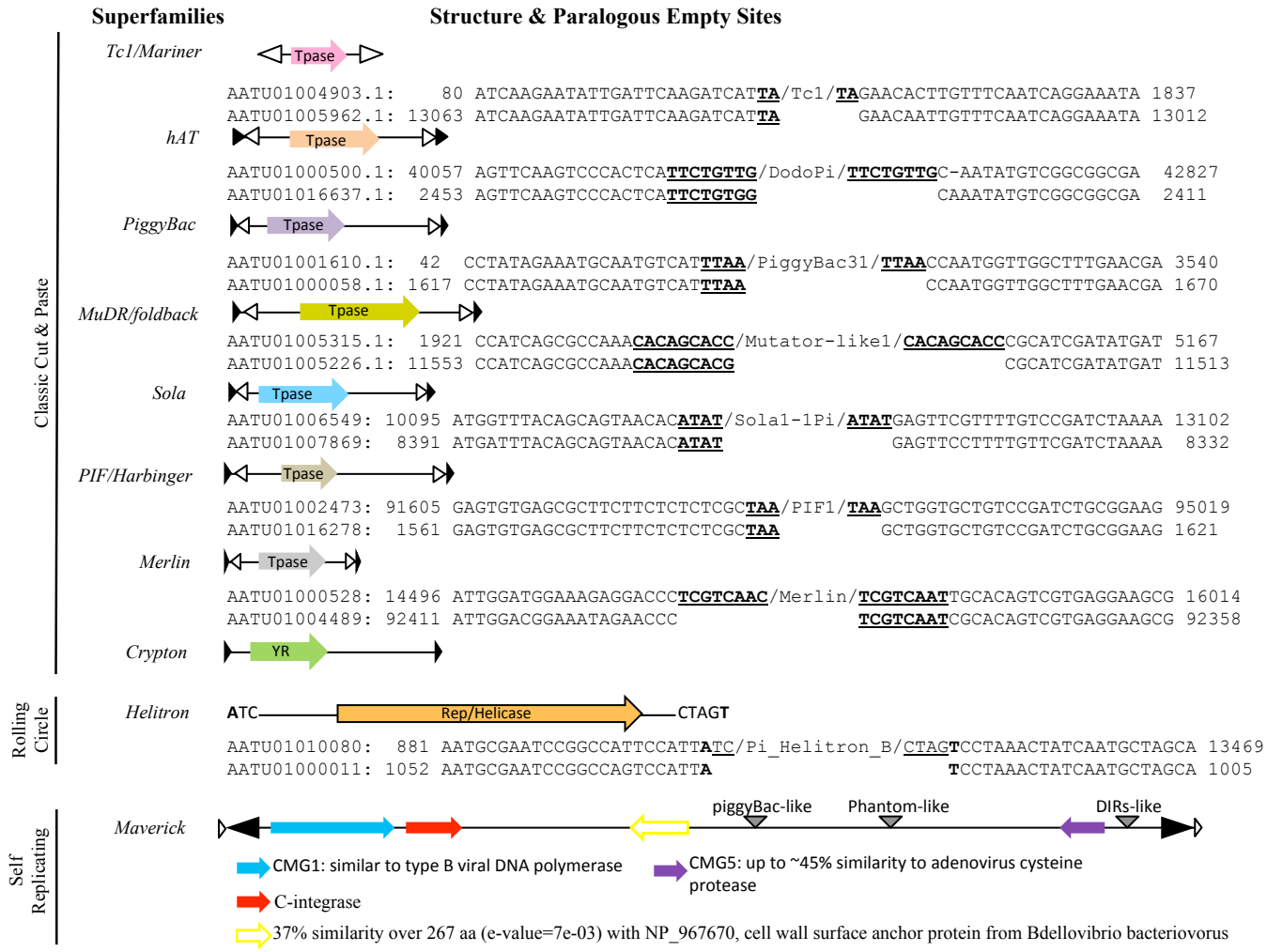


Figure 2.2: Diversity of DNA transposons in *P. infestans*. Direct triangles = target sites duplications (TSDs), inverted black triangles = terminal inverted repeats (TIRs), YR = tyrosine recombinase, Tpase = transposase.

Additionally, we report on a new superfamily of cut and paste DNA elements that have not been previously characterized: *PIF/Harbinger*. There are total of eight families of *PIF* elements that account for 0.391% of the genome. These elements have TIRs ranging from 17-56 bp, induce 3 bp TSD upon insertion and encode a putative transposase typical of *PIF* superfamily (Figure 2.2).

To find evidence for mobility for DNA elements, a series of blastN searches were conducted by fusing ~50 bp upstream and downstream region flanking the TE to identify sites in the genome that are devoid of the TE insertion (see methods). We were able to identify paralogous empty sites for 8 superfamilies of DNA transposons (Figure 2.2). Overall, DNA elements make up approximately 8.9% of the *P. infestans* genome with *Helitrons* and *PiggyBac* accounting for ~2.66% and ~2.81% respectively.

2.3.3 Non-canonical TEs in P. infestans

In the process of annotating the TEs, we found an interesting pattern with a number of elements representing four different superfamilies (*PiggyBac, MULE, Helentron* and *PIF*) that encoded extra genes in addition to the typical transposase gene. Since these elements were 'packed' with atypical ORFs, we named these superfamilies: *PackPiggyBac, PackMULE, PackHelentron* and *PackPIF*.

2.3.3.1 *PackPiggyBac*

We identified four families of *PiggyBac* that encoded additional ORFs. To confirm that these ORFs are incorporated within the boundaries of transposon, we carefully examined the flanking sequences and found terminal inverted repeats ~10-14 bp long that were flanked by TTAA target site duplications characteristic of *PiggyBac*

elements. Hence these families were named, *PiPackPB1, PiPackPB2, PiPackPB3* and *PiPackPB4* (Figure 2.3). To determine the nature of these ORFs conserved domain (CD) and PSI blast searches were conducted (see methods). The first three families harbor proteins that encode an AdoMet domain (CD: cl12011, pfam08123) characteristic of the class 1 S-adenosylmethionine-dependent methyltransferase (referred to as SAM or AdoMet-MTases). *PiPackPB1* encodes ~163 aa SAM that is annotated as gene (PITG_11355) that is 539 nt long with one intron. Likewise, *PiPackPB2* and *PiPackPB3* encode ~ 262aa and ~231 aa SAM that are also annotated as genes, PITG_20021 and PITG_14184 respectively. On the other hand, *PiPackPB4* encodes a protein that is 92% identical over ~292 aa residues to a pleiotropic drug resistance protein (PDR), EEY56124.1 (Figure 2.3). However no putative domains were identified.

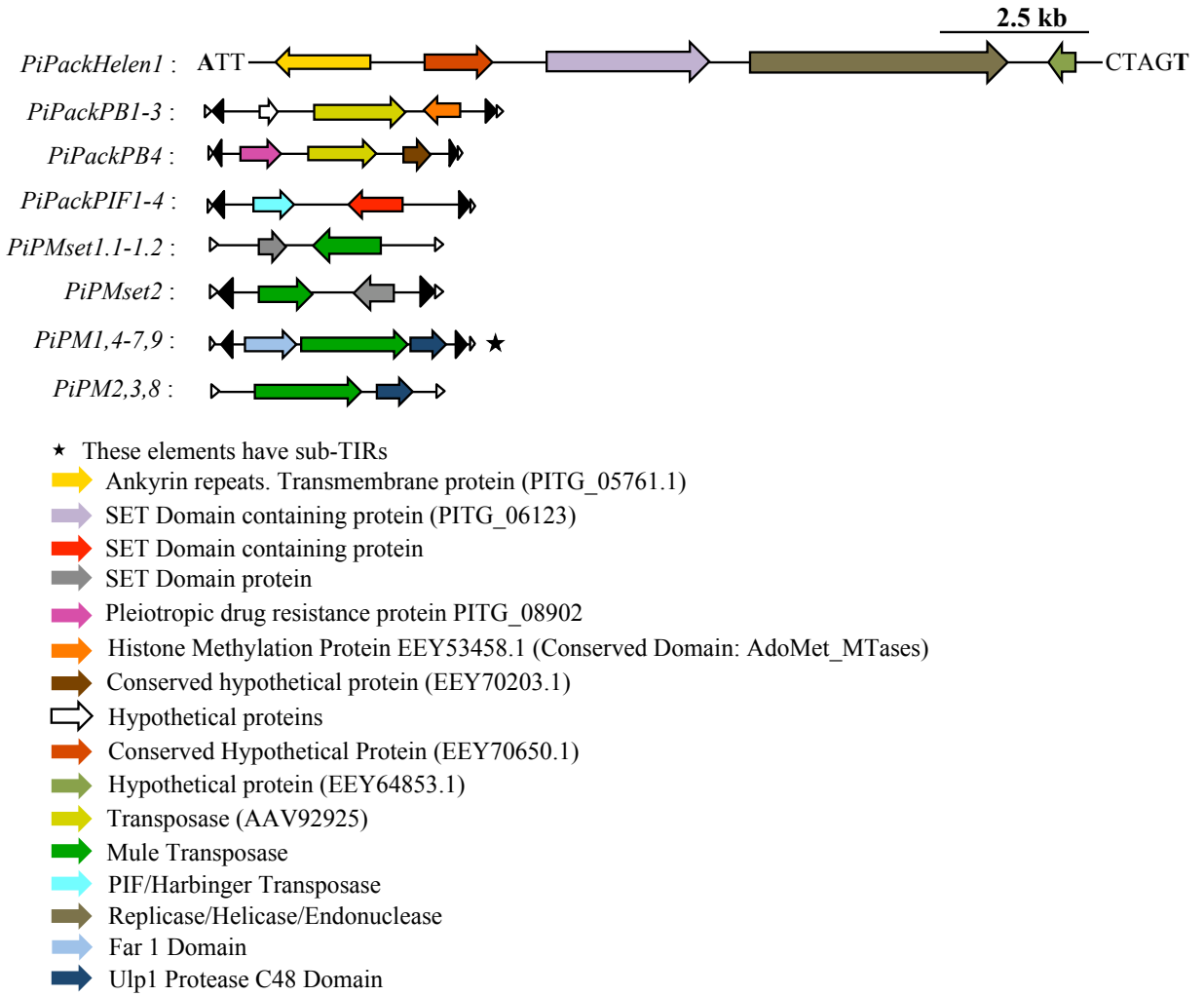


Figure 2.3: Structural features of various transduplicates in *P. infestans*. Direct triangles = target site duplications (TSDs), inverted black triangles = terminal inverted repeats (TIRs).

To identify the parental copy of the gene in the genome, the captured genes were used as query to conduct multiple tblastn searches against *P. infestans* WGS database at NCBI (see methods). We found that SAM is a single copy gene in *P. infestans* (PITG_00145) that shares between 41-48% amino acid identity to the captured genes. We also searched other oomycetes genome to find putative homologs, and found that SAM is present as a single copy in *P. ramorum, P. sojae* and *Hyaloperonospora parasitica* as well. Our phylogenetic analysis suggests that there was a single capture of SAM in past by *PiPackPB1*, that led to diversification and birth of two families, *PiPackPB2* and *PiPackPB3*, resulting in expansion of a single copy gene (Figure 2.4).

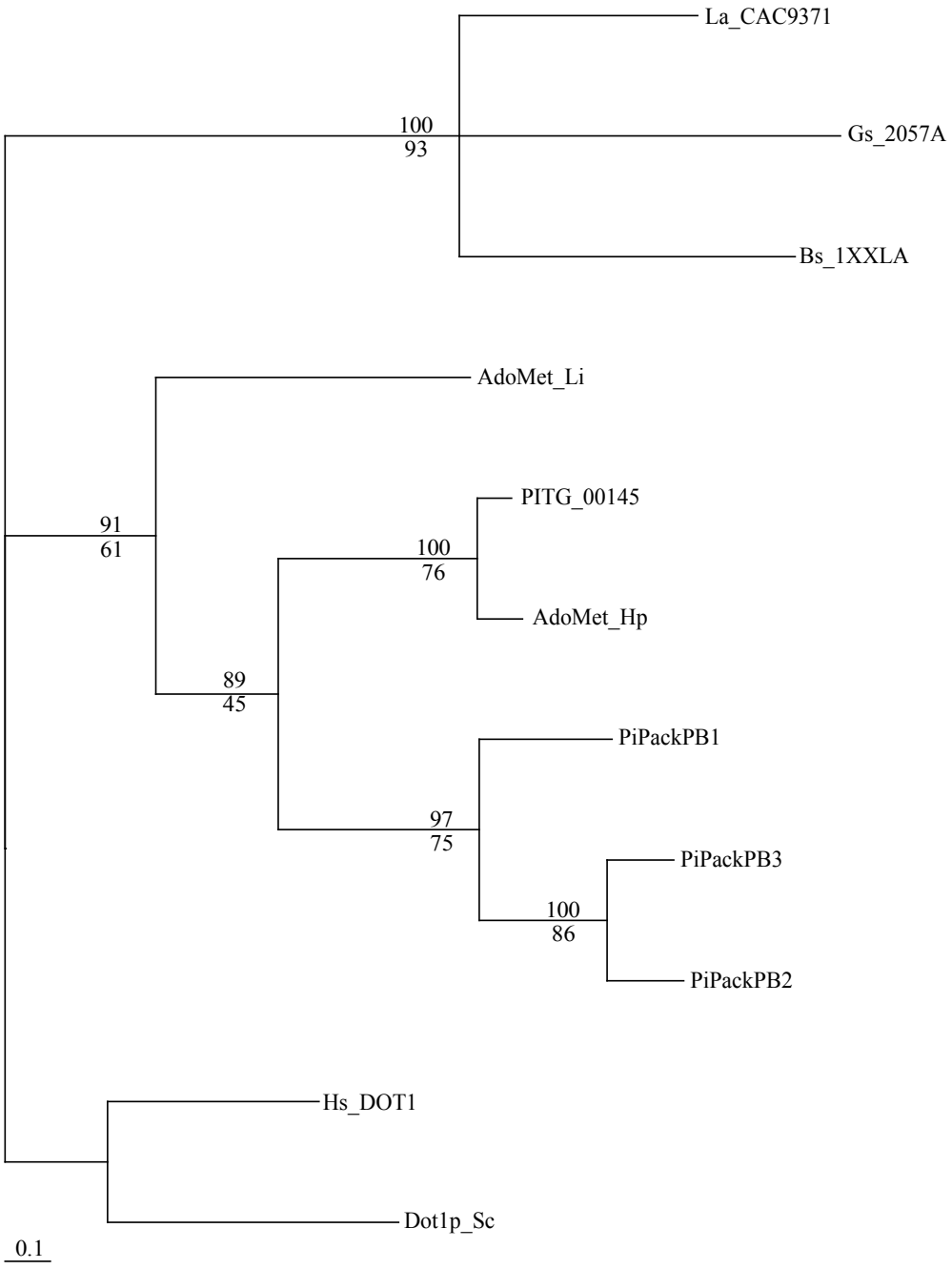


Figure 2.4: Phylogeny of AdoMet-dependent methyltransferases. La = *Lechevalieria aerocolonigenes* (CAC93718.1), Gs = *Galdieria Sulphuraria* (gi:122921433), Bs = *Bacillus Subitilis* (gi:56967300), Sc = *Saccharomyces cerevisiae* (NP_010728.1), Hs = *Homo sapiens* (gi:29726781), Li= *Leishmania infantum* (XP_001468293.1), Hp = *Hyaloperonospora parasitica* (ABWE01000340.1), Pi/PITG= *P. infestans*. Numbers on top =Bayesian posterior probabilities, bottom= maximum-likelihood bootstrap values.

To date this capture event, we surveyed other *Phytophthora* genomes for the presence of these families. We find presence of these elements in *P. ramorum* and *P. sojae* illustrating that the capture of SAM pre dated split of these species (Table 2.1).

Table 2.1: Cross-species analysis of transduplicates in other *Phytophthora* species.

Superfamily	**Genes Captured**	**Pi**	**Ps**	**Pr**
PIF	SET-Domain MTase	+	+	+
PiggyBac	AdoMet-dependent MTase	+	+	+
PiggyBac	Pleiotropic Drug Resistance	+	-	-
Helentron	SET-Domain MTase + Transmembrane	+	+	+
MULEs	SET-Domain MTase	+	+	+
MULEs	Ulp1 Protease	+	+	+

Pi = *Phytophthora infestans*; Ps = *Phytophthora sojae*; Pr = *Phytophthora ramorum*

On the other hand, *PiPackPB4*, has captured the first exon (~861 bp) of a pleiotropic drug resistance gene (PDR), PITG_08902 (Figure 2.5a). This gene fragment shares 95% identity at DNA level to the progenitor copy. Moreover, our cross species analysis shows absence of this family in other *Phytophthora* species demonstrating that this capture event is fairly recent (Table 2.1).

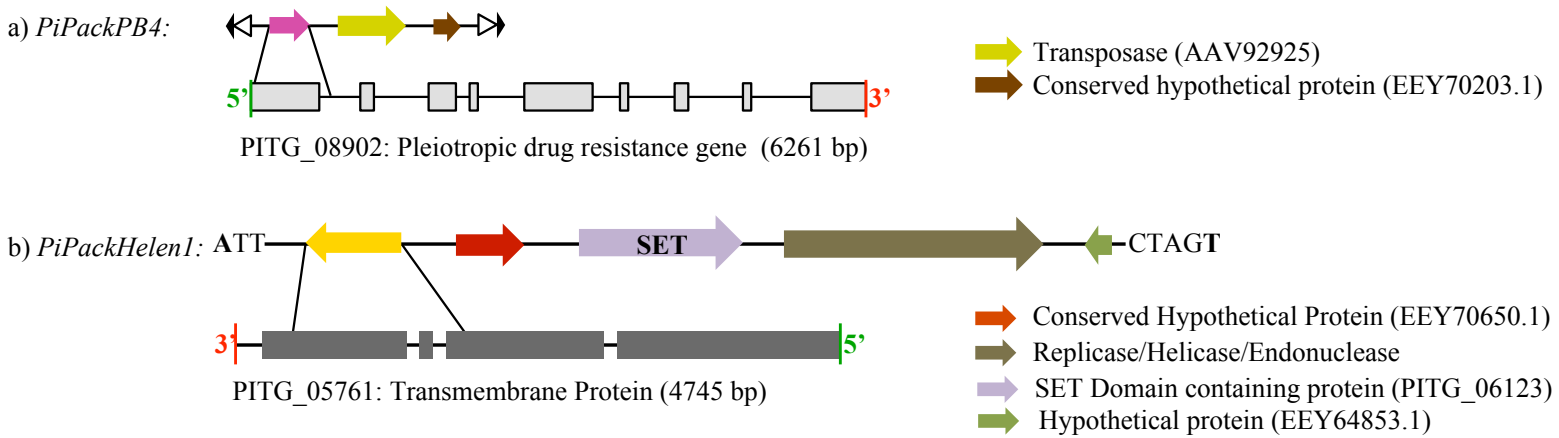


Figure 2.5: a) Capture of pleiotropic drug resistance (PDR) gene fragment by *PiPackPB4*. b) Acquisition of transmembrane gene fragment by *PiPackHelen1*. Boxes = exons, lines connecting boxes = introns.

In order to estimate the copy number and total bp accounted by each family of *PackPiggyBac*, we used RepeatMasker and a perl script (see methods). Overall, all four families occupy 1.4 Mb (~0.77%) of genome (Table 2.2).

Table 2.2: Copy number estimate and total bp count of various families of transduplicates in *P. infestans*.

TE family	Genes Captured	Total bp	Copy Number
PiPackHelen1	Transmembrane, SET-Domain MTase	234557	25
PiPackPB1	AdoMet-dependent Mtase	372595	78
PiPackPB2	AdoMet-dependent Mtase	252763	51
PiPackPB3	AdoMet-dependent Mtase	37487	11
PiPackPB4	Pleiotropic Drug Resistance	800842	195
PiPackPIF1	SET-Domain MTase	136715	33
PiPackPIF2	SET-Domain MTase	56860	32
PiPackPIF3	SET-Domain MTase	108677	28
PiPackPIF4	SET-Domain MTase	101625	30
PiPMset1.1	SET-Domain MTase	263059	90
PiPMset1.2	SET-Domain MTase	325961	129
PiPMset2	SET-Domain MTase	86917	26
PiPM1	FAR1, Ulp1 protease	264049	92
PiPM2	Ulp1 Protease	195285	115
PiPM3	Ulp1 Protease	213515	119
PiPM4	FAR1, Ulp1 protease	143072	79
PiPM5	FAR1, Ulp1 protease	128613	31
PiPM6	FAR1, Ulp1 protease	76682	14
PiPM7	FAR1, Ulp1 protease	452643	118
PiPM8	Ulp1 Protease	159052	45
PiPM9	FAR1, Ulp1 protease	191967	50

To determine the potential functionality of the captured genes, we analyzed protein domains of AdoMet-MTases. These proteins contain five conserved motifs and our protein alignment shows high conservation of generic GxGxG domain within motif I (Figure 2.6).

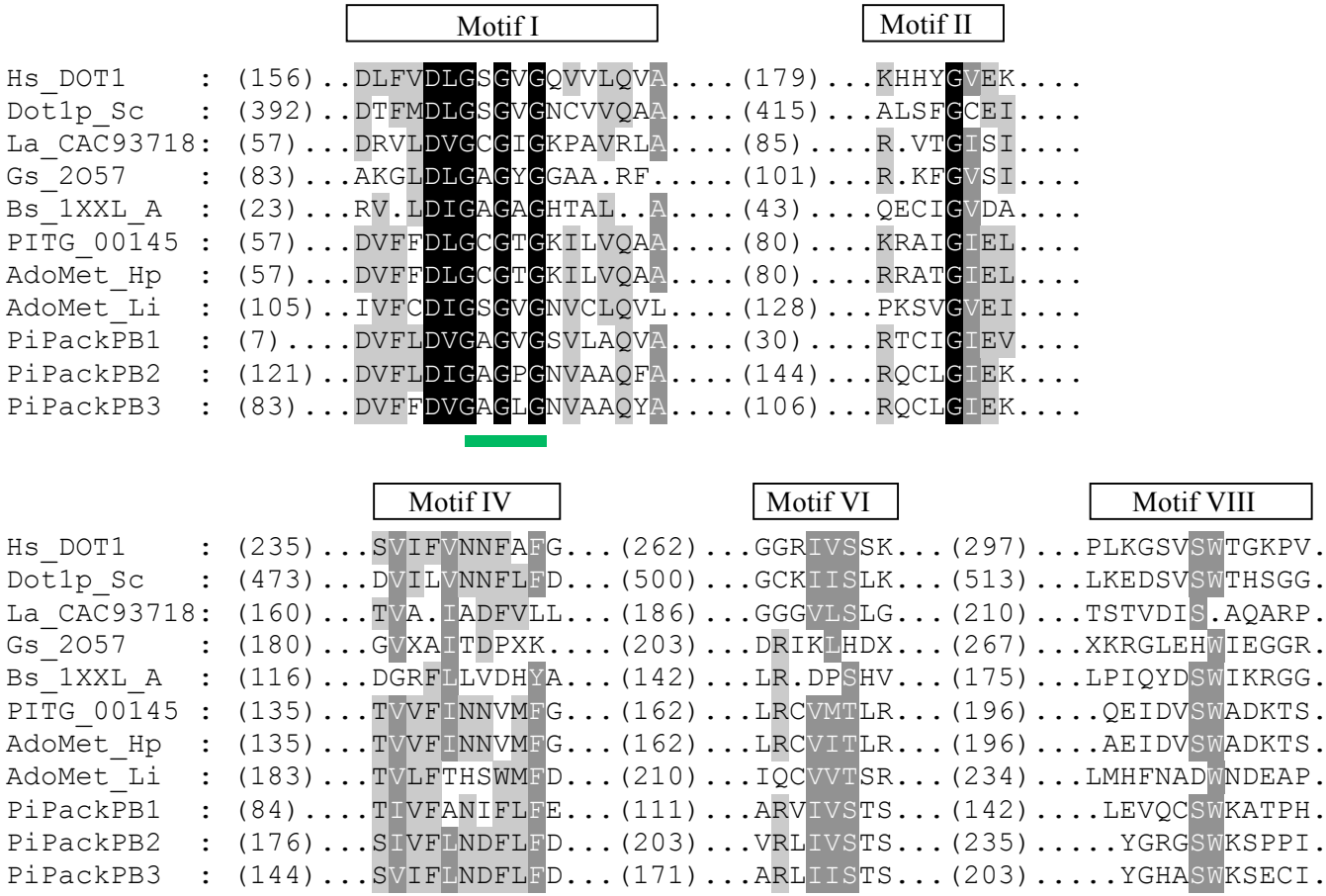


Figure 2.6: Protein alignment of AdoMet-dependent methyltransferases. The boxes on top represent the five important motifs of the protein. Green line indicates GxGxG nucleotide-binding domain. La = *Lechevalieria aerocolonigenes* (CAC93718.1), Gs = *Galdieria Sulphuraria* (gi:122921433), Bs = *Bacillus Subitilis* (gi:56967300), Sc = *Saccharomyces cerevisiae* (NP_010728.1), Hs = *Homo sapiens* (gi:29726781), Li= *Leishmania infantum* (XP_001468293.1), Hp = *Hyaloperonospora parasitica* (ABWE01000340.1), Pi/PITG= *P. infestans*.

This domain is proposed to serve a role in nucleotide binding (Schubert et al. 2003); however further biochemical studies should be undertaken to validate the function.

2.3.3.2 *PackMULEs*

We identified nine families of *MULEs* (*PiPM1-9*) that encoded atypical ORFs besides a transposase gene. To analyze these ORFs, we conducted PSI and CD blast searches. We found that these proteins corresponded to Ulp1 cysteine proteases with C48 peptidase domain. Moreover, six out of nine families harbored an ORF upstream of the putative transposase gene with a FAR1 domain (Figure 2.3). These elements possess ~37 bp long sub-TIRs (with imperfect termini), whereas the elements lacking the ORF that corresponds to the FAR1 domain are TIRless (contain no detectable terminal inverted repeats). To identify the parental copy of the Ulp1 cysteine protease gene, we extensively searched the *P. infestans* genome. In many instances, the annotated genes were actually part of a transposon. However, we found two copies of genes in the genome that were not (PITG_18649, PITG_13913). Furthermore, *Pack-MULEs* with Ulp1 cysteine protease are also documented in plants (Hoen et al. 2006; van Leeuwen et al. 2007). Therefore, to determine the relationship of these plant *PackMULEs* to elements from *P. infestans*, we conducted phylogenetic analysis using bayesian and maximum-likelihood method. Our phylogeny reveals that these elements might represent an ancient clade of *MULE* transposons (Figure 2.7). This is in agreement with our cross species analyses of *P. sojae* and *P. ramorum* genomes where we find presence of these enigmatic elements (Table 2.1). Also another possibility could be that these elements were acquired through horizontal transfer from plants.

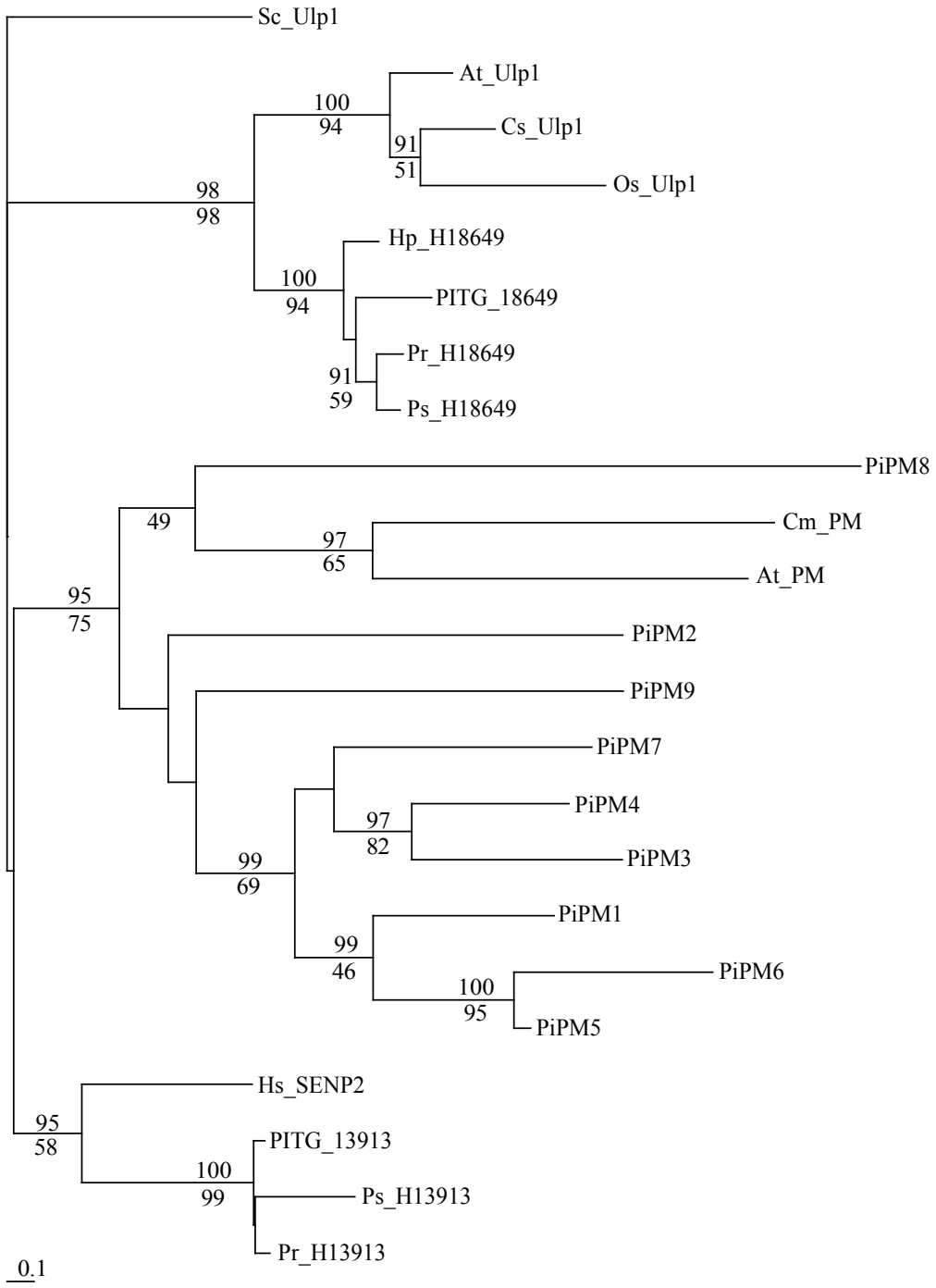


Figure 2.7: Phylogeny of Ulp1 protease. Numbers on top represent Bayesian posterior probabilities and bottom numbers are maximum-likelihood score. Pi/PITG= *P. infestans*, Ps= *P. sojae*, Pr= *P. ramorum*, Hp= *H. parasitica*, At= *Arabidopsis thaliana*, Cs= *Cucumis sativus*, Cm= *Cucumis melo,* Os= *Oryza sativa*, Hs= *Homo* sapiens, Sc= *Saccharomyces cerevisiae*.

Furthermore to determine the functional capacity of Ulp1 cysteine protease, we aligned the captured genes with biochemically studied Ulp1 cysteine proteases. We find that the catalytic residues: H (histidine), D (aspartyl) and C (cysteine) are highly conserved (Figure 2.8). This triad is implicated to form an active site of the protein (Li and Hochstrasser 1999). Together, we show that the captured genes exhibit high degree conservation of catalytic residues and maintenance of essential domains. These results suggest that the captured proteins might be functional.

In the process of mining *MULEs* (**MU**tator **L**ike **E**lements) in *P. infestans*, we found multiple copies in genome that possess an ORF that is predicted to encode a SET-domain protein. This additional ORF is located upstream of an ORF encoding the predicted *MULE* transposase (CAI72252). To confirm that this SET-domain protein is part of the *MULE* transposon, we analyzed the flanking sequences and found 9bp TSD. Using this sequence, we queried the WGS database and were able to retrieve multiple copies in the genome. Thus, we were able to identify three families of *MULEs* that were packed with SET-domain proteins and named them *PiPMset1.1*, *PiPMset1.2* and *PiPMset2* (Figure 2.3). The *PiPMset1.1* and *PiPMset1.2* families are predicted to encode an ORF of ~529 aa SET-domain protein (CAI72344), whereas *PiPMset2* encodes ~134 aa SET-domain protein (PITG_12633). These proteins are very diverse and align only at amino acid level with 35% identity. The first two families do not possess any TIRs; however they induce characteristic 9 bp TSD. On the other hand, *PiPMset2* family has ~24 bp TIRs and induces 9 bp TSD upon insertion.

Figure 2.8: Protein alignment of C48 peptidase domain of Ulp1 protease. The highly conserved residues H, D and C represent the catalytic triad. Pi/PITG= *P. infestans*, Ps= *P. sojae*, Pr= *P. ramorum*, Hp= *H. parasitica*, At= *Arabidopsis thaliana*, Cs= *Cucumis sativus*, Cm= *Cucumis melo,* Os= *Oryza sativa*, Hs= *Homo* sapiens, Sc= *Saccharomyces cerevisiae*.

2.3.3.3 *Pack-PIF*

There are total of eight families of *PIF* elements. Out of the eight families, four are enriched with an ORF (~170 aa) that is predicted to have a SET-domain (CD: cl02566). These families were named *PiPackPIF1, PiPackPIF2, PiPackPIF3* and *PiPackPIF4*. These families are typically 3-5 kb in length, engender 3 bp TSDs and have TIRs ranging from 50-56 bp (Figure 2.3). To determine the nature of this putative SET-domain protein, we conducted blast searches against *P. infestans* genome. We found that the captured ORF is annotated as a putative intronless gene ~ 513 nt (PITG_15246). Moreover, we find transcript evidence for this gene suggesting it is expressed in the genome.

2.3.3.4 *Pack-Helentron*

We identified a family of rolling circle transposon, *Helentron* that had multiple ORFs besides the typical rep/helicase/endonuclease proteins. To determine if these ORFs are incorporated within the boundaries of the *Helentron* element, we examined the flanking sequences. We were able to identify characteristic features like 5'TC and 3'CTAG termini. This family was called *PiPackHelen1* since it was 'packed' with multiple atypical ORFs (Figure 2.3). In order to identify the features of these additional proteins, CD and PSI blast were conducted. We found that one of the ORF containing ~248 aa corresponded to SET-domain protein that was annotated as gene, PITG_06123 (2874 bp, 7 introns). And additional ORF towards the 5' end shared 72% identity over 1,740 nt to transmembrane gene (locus, PITG_05761) (Figure 2.5b). Hence, *PiPackHelen1* has not only captured a whole SET-domain gene but also a

transmembrane gene fragment. *PiPackHelen1* family averages ~14kb in length and is dispersed in 25 copies in the genome (Table 2.2).

Overall, we observed three superfamilies of TEs harboring SET-domain proteins: *Pack-MULEs*, *Pack-PIF* and *Pack-Helentron*. To identify the parental SET-domain gene, we rigorously surveyed the genome using captured SET-domain proteins as query (see methods). We found two copies of SET-domain containing gene in *P. infestans*, PITG_02096 and PITG_13756. To determine the relationship between the captured SET-domains and parental genes, we performed Bayesian and Maximum-likelihood analysis. Our phylogeny demonstrates that there have been multiple, independent capture events of SET-domains by these three superfamilies (*Pack-MULEs, Pack-PIF* and *Pack-Helentron*) (Figure 2.9). After the initial capture these SET domains diverged greatly and hence are distinct from one another. To date these capture events, we surveyed *P. sojae* and *P. ramorum* WGS database at NCBI. We find presence of all three superfamilies encoding SET domain genes (*PackMULEs, PackPIF* and *PackHelentron*) (Table 2.1). This suggests that the capture of SET domain genes predated the split of *P. infestans* from *P. sojae* and *P. ramorum*.

To assess the putative function of captured gene products, we analyzed the SET-domain proteins. Our results demonstrate that they possess both N- and C- terminal regions (SET-N and SET-C). Also, the essential cofactor binding sites in SET-N and SET-C termini (Marmorstein 2003) show high degree of conservation (Figure 2.10). However, to gain insights into substrate binding and assess the methyltransferase activities of SET domains biochemical studies should be supplemented.

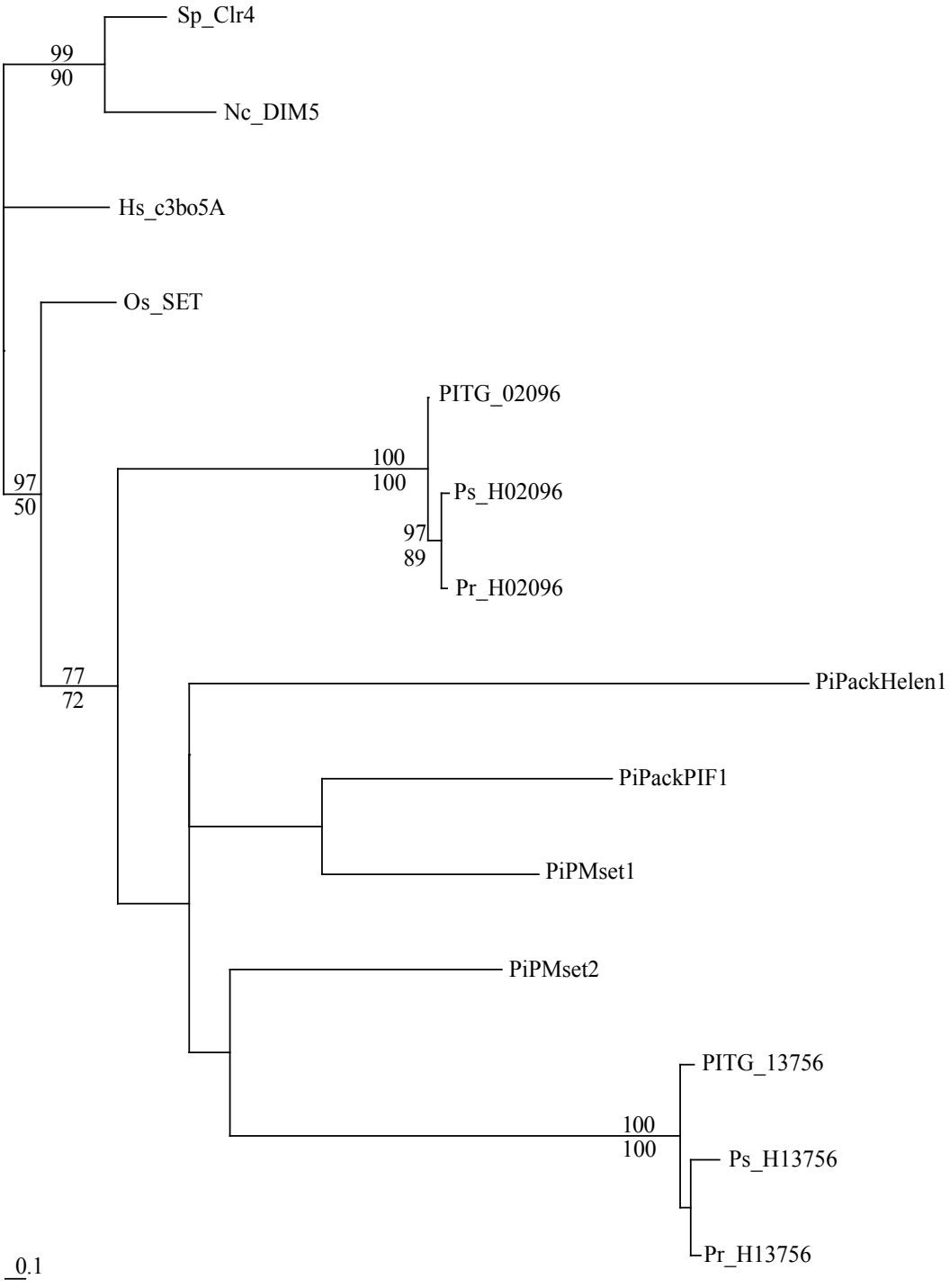
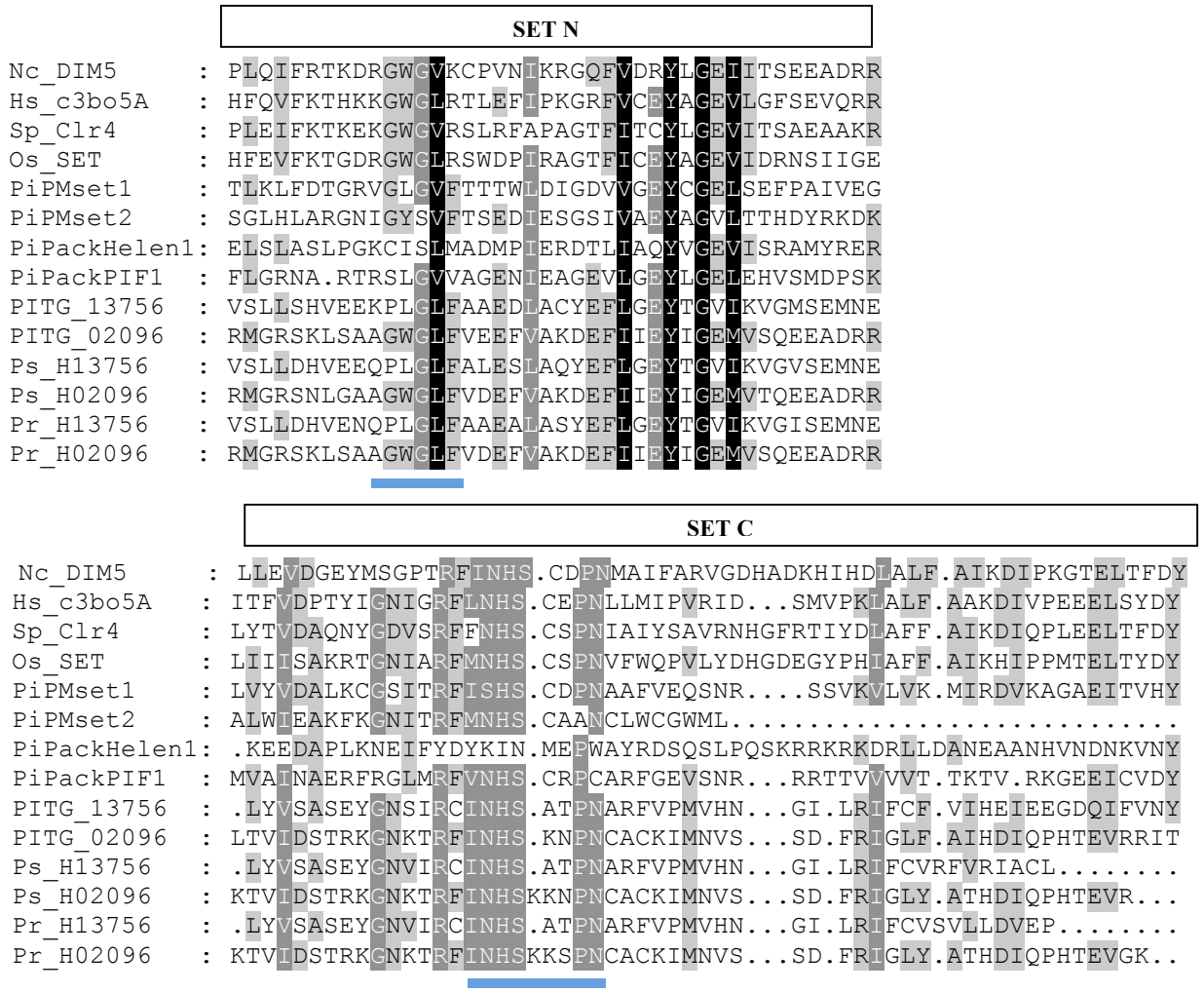


Figure 2.9: Phylogeny of SET-domain proteins. Numbers on top represent Bayesian posterior probabilities and bottom numbers are maximum-likelihood score. Pi/PITG=*P. infestans*, Ps=*P. sojae*, Pr=*P. ramorum*, Os=*Oryza sativa*, Nc=*Neurospora crassa*, Hs=*Homo sapiens*, Sp=*Schizosaccharomyces pombe.*



2.10: Protein alignment of SET-domain. The boxes represent N- and C- terminal domains. Blue lines indicate cofactor binding site. Nc=*Neurospora crassa*, Hs=*Homo sapiens*, Sp=*Schizosaccharomyces pombe,* Os=*Oryza sativa*, Pi/PITG=*P. infestans*, Ps=*P. sojae*, Pr=*P. ramorum.*

2.3.3.5 Evidence of mobility & presence of active elements in the genome

To find evidence of past mobility of these transduplicates, we constructed a chimeric sequence fusing upstream and downstream flanking sequences of TEs. A series of blastN searches were conducted using this sequence to identify sites in the genomes devoid of TE insertion. We found evidence of paralogous empty sites for all four superfamilies (Figure 2.11). This further reiterates that genes and gene fragments are incorporated within the transposon and subsequently mobilized within the genome.

To estimate the age of these non-canonical TEs, we conducted analysis to calculate sequence divergence from consensus. This was done by using repeatmasker generated output (see methods). We find both young, potentially active, elements that are less than three % diverged as well as relatively old elements present in the genome (Figure A1-7).

a) AATU01000913: 23184 AAAGCATAGTAACAATAATTTATTA**TTAA**/*PiPackPB2*/**TTAA**AAGAGCAATATTATCAGTCTGTCAGT 29703
AATU01001183: 20682 AAAGCATAATAACAATAATTTATTA**TTAA** AAGAGCAATAGTATCAGTCTGTCAGT 20746

b) AATU01005398: 991 CGGGTGCTTCATTGGGAAGCGAGGC**TTAA**/*PiPackPB4*/**TTAA**GGGCAGATCCTGCCAAAGTCAAGGCCA 5053
AATU01003476: 3388 CGGGTGCTTCATTGGGAAGCGAGGC**TTAA** GGGCAGATCCTGCCAAAGTCAAGGCCA 3453

c) AATU01002161: 16953 CTGATGATAAGGCACGTAGTATTTTTAAA**A**/*PiPackHelen1*/**T**AAACACTGTTTTGAACCTGCTGGTT 32080
AATU01008688: 6865 CTCAT-ATCAGGCACGTAGTATTTTTTAA**A** **T**AAACACTGTTTTGAACCTGCTGGTT 6926

d) AATU01004555: 44273 TCCCTTTTGTAACGGGGTACAATCC**AGT**/*PiPackPIF1*/**AGT**GTGCCTCAT-GTTACATAAAGAGTGAC 49648
AATU01000605: 80405 ACCATTTTGTAACGGGGTACAATCC**AGT** GTACCCTCAGGTTACGTATAGAGTGAC 80344

e) AATU01002680: 28272 GCATCGGACATAGCCACCCT**CTTGGACAC**/*PiPMset*/**CTTGGACAC**AACCCATTGCTTACACCTCCAT 32010
AATU01004285: 115723 GCATTGGACATAGCCACCCT**CTTGGACAC** AACCCATTGCTTACACCTCCAT 115782

Figure 2.11: Paralogous empty sites of different transduplicates in *P. infestans*.

2.4 Discussion

2.4.1 PIF and PiggyBac make a smashing debut on the transduplicate scene

Our results add to the growing body of evidence that TEs are important contributors to the genome evolution despite deemed as "junk" DNA. One of the spectacular ability of TEs is their capacity to capture and mobilize genes thereby leading to genetic variability. To date, the transduplicate scene was dominated by rolling circle *Helitrons*, cut and paste *CACTA* and *MULE* elements (Kawasaki and Nitasaka 2004; Lal and Hannah 2005; Hanada et al. 2009). Therefore, it was not surprising to have encountered *MULEs* and *Helentrons* harboring gene fragments in our extensive survey of *P. infestans* genome. More surprising is the discovery of *PiggyBac* and *PIF* transposons enriched with genes suggesting that transduplication might be an intrinsic ability of TEs and not restricted to only few TE superfamilies. Though the mechanism of gene capture still remains to be elucidated, this process most likely involves capture at DNA level because of observed retention of introns in some of the captured genes.

2.4.2 What are the driving forces behind abduction of host genes?

Previous studies of transduplicates have reported captures of gene fragments as opposed to complete genes (Jiang et al. 2004; Gupta et al. 2005; Zabala and Vodkin 2005 & 2007). In this study, we demonstrate that TEs have not only captured gene fragments but have also abducted whole copies of genes. Moreover, we see a strong preference for acquisition of genes that are involved in epigenetic regulation like DNA methyltransferases (MTases). AdoMet-MTases (class 1 DNA MTases) captured by

PiggyBac in *P. infestans* shares similarity to biochemically studied AdoMet-MTases from other organisms where they serve a wide array of functions including signal transduction, regulation of chromatin, gene silencing etc (Schubert et al. 2003). From the amino acid sequence analysis, we report the maintenance of critical domains suggesting that these proteins in *P. infestans* might be serving similar function in methylating DNA and impacting expression of genes, however further biochemical analysis needs to be supplemented. On the other hand, *PIF*, *MULEs* and *Helentrons* have captured class 2 DNA MTases that possess characteristic SET-domain. SET-domain proteins methylate lysines at various positions (4, 9, 27 and 36) on histones 3 and 4 (Sawada et al. 2004). Our phylogenetic analysis shows that there have been multiple, independent capture events of SET-domain MTases in past by these TEs. One probable explanation of this scenario is that the capture of these DNA methyltransferases allows access to heterochromatin or provides chromatin remodeling thereby ensuring long term survival of TEs and less profound effect on host by avoiding gene dense regions. Another possibility that could have facilitated capture of host genes is the inherent competition between TEs in the genome. It is known that retrotransposon, copy and paste, have attained extremely high copy number and account for ~38% of the genome (Haas et al. 2009). Therefore, this competition for space among TEs could have encouraged the capture of host genes, like DNA MTases, to enable TEs to persist in the genome for a longer period.

2.4.3 Evolutionary implications of massive scale transduplication: a boon or a curse

Our results show that SAM (AdoMet-MTases) is a single copy gene that was captured by *PiPackPB1* family thereby leading to expansion of this gene and also birth of *PiPackPB2* and *PiPackPB3*. Moreover, the sequence analysis of captured gene to its progenitor gene shows a high degree of divergence demonstrating that capture by the TEs not only resulted in expansion but also diversification of gene. We also show transcript evidence of the captured AdoMet-MTases demonstrating that they are transcriptionally expressed suggesting a possible host function. It is not clear what the consequences on the host might be if the putative proteins retain similar function. It is possible that the putative proteins if translated have acquired slightly different function.

Other examples of transduplicates involved the capture of first exon of pleiotropic drug resistance gene (PDR) by *PiPackPB4*. PDR genes belong to the superfamily of ABC transporters that play a critical role in plant pathogenesis (Haas et al. 2009). There are about 156 members of ABC transporters annotated by the genome analysis. Due to the high sequence identity (95% with retention of 5' splice site) to the progenitor gene, the presence of young elements in the genome, and absence of this family in other *Phytophthora* species (*P. sojae and P. ramorum)*, we predict this capture event was very recent. The implication of this exon shuffling is not known, however, we speculate that it could potentially mediate the formation of novel genes. Another spectacular example is the abduction of Ulp1 protease by *PackMULEs*. Ulp1 proteases fall into family of cysteine protease that represents one of the six major classes of proteases. These proteases have been known to have implication in post translational

modifications of cellular proteins and represent one of the important mechanisms involved in plant-pathogen interaction (Avrova et al. 1999; Xia 2004). In *Phytophthora* species, there is an extensive repertoire of cysteine proteases and it remains one of the important protein families involved in plant pathogenesis (Haas et al. 2009). This expansion of cysteine proteases might be attributed to the activity of *PackMULEs* that have captured and subsequently amplified to high copy number in the genome. Also, our phylogenetic analysis shows diversification of Ulp1 proteases, demonstrating rapid evolution of this critical gene family involved in plant infection. This might in turn provide the pathogen with an arsenal to quickly develop immunity to plant defenses. Moreover, these Ulp1 proteases might be serving a potential proteolytic function in the lifecycle of *PackMULEs* therefore benefiting the TE as well.

It is clear that TEs account for a major portion of repetitive repertoire of *P. infestans* and their proliferation and persistence has helped shaped the genome architecture from time to time. Besides sculpting genomes with their proliferations, TEs can donate genes to the host in a process termed "molecular domestication or exaptation". This phenomenon has been very well documented whereby the TE related protein has adopted novel function in host (Feschotte and Pritham 2007). Hence, the TE-mediated capture, diversification and expansion of genes followed by domestication would in fact enrich and replenish the protein-coding repertoire of *P. infestans* thereby potentially giving it a heads up in the evolutionary arms race with the plant host. Undoubtedly, TEs have proven to be a rich source of genetic material and have had a major impact in shaping the evolutionary trajectory of this insidious parasite.

Therefore, it is tempting to speculate that the dynamic portion of the genome holds a key to shut down this phytopathogen that poses a significant threat to potato and tomato crops worldwide.

CHAPTER 3

PACK-HATS, A NOVEL FAMILY OF TRANSPOSABLE ELEMENTS IN THE UNICELLULAR PARASITE, *PHYTOPHTHORA RAMORUM*.

3.1 Introduction

The genus *Phytophthora,* a lineage that evolved independent of fungi, harbors some notorious plant pathogens that pose great threat to crops worldwide (Sogin and Silberman 1998). The interesting feature of these phytopathogens is their exquisite ability to quickly adapt to their ever-changing environment, leading to serious agricultural and environmental challenge. The year 2006 marked release of one of the *Phytophthora* species genome sequences, *P. ramorum*, causal behind sudden oak death and ramorum blight in woody ornamentals (Tyler et al. 2006). The wealth of data from this genome-sequencing project brought to light that about 28% of its 65 Mb genome is composed of repetitive DNA (Haas et al. 2009). Our analysis of the repetitive repertoire showed that transposable elements, mobile genetic entities, make up ~ 22% of *P. ramorum* genome (data now shown; unpublished).

Transposable elements are dynamic in nature and make up significant fraction of many eukaryotic genomes. Their movement and proliferation in the genome can impart negative fitness effects on the host. Despite being viewed as potential mutagens, there is compelling evidence demonstrating that TEs can be a source of raw genetic material leading to genome innovation. In recent times, one mechanism that has garnered a lot

of curiosity is transduplication, a process whereby host genes are captured by transposons. This mechanism is highly intriguing considering the propensity of TEs to mobilize the captured gene fragments leading to diversification and expansion of host genes. So far the TEs that have been known to transduce gene fragments include rolling circle *Helitrons*, classic cut and paste *CACTA* and *MULE* elements (Kawasaki and Nitasaka 2004; Lal and Hannah 2005; Hanada et al. 2009). There is mounting evidence suggesting that transduplication occurs on a massive scale and is extremely rampant in many plant genomes (Jiang et al. 2004; Feschotte and Pritham 2009).

We employed computational tools to analyze and study the dynamics of TE populations in genome evolution of *P. ramorum*. Here, we report a family of the classic cut and paste DNA transposons, *hAT* superfamily, 'packed' with transglutaminase elicitor gene fragments (TGase) that we named *Pack-hAT*. We present on the structure and abundance of these elements and the role they may have had in shaping the evolutionary trajectory of this pathogen.

3.2 Material and Methods

3.2.1 Mining of Pack-hATs in P. ramorum

Pack-hATs were fortuitously discovered while annotating *P. ramorum* repeat library. The initial query (R # 95) was used to conduct blastn searches to retrieve more copies in the *P. ramorum* genome. Thereafter, a series of blastn searches were conducted by fusing ~50 bp upstream and downstream region flanking the transposon to identify any potential paralogous (within genome) empty sites. An empty site was

annotated when another region is identified in the same genome that lacks the insertion yet contains the unduplicated target site.

3.2.2 Sequence analysis of transglutaminase elicitor gene fragments and hAT transposase

Pack-hAT sequences were translated in six reading frames using online expasy translate tool (http://ca.expasy.org/tools/dna.html). Thereafter, TGase elicitor sequences and *hAT* transposases were aligned using ClustalW (http://www.ebi.ac.uk/Tools/clustalw/) and alignments were refined manually using GeneDoc version 2.6.02 (Larkin et al. 2007, Nicholas et al. 1997).

3.2.3 Identification of parental copy of gene

Tblastn searches were conducted using TGase elicitor sequence from the *Pack-hATs*. The flanking sequences of the resulting significant hits (e value < 10^-4) were exhaustively examined to identify any structural features (TIRs, TSDs) characteristic of *hAT* superfamily. If no such structural characteristics were identified then these entries were classified as a progenitor TGase elicitor copy.

3.2.4 Cross species analysis for presence of Pack-hATs

Tblastn and blastn searches were conducted using *Pack-hAT* nucleotide and protein query respectively to query *Phytophthora sojae* and *Phytophthora infestans* genome at NCBI (WGS data set). Each significant hit was inspected to identify protein domains and structural features similar to *hAT* superfamily.

3.3 Results

3.3.1 Discovery of Pack-hATs in P. ramorum

Manual annotation using blastx search of the *P. ramorum* repeat library revealed a repeat (#95) that yielded hits to a transglutaminase elicitor (TGase) with 71% identity over stretch of 51 amino acid residues (e-value 2e-16). Series of blastn searches against *P. ramorum* revealed that there were multiple copies in the genomes. Close inspection of these hits led to the discovery of an ORF corresponding to a *hAT* transposase downstream of TGase elicitor protein. To determine if this TGase elicitor fragment is part of the *hAT* transposon, we examined the flanking sequences to check for the presence of structural boundaries like TIRs and TSDs. Together, this repeat was flanked by 17-18 bp terminal inverted repeats (TIRs) and target site duplications (TSDs) characteristic of the *hAT* superfamily of DNA transposons. Since, these *hAT* elements were 'packed' with TGase elicitor gene fragments we named them *Pack-hATs*. To verify the structure, blastn search was conducted by fusing the sequence 50 bp upstream and downstream of the TIRs, including one of the TSD, to identify an empty site. This search revealed a paralogous site, devoid of the transposon in the genome illustrating the past mobility of these elements and confirming that the TGase elicitor fragments are indeed part of *Pack-hATs* (Figure 3.1). Within the *Pack-hAT* family, we observe a spectrum of elements: harboring both TGase elicitor and *hAT* transposase fragments (*Pack-hAT1.1*, *Pack-hAT1.2, and Pack-hAT1.4*), elements with TGase elicitor fragments (*Pack-hAT1.3*) and elements containing *hAT* transposase fragments only (*Pack-hAT1.5* and *Pack-hAT1.6*).

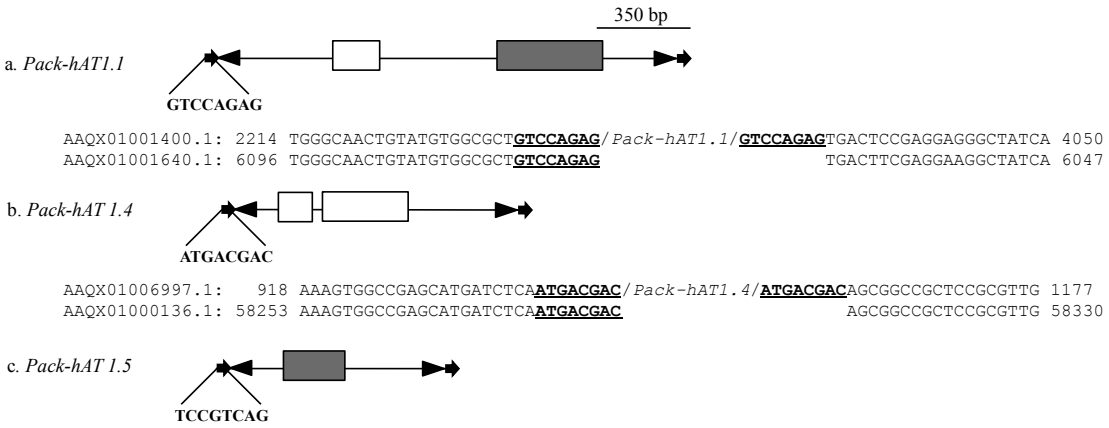


Figure 3.1: Structure of *Pack-hATs* in *P. ramorum*. Direct black arrows represent the target site duplications (TSDs), whereas inverted black arrows illustrate the terminal inverted repeats (TIRs). Grey box = *hAT* dimerization domain, white box = TGase elicitor gene fragments.

Overall, *Pack-hATs* range from 858-1982 bp in length, possess 17-18 bp TIRs that are well conserved (Figure 3.2), and induce unique 8 bp target site duplications.

```
PackhAT1.1:  TAGAGGTGGCGACGTTA.
PackhAT1.2:  TAGAGGTGGCGACGTTAC
PackhAT1.3:  TAGAGGTGGCGACGTTAC
PackhAT1.4:  TAGGGGTGGCGACGTTAC
PackhAT1.5:  TAGGGGTGGCGACGTTA.
PackhAT1.6:  TAGAGGTGGCGACGTTA.
```

Figure 3.2: Alignment of *Pack-hATs* terminal inverted repeats (TIRs).

To estimate the copy number, we used blast tools to extensively search *P. ramorum* genome deposited in WGS database at NCBI. We identified a total of 27 elements that comprised of some full length and partial copies in the genome (Table 3.1).

Table 3.1: Copy number estimate of *Pack-hATs* in *P. ramorum*.

***Pack-hAT* family**	**Copy Number**
Pack-hATs (TGase elicitor and transposase)	7
Pack-hATs (only TGase elicitor)	14
Pack-hATs (only transposase)	6

This copy number estimate also includes truncated elements with 5' and 3'ends.

3.3.2 Analysis of captured transglutaminase elicitor gene fragments

To identify the progenitor gene, we surveyed *P. ramorum* genome sequences using *Pack-hAT* TGase elicitor fragment as a query (see methods). The *P. ramorum* TGase elicitor gene is about 2327 bp long and contains no introns. The captured gene fragments vary in size from 319-537 nt (Figure 3.3). Moreover, these captured fragments share 52-64% identity at the amino acid level and 70-74% identity at nucleotide level to the parental gene.

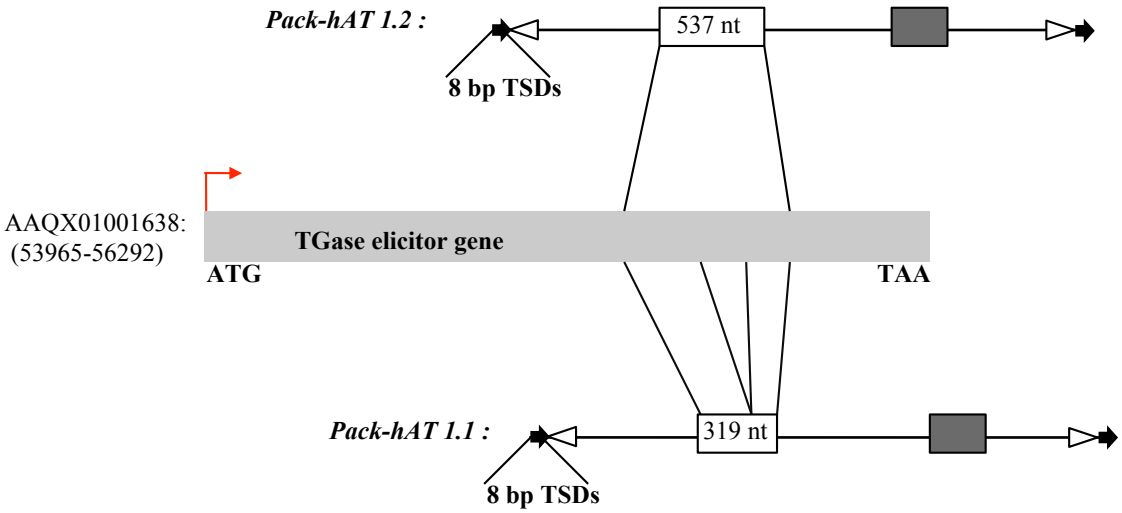


Figure 3.3: Capture of TGase elicitor gene fragments by *Pack-hATs*. Numbers in the white box represent the size of gene fragment captured. Red arrow indicates the start site. Inverted black arrows = terminal inverted repeats, direct arrows are target site duplications (TSDs). Grey box = *hAT* dimerization domain.

TGase elicitor belongs to the elicitor-like gene family in *Phytophthora* species. Biochemical studies have revealed that the 13 aa residue domain, pep-13, in TGase elicitor protein is essential for elicitor function and helps initiate defense responses in hosts (Brunner et al. 2002). Comparison of *Pack-hAT* gene fragments to known TGase elicitor proteins from other *Phytophthora* species demonstrated the absence of pep-13 domain (Figure 3.4). Moreover, TGase elicitor fragments in *Pack-hATs* are frequently interrupted by stop codons and lack the pep-13 domain of TGase elicitors, thereby suggesting that no functional proteins are likely produced.

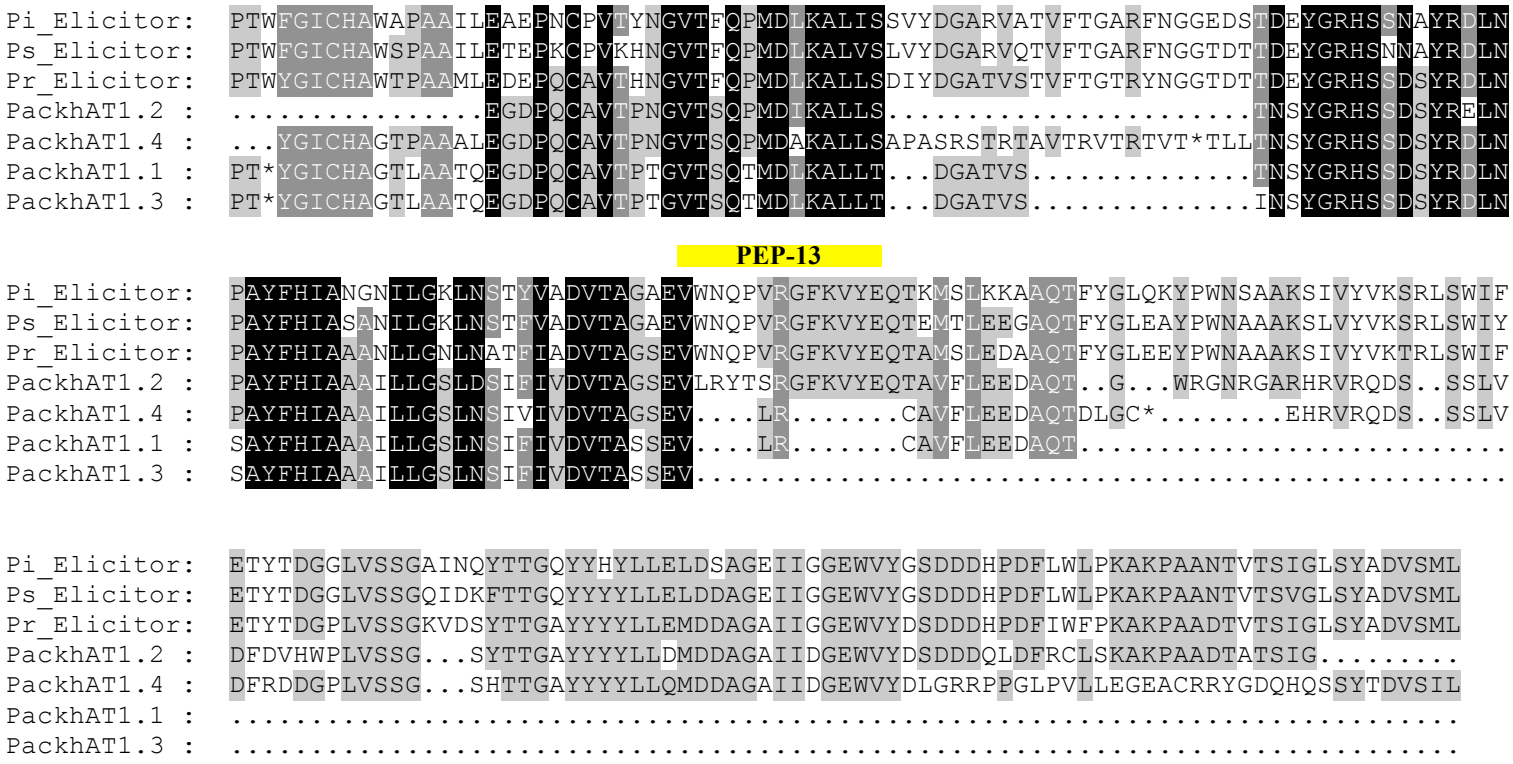


Figure 3.4: Sequence analysis of TGase elicitor gene fragments of *Pack-hATs*. Pi=*Phytophthora infestans*; Ps=*Phytophthora sojae*; Pr=*Phytophthora ramorum*. Asterisk represents stop codons.

3.3.3 Transpositional capacity of hAT transposase

To determine the transposase coding capacity of these elements, we carefully examined the transposase (tpase) protein. These elements possess only a fragment of tpase, which is frequently interrupted by pre mature stop codons. This ~95 aa protein corresponds to *hAT*-dimerization domain of tpase. This dimerization domain, located near the C terminus of the protein, is known to form an integral part of transposases belonging to the *hAT* superfamily (Essers et al. 2000). It has been predicted that some of the most highly conserved residues are located in the regions that are more likely to assume a helical conformation. Site directed mutagenesis have confirmed that the amino acids in the helical regions are involved in the formation of tpase dimer (Essers et al. 2000). We aligned the dimerization domain from the *Pack-hATs* to known tpases like *Ac* (maize), *Tam3* (snapdragon), *Hermes* (*Musca domestica*) and *Hobo* (*Drosophila melanogaster*). The signature motif WWxxxxxxxPxLxxxAxxxL of *hAT* superfamily of tpases is somewhat conserved among *Pack-hATs* (Figure 3.5). However, the first W residue of this motif is replaced by a stop codon in *Pack-hATs*. Therefore, we suggest that these are non-autonomous elements that do not possess any coding capacity and might have subsequently lost it due to persistence in genome for a long time.

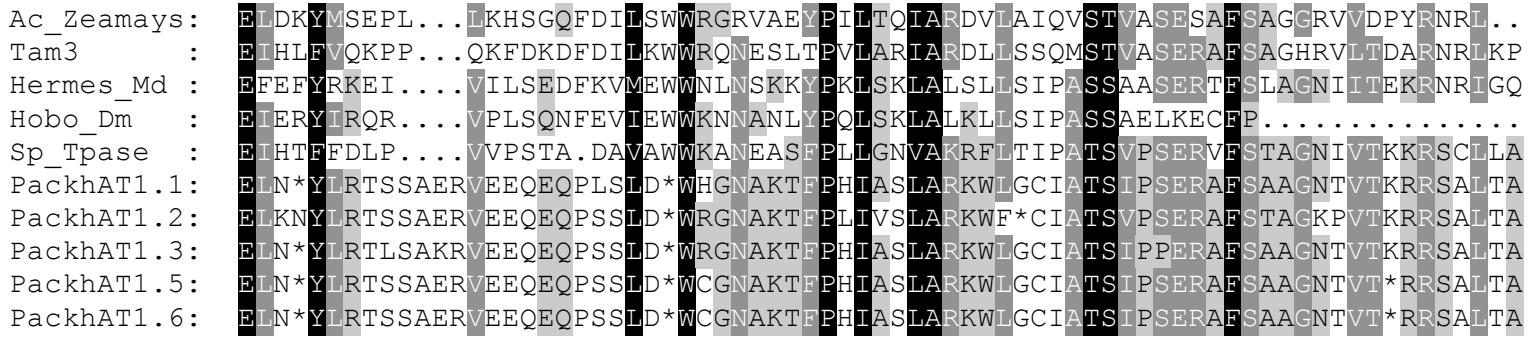


Figure 3.5: Alignment of the conserved *hAT* dimerization domain of *Pack-hATs* to other known *hAT* tpases. The * represent pre mature stop codons in the reading frames. Ac_Zeamays (CAA29005.1); Tam3 = *Antirrhinum majus* (CAA38906.1); Hermes_Md = *Musca domestica* (AAC37217.1); Hobo_Dm= *Drosophila melanogaster* (CAA28410.1); Sp_Tpase= *Strongylocentrotus purpuratus* (XP_001191156.1).

3.3.4 Absence of Pack-hATs in other Phytophthora species

To assess the presence of *Pack-hATs* in other *Phytophthora* species, blastn and tblasntn searches were conducted to query WGS database at NCBI. The searches were limited to *Phytophthora* genus (taxid: 4783). No significant hits revealing the presence of *Pack-hATs* elements were found in the other *Phytophthora* species. These results indicate that *Pack-hATs* are unique to *P. ramorum* and the capture of TGase elicitor gene fragments likely occurred after the split of *P. ramorum* from other *Phytophthora* species.

3.4 Discussion

3.4.1 Gene capture mediated by hAT superfamily

Here we describe TEs of the *hAT* superfamily that are carrying TGase elicitor gene fragments. We call this family *Pack-hATs*. There are many genus and species-specific elicitor that are essential in the life cycle of *Phytophthora* species. The abundance of cell wall TGase elicitor might have facilitated the capture of gene fragments by *hAT* transposons in *P. ramorum*. Another possibility could be that the transposons picked up TGase elicitor gene essential to host to ensure survival in the genome. There is mounting evidence showing that many plant pathogens secrete unique molecules, called elicitors that elicit defense responses in plants (Jiang et al. 2005). Many of these elicitor-like proteins exhibit TGase activity, where only a 13-aa long pep domain is sufficient to exhibit elicitor function (Brunner et al. 2002). Moreover, the captured gene fragments appear to be frequently interrupted by stop codons, and could possibly be non-functional. However, it could be that they have

evolved some novel function that no longer requires the maintenance of coding capacity as the progenitor gene. Therefore, further study to evaluate the functional constraint on captured gene fragments should be undertaken.

3.4.2 Evolutionary implication of Pack-hATs

Pack-hAT elements lack a functional tpase for transposition. In almost all cases, we find a highly mutated *hAT* dimerization domain, ~95 aa long. It could possibly be that these elements lost the coding capacity during double stranded break repair after excision. Moreover, there are only handfuls of full length elements in the genome, the rest have either truncated 5' or 3' end possibly due to DNA deletions. It is indicative that *Pack-hAT* family is relatively old and has acquired mutations over evolutionary period. Also, *Pack-hAT* family is not present in other *Phytophthora* species where the genome sequences are available (*P. infestans* and *P. sojae*). Therefore, the acquisition of TGase elicitor gene fragments probably occurred after the split of *P. ramorum* from these species. The implication of this capture is unknown however it could have potentially contributed to the elicitor repertoire of *P. ramorum*.

APPENDIX A

SUPPLEMENTARY INFORMATION - CHAPTER 2

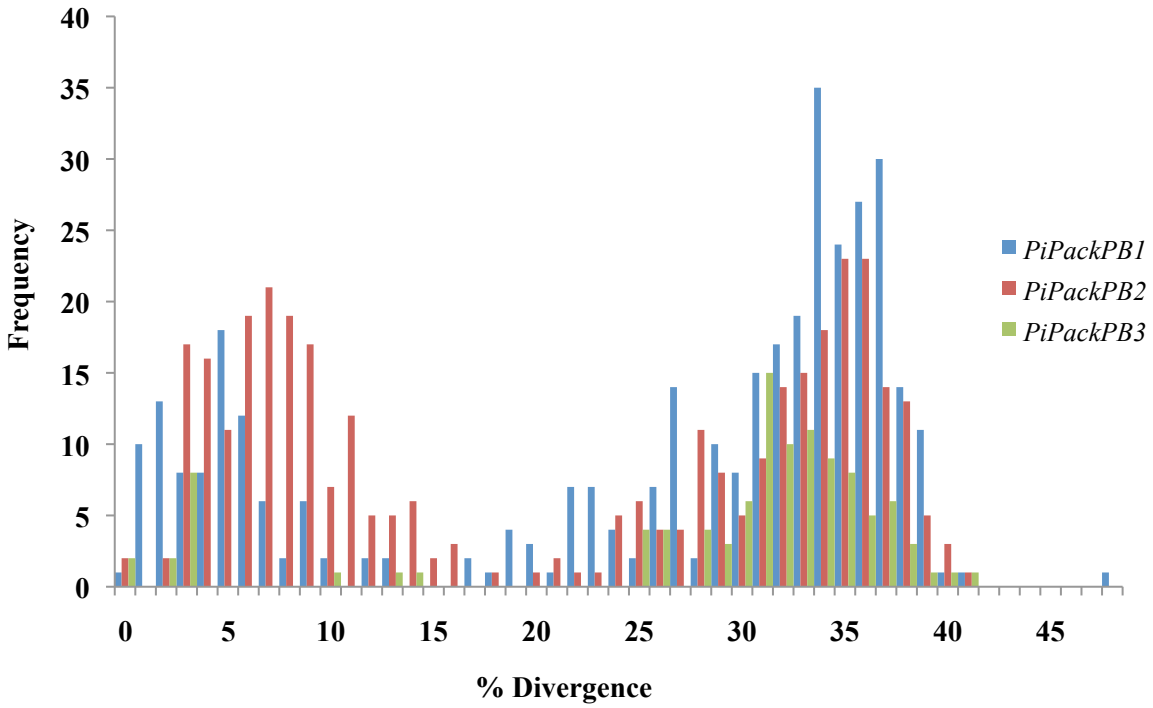


Figure A1: Sequence divergence to consensus of *PiPackPB1, PiPackPB2, and PiPackPB3* family with AdoMet-dependent methyltransferases.

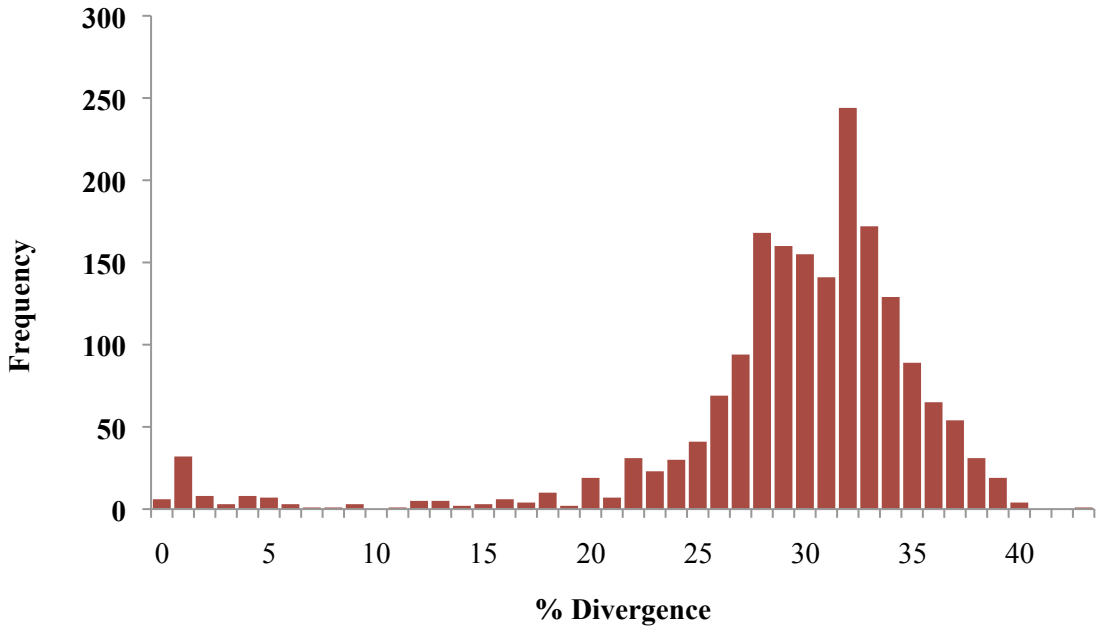


Figure A2: Sequence divergence to consensus of *PiPackPB4* family.

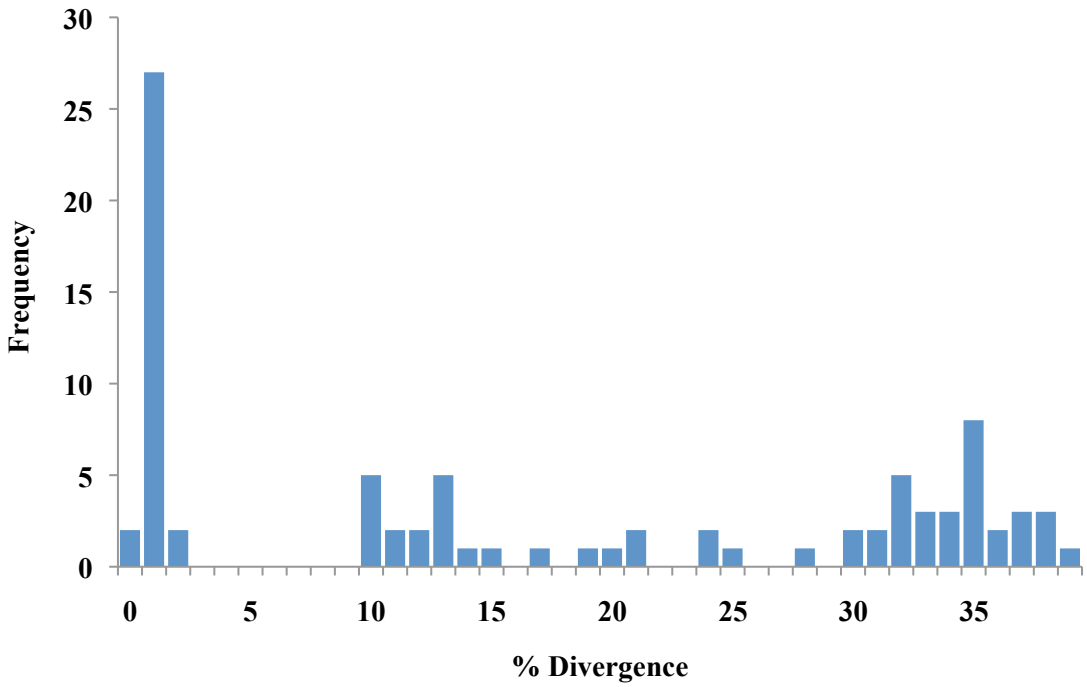


Figure A3: Sequence divergence to consensus of *PackHelen1*.

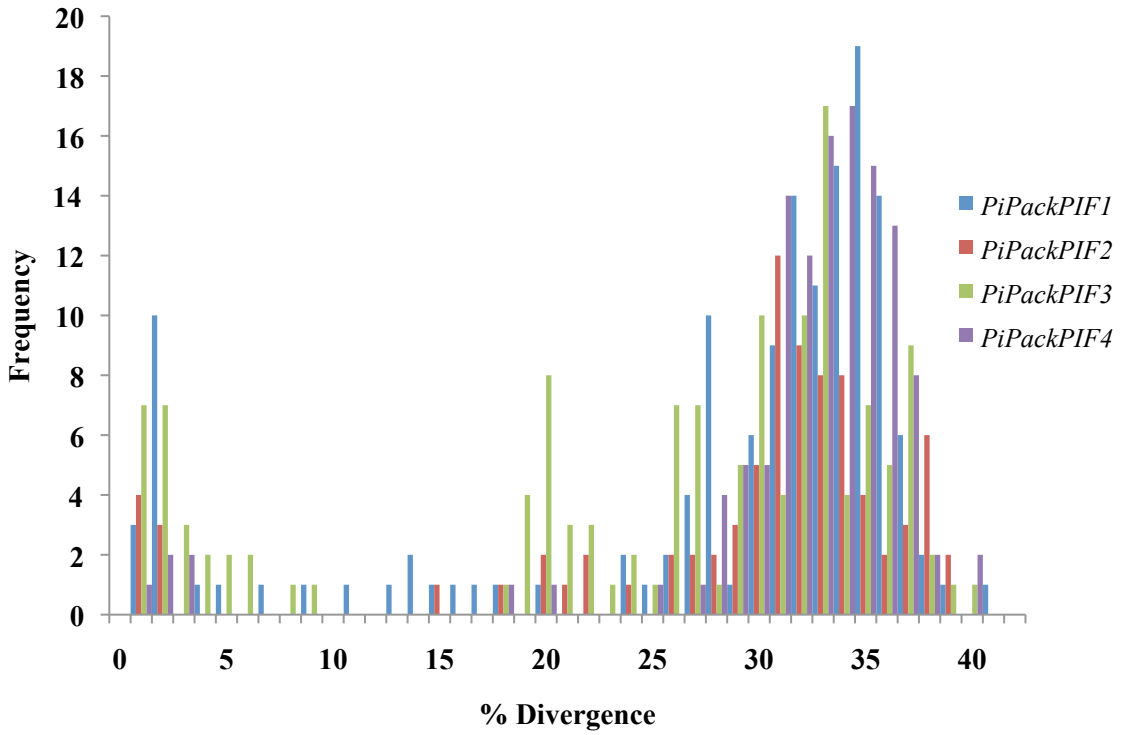


Figure A4: Sequence divergence of *PiPackPIF1, PiPackPIF2, PiPackPIF3* and *PiPackPIF4* families.

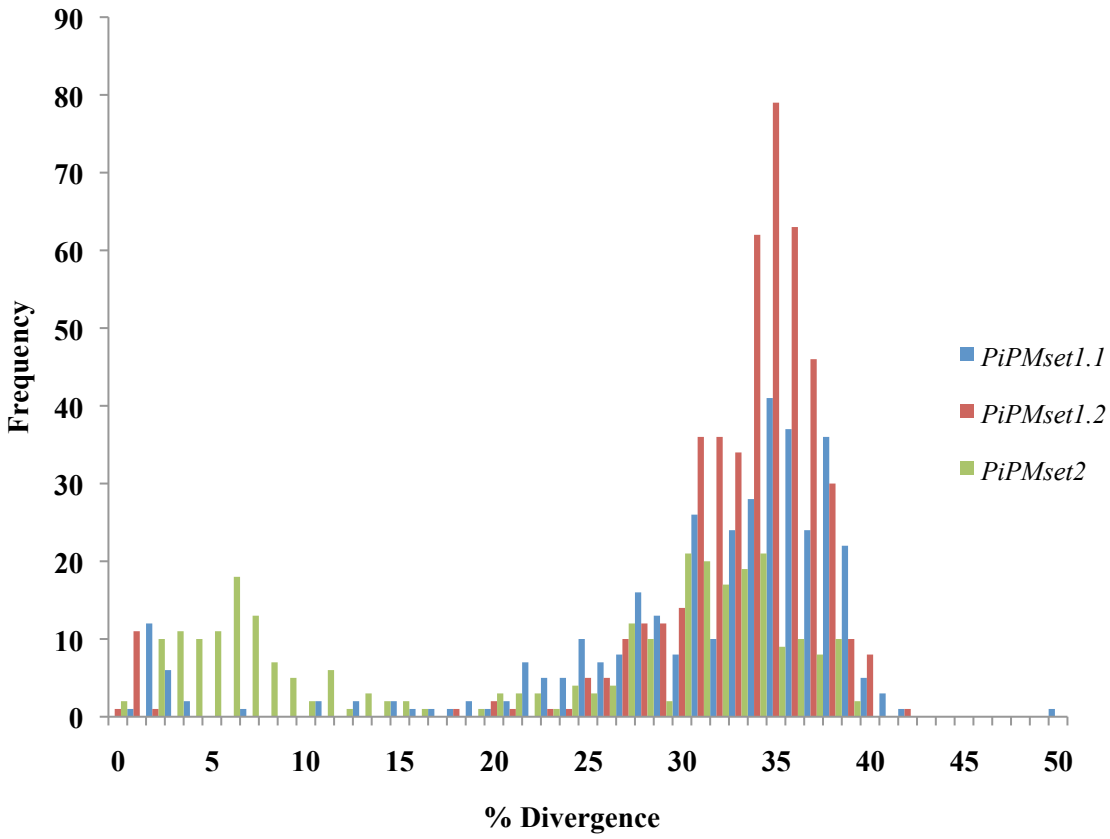


Figure A5: Sequence divergence of *PiPMset1.1, PiPMset1.2* and *PiPMset2.*

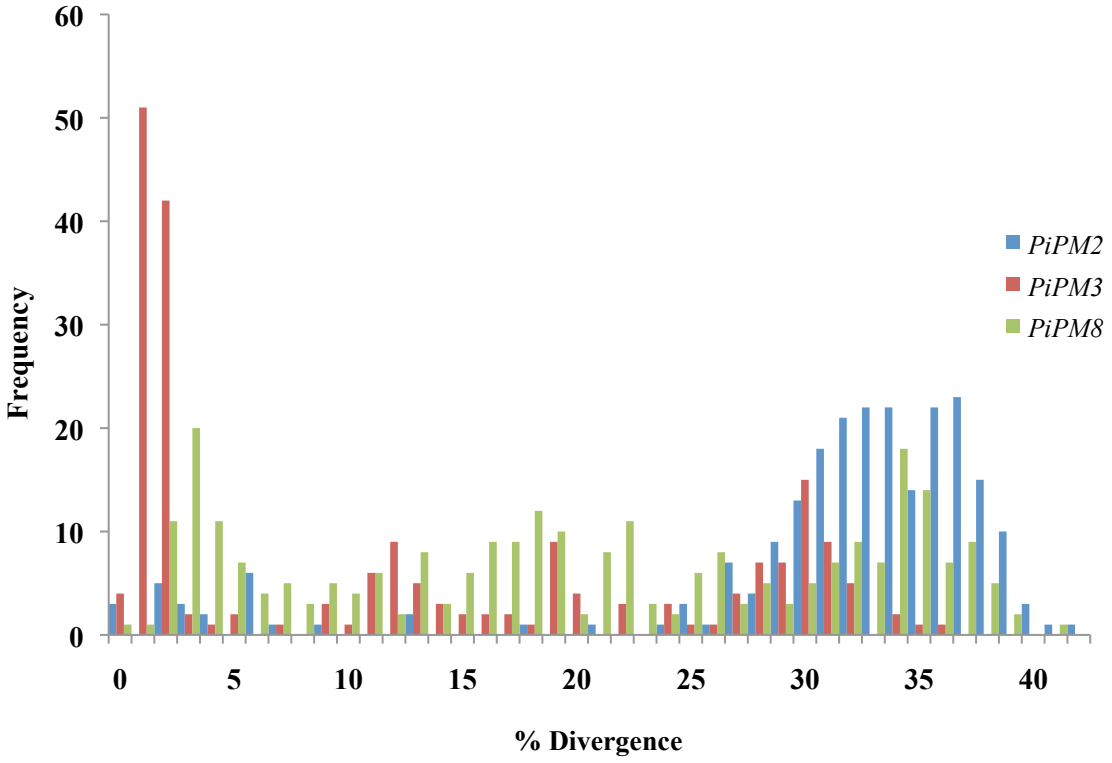


Figure A6: Sequence divergence of *PiPM2*, *PiPM3* and *PiPM8* family of *PackMULEs* with Ulp1 protease.

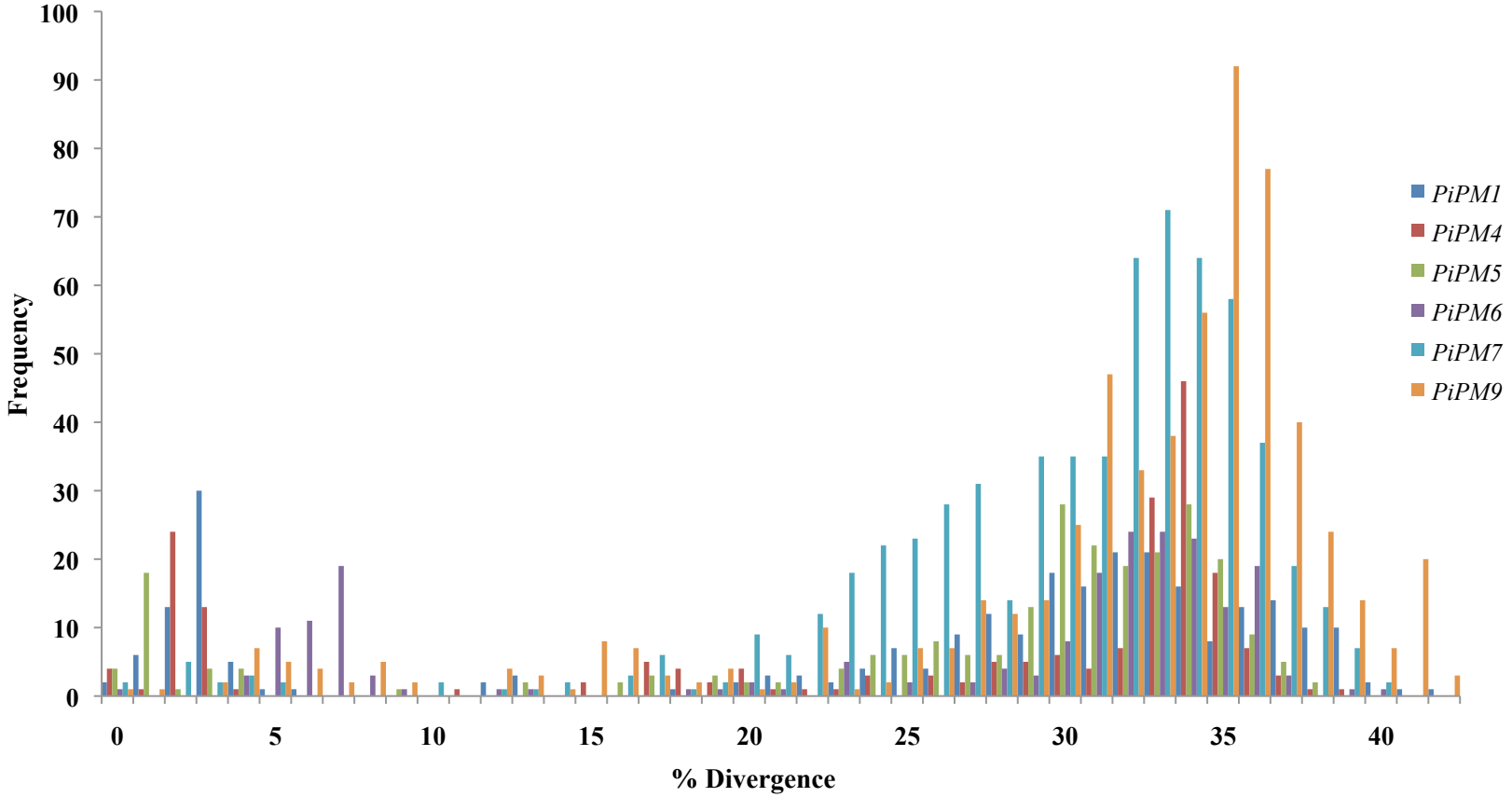


Figure A7: Sequence divergence of *PackMULEs* with FAR1 domain and Ulp1 protease.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Avrova AO, Stewart HE, De Jong WD, Heilbronn J, Lyon GD, et al. 1999. A cysteine protease gene is expressed early in resistant potato interactions with *Phytophthora infestans*. Mol Plant Microbe Interact. 12: 1114-1119.

Bao W, Jurka MG, Kapitonov VV, Jurka J. 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol Biol Evol. 26: 983-993.

Blair JE, Coffey MD, Park SY, Geiser DM, Kang S. 2008. A multi-locus phylogeny for *Phytophthora* utilizing markers derived from complete genome sequences. Fungal Genet Biol. 45: 266-277.

Brasier CM. 1992. Evolutionary Biology of *Phytophthora*: Genetic System, Sexuality and the Generation of Variation. Annual Review Phytopathology. 30: 153-171.

Brunner F, Rosahl S, Lee J, Rudd JJ, Geiler C, et al. 2002. Pep-13, a plant defense-inducing pathogen-associated pattern from *Phytophthora* transglutaminases. EMBO J. 21: 6681-6688.

Erwin, D.C., S. Bartnicki-Garcia, and P. H. Tsao. 1983. *Phytophthora*: its biology, taxonomy, ecology, and pathology. APS Press, St. Paul, Minn.

Essers L, Adolphs RH, Kunze R. 2000. A highly conserved domain of the maize

activator transposase is involved in dimerization. Plant Cell. 12: 211-224.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 9: 397-405.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41: 331-368.

Feschotte C, Pritham EJ. 2009. A cornucopia of *Helitrons* shapes the maize genome. Proc Natl Acad Sci U S A. 106: 19747-19748.

Feschotte C, Ranganathan N, Guibotsy ML, Levine D. 2009. Exploring Repetitive DNA Landscapes Using REPCLASS, a Tool That Automates the Classification of Transposable Elements in Eukaryotic Genomes. Genome Biology and Evolution. 21: 205-220.

Feschotte C, Wessler SR. 2001. Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci U S A. 98: 8923-8924.

Gijzen M. 2009. Runaway repeats force expansion of the *Phytophthora infestans* genome. Genome Biol. 10: 241.

Goodwin TJ, Butler MI, Poulter RT. 2003. Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiology. 149: 3099-3109.

Govers F, Gijzen M. 2006. *Phytophthora* genomics: the plant destroyers' genome decoded. Mol Plant Microbe Interact. 19: 1295-1301.

Govers F. 2001. Misclassification of pest as 'fungus' puts vital research on wrong track. Nature. 411: 633.

Grunwald NJ, Goss EM, Press CM. 2008. *Phytophthora ramorum*: a pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals. Mol Plant Pathol. 9: 729-740.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33: W557-559.

Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK. 2005. A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol. 57: 115-127.

Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature. 461: 393-398.

Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, et al. 2009. The functional role of *Pack-MULEs* in rice inferred from purifying selection and expression profile. Plant Cell. 21: 25-38.

Haverkort AJ, Boonekamp PM, Hutten R, Jacobsen E, Lotz LAP, et al. 2008. Societal Costs of Late Blight in Potato and Prospects of Durable Resistance Through Cisgenic Modification. Potato Research. 51: 47-57.

Hildebrand, A. A. 1959. A root and stalk rot of soybeans caused by *Phytophthora megasperma* Drechsler var. *sojae* var. nov. Can. J. Bot. 37:927-957.

Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, et al. 2006. Transposon-mediated expansion and diversification of a family of ULP-like genes. Mol Biol Evol. 23:

1254-1268.

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR. 2006. The transposable element landscape of the model legume *Lotus japonicus*. Genetics. 174: 2215-2228.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17: 754-755.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. *Pack-MULE* transposable elements mediate gene evolution in plants. Nature. 431: 569-573.

Jiang RH, Tyler BM, Whisson SC, Hardham AR, Govers F. 2006. Ancient origin of elicitin gene clusters in *Phytophthora* genomes. Mol Biol Evol. 23: 338-351.

Judelson HS. 1997. The genetics and biology of *Phytophthora infestans*: modern approaches to a historical challenge. Fungal Genet Biol. 22: 65-76.

Judelson HS. 2007. Genomics of the plant pathogenic oomycete *Phytophthora*: insights into biology and evolution. Adv Genet. 57: 97-141.

Kamoun S. 2003. Molecular genetics of pathogenic oomycetes. Eukaryot Cell. 2: 191-199.

Kawasaki S, Nitasaka E. 2004. Characterization of Tpn1 family in the Japanese morning glory: En/Spm-related transposable elements capturing host genes. Plant Cell Physiol. 45: 933-944.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, et al. 2005. The tree of eukaryotes. Trends Ecol Evol. 20: 670-676.

Lal SK, Hannah LC. 2005. *Helitrons* contribute to the lack of gene colinearity observed

in modern maize inbreds. Proc Natl Acad Sci U S A. 102: 9993-9994.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409: 860-921.

Large, E. C. 1940. The advance of the fungi. Jonathan Cape Ltd., London.

Li SJ, Hochstrasser M. 1999. A new protease required for cell-cycle progression in yeast. Nature. 398: 246-251.

Makalowski W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene. 259: 61-67.

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37: D205-210.

Marmorstein R. 2003. Structure of SET domain proteins: a new twist on histone methylation. Trends Biochem Sci. 28: 59-62.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 36: 344-355.

Moran JV, DeBerardinis RJ, Kazazian HH, Jr. 1999. Exon shuffling by L1 retrotransposition. Science. 283: 1530-1534.

Nicholas, K.B., Nicholas H.B. Jr., and Deerfield, D.W. II. GeneDoc: Analysis and Visualization of Genetic Variation, *EMBNEW.NEWS* **4**,14 1997.

Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci. 12: 357-358.

Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D, et al. 2006. Evaluating

support for the current classification of eukaryotic diversity. PLoS Genet. 2: e220.

Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. Bioinformatics. 21 Suppl. 1: i351-358.

Pritham EJ. 2009. Transposable elements and factors influencing their success in eukaryotes. J Hered. 100: 648-655.

Pritham EJ, Putliwala T, Feschotte C. 2007. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 390: 3-17.

Reader, J. Potato: A history of the propitious esculent Yale Univ. Press, 2009.

Sawada K, Yang Z, Horton JR, Collins RE, Zhang X, et al. 2004. Structure of the conserved core of the yeast Dot1p, a nucleosomal histone H3 lysine 79 methyltransferase. J Biol Chem. 279: 43296-43306.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326: 1112-1115.

Schubert HL, Blumenthal RM, Cheng X. 2003. Many paths to methyltransfer: a chronicle of convergence. Trends Biochem Sci. 28: 329-335.

Sogin ML, Silberman JD. 1998. Evolution of the protists and protistan parasites from the perspective of molecular systematics. Int J Parasitol. 28: 11-20.

Tyler BM. 2007. *Phytophthora sojae*: root rot pathogen of soybean and model oomycete. Molecular Plant Pathology. 8: 1-8.

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, et al. 2006. *Phytophthora* genome

sequences uncover evolutionary origins and mechanisms of pathogenesis. Science. 313: 1261-1266.

van Leeuwen H, Monfort A, Puigdomenech P. 2007. *Mutator*-like elements identified in melon, Arabidopsis and rice contain ULP1 protease domains. Mol Genet Genomics. 277: 357-364.

Whisson SC, Avrova AO, Lavrova O, Pritchard L. 2005. Families of short interspersed elements in the genome of the oomycete plant pathogen, *Phytophthora infestans*. Fungal Genet Biol. 42: 351-365.

Xia Y. 2004. Proteases in pathogenesis and plant defense. Cellular Microbiology. 6:906-913.

Zabala G, Vodkin L. 2007. Novel exon combinations generated by alternative splicing of gene fragments mobilized by a *CACTA* transposon in *Glycine max*. BMC Plant Biol. 7: 38.

Zabala G, Vodkin LO. 2005. The wp mutation of *Glycine max* carries a gene-fragment-rich transposon of the *CACTA* superfamily. Plant Cell. 17: 2619-2632.

BIOGRAPHICAL INFORMATION

Komal Vadnagara was born in the city of Ahmedabad, Gujarat, India. She immigrated to the US with her family and grew up in Texas. She received her high school diploma in 2004 from Aledo High School, Aledo, TX. She received her Bachelor's degree in Biology with minor in Chemistry in May 2008, and Master's in Biology in May 2010 from The University of Texas, Arlington. During her undergrad years, she worked on an independent research project studying Mobile DNA in Pritham Lab and decided to pursue graduate school. She received Phi Sigma travel grant in 2009 and William L. & Martha Hughes study in Biology award in 2010. Her research is focused on elucidating the impact of transposable elements in phytopathogens' genome evolution. Her future plans include pursuing a career in medical sciences involving patient care, conducting clinical research, teaching and travelling that will allow her to contribute to the society.