UNIVERSITY OF CALIFORNIA
RIVERSIDE

A Multidisciplinary Approach for Identifying Stage-specific Transcription Factor Binding Sites in the Irish Potato Famine Pathogen, *Phytophthora Infestans*

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

Graduate Program in

Genetics, Genomics and Bioinformatics

by

Sourav Roy

December 2011

Dissertation Committee:
Howard S. Judelson, Chairperson
Katherine A. Borkovich
Jason E. Stajich

The Dissertation of Sourav Roy is approved:

_____

_____

_____
(Chairperson)

University of California, Riverside

Acknowledgements

During the course of my graduate studies a number of people have inspired, helped and supported me. I must acknowledge that I would not have reached this far had they not been there, by my side. The list is long but I want to acknowledge their contributions towards my development as a researcher.

My greatest appreciation is reserved for my major advisor, Dr. Howard Judelson, for being such a strong and supportive mentor. He taught me how to think and write scientifically, and gave me great freedom to express my thoughts. Without his guidance and persistent support, this dissertation would not have been possible.

I would also like to thank my dissertation committee member and qualifying committee chair, Dr. Katherine Borkovich, for providing valuable scientific advice, help and encouragement over the years. My other dissertation committee member Dr. Jason Stajich, deserves special thanks for agreeing to be on my committee at a short notice, and for helping me with the questions related to bioinformatics.

I would also like to acknowledge the help that I have received from my guidance and former dissertation committee member Dr. Tao Jiang, over the entire course of this study. I am grateful for his support . I should also thank Dr. Patricia Springer, who was my guidance committee member, during the first two years, for her valuable inputs.

The Broad Institute deserves a special mention for providing me with the access to their *Phytophthora infestans* database.

I appreciate the help and support that I received from all the past and current members of the Judelson lab, during the last five years. It was a joy sharing a lab with such wonderful, supportive and friendly people. I am grateful to all the faculty members, staffs and fellow students in the Genetics, Genomics and Bioinformatics Graduate Program for their help, encouragement and friendship.

I would like to convey my deepest feelings of gratitude towards my parents for everything they have given me during my life, without the happy upbringing that they have provided, I would not have reached this far.

Finally, I would like to thank my wife, for her love, help and unwavering support, and my three and a half years old daughter, who is a bundle of joy, and has been a major stress reliever during those difficult times.

ABSTRACT OF THE DISSERTATION

A Multidisciplinary Approach for Identifying Stage-specific Transcription Factor Binding Sites in the Irish Potato Famine Pathogen, *Phytophthora Infestans*

by

Sourav Roy

Doctor of Philosophy, Genetics, Genomics and Bioinformatics Graduate Program
University of California, Riverside, December 2011
Dr. Howard Judelson, Chairperson

Phytophthora infestans, an oomycete within the phylum Heterokontophyta is one of the most devastating phytopathogens, causing late blight in potato and tomato. Its pathogenic success depends on the formation of different asexual spores such as sporangia and zoospores. My goal was to identify what regulates transition between each of the five different asexual stages viz. hyphae, spores, cleaving sporangia, swimming zoospores and germinating cysts, by understanding what determines stage-specific transcription. To help accomplish this, I have identified potential binding sites for regulatory elements within the promoters of stage-specific, co-expressed *Phytophthora infestans* genes, by integrating bioinformatics and traditional molecular biology techniques. Promoter sets, of co-expressed genes identified from expression data, were searched for over-represented motifs using motif discovery algorithms. Approximately 15 to 30 over-represented motifs were detected for each of the five stages. Phylogenetic footprinting and positional bias analyses increased the robustness of

Transcription Factor Binding Site (TFBS) predictions. Some of the over-represented motifs, which were evolutionarily conserved and showed bias for a certain position within the promoters, were tested for their functionality. Putative TFBSs related to each of the above-mentioned asexual stages, other than the short and transient swimming zoospore stage, were shown to be biologically active. Molecular biology techniques like promoter-reporter fusion assay, serial deletion, target-specific mutation, RNA-blotting and electromobility shift assay, were used for the functional validation of these potential TFBSs, all of which acted as proximal promoter elements. In addition to these elements, we also looked at the core promoter elements of *P. infestans* genes. A novel core promoter element, specific to the group Pythiales and named DPEpyth, was identified. We have also come up with better Phytophthora specific definitions (consensus sequence), for 'FPR' that has been previously detected within the oomycete core promoters and also for the very well known Initiator element ('Inr'). I believe that the identification of these putative TFBSs should lead to a better understanding of signaling pathways regulating spore and infection structure development and provide insight into new disease control strategies in the future.

Table of Contents

CHAPTER III

List of Figures

CHAPTER III

Core promoter elements in Irish potato famine pathogen *Phytophthora infestans:* their consensus, effect in gene expression and distribution within the Heterokontophyta.

List of Tables

CHAPTER I

CHAPTER II

Appendices:

CHAPTER III

Core promoter elements in Irish potato famine pathogen *Phytophthora infestans:* their consensus, effect in gene expression and distribution within the Heterokontophyta

Introduction

Developmental regulation has always been a key area of research and the involvement of extracellular and intracellular signals has intrigued researchers for decades. Much of the pathogenic success of eukaryotic microbes like *Phytophthora infestans* depends on events such as sporulation and infection structure formation. These stages are induced by a variety of environmental and physiological factors. However, very little is known at the molecular level of how such signals are translated into the cellular responses, but the end result is usually a changed transcriptome due to direct or indirect alteration of transcription factors (TFs). We are interested in knowing what causes transition from one developmental stage to another, but our knowledge is limited only to sets of genes that are up or down regulated during these transitions. Studies of genes induced during different stages of development, especially very early in these stages, including analyses of their promoter motifs that regulate transcription, will advance our understanding of what triggers development. Finding the molecular mechanism behind the activation of these transcription factors (TFs) and identification of their binding sites (TFBSs) will lead to new ways of disease control. Here in this study a multidisciplinary approach has been taken to learn what regulates life-stage differentiation in *P. infestans* with special emphasis on sporulation and zoosporogenesis-induced transcription. In the following sections the concepts important for this study are discussed.

Heterokonts:

Heterokonts make up a major eukaryotic phylum that contains more than 100,000 known species (van den et al., 1995), many of which are diatoms and are part of the Chromalveolates supergroup (Kemen et al., 2011). The name heterokonts is due to the presence of two differently-shaped flagella in a life cycle stage, in which the cells are motile. The flagella include a posteriorly directed whiplash and an anteriorly directed fibrous and ciliated one (Rossman and Palm, 2006). Heterokonts also include brown algae and many important pathogens and saprophytes in the oomycete ("water mold") class that is discussed in the next section.

Oomycetes:

Oomycetes are a highly diverse class of eukaryotic organisms and are found all over the world; from terrestrial mountains to open sea environments (Thines and Kamoun, 2010), from the hot deserts of Iran (Mirazee et al., 2009) to the freezing arctic regions including Antarctica (Bridge et al., 2008, Hughes et al., 2003). The large round oogonia, structures that contain the female gametes, are responsible for the name oomycete or “Oomycota” which means “egg fungi”. Most of these are filamentous microorganisms and are similar to fungus in morphology but have evolved independently (Gijzen, 2009). Unlike true fungi that are unikonts (have one flagellum) and are related to animals, oomycetes as mentioned before are heterokonts and belong to the chromalveolates (Gijzen,

2009). There are striking similarities between true fungi and oomycetes. Dissemination by spores, filamentous growth, pathogenic types and lifestyles are some of the basic similarities. Some of the major differences between oomycetes and true fungi include diploidy instead of haploidy, nonseptate hyphae, cellulose instead of chitin in the cell wall and different lysine synthesis pathways (Latijnhouwers et al., 2003). Even though this group is a collection of diverse species that include saprophytes and pathogens of plants, vertebrates, insects, fishes and microbes, more than 60% of the oomycete species are plant parasites (Thines and Kamoun, 2010) and *Phytophthora infestans* is one of the most devastating ones.

Phytophthora:

Within the class Oomycota, Phytophthora ("the plant-destroyer") is a genus that consists of species responsible for damaging plants. In 1875, Heinrich Anton de Bary first described this genus that we now know contains over 100 species of plant pathogens causing serious economic and environmental damage. Even though Phytophthora species in most cases are pathogenic to dicotyledons, some infect monocots too. These are relatively host-specific parasites, with the exception of a few species like *P. cinnamomi* and *P. palmivora* that can infect more than 900 (Zentmyer, 1980), and more than 130 (Chee, 1969) different hosts respectively. Some of the major diseases that the Phytophthora species cause to the economically important plants are late blight in potato and tomato by *P. infestans*, root and stem rot in Soya bean by *P. sojae*, and sudden

oak death by *P. ramorum*. In general, the two main strategies for control of the plant diseases caused by Phytophthora species are growing resistant cultivars (Forbes and Jarvis, 1994) and chemical control (Fernández-Northcote et al., 2000). These management strategies are not optimal, due to insufficient resistance levels of available cultivars (Andrade-Piedra et al, 2005) and growing fungicide resistance (Hakiza, 1999). Phytophthora being diploid has a genetic system that is more similar to that of higher organisms than true fungi. This along with its economical importance and a range of reproductive mechanisms make Phytophthora an interesting genus for research (Braiser, 1992).

***Phytophthora infestans*:**

P. infestans is a heterothallic oomycete that causes late blight disease in potatoes, tomatoes and some other members of the Solanaceae family. *P. infestans* was the causal agent for the great Irish potato famine of the 1840s, which resulted in a loss of 1.5 million lives. Another 1.5 million people had to immigrate to other parts of the world (Bourke, 1964). The great Irish famine, continues to be one of the most destructive plant disease epidemic ever documented. Late blight was also one of the first documented plant diseases linked to a microbe (Berkeley, 1846; DeBary, 1876). But, this organism is not just historically important; it is one of the most economically important plant pathogens too. Current conservative estimates show an annual worldwide loss of ~$6.7 billion (Haverkort et al., 2008) in the yield of potato, the fourth largest food crop (Reader, 2009), due to late blight. This pathogen is known for its adaptive

capability to control strategies like genetically resistant cultivars (Fry, 2008), thereby making its management extremely difficult. Its rapid growth within the susceptible host tissue is due to the asexual cycle (Fry, 2008) and therefore, the asexual cycle is extremely important for successful pathogenicity.

Asexual cycle:

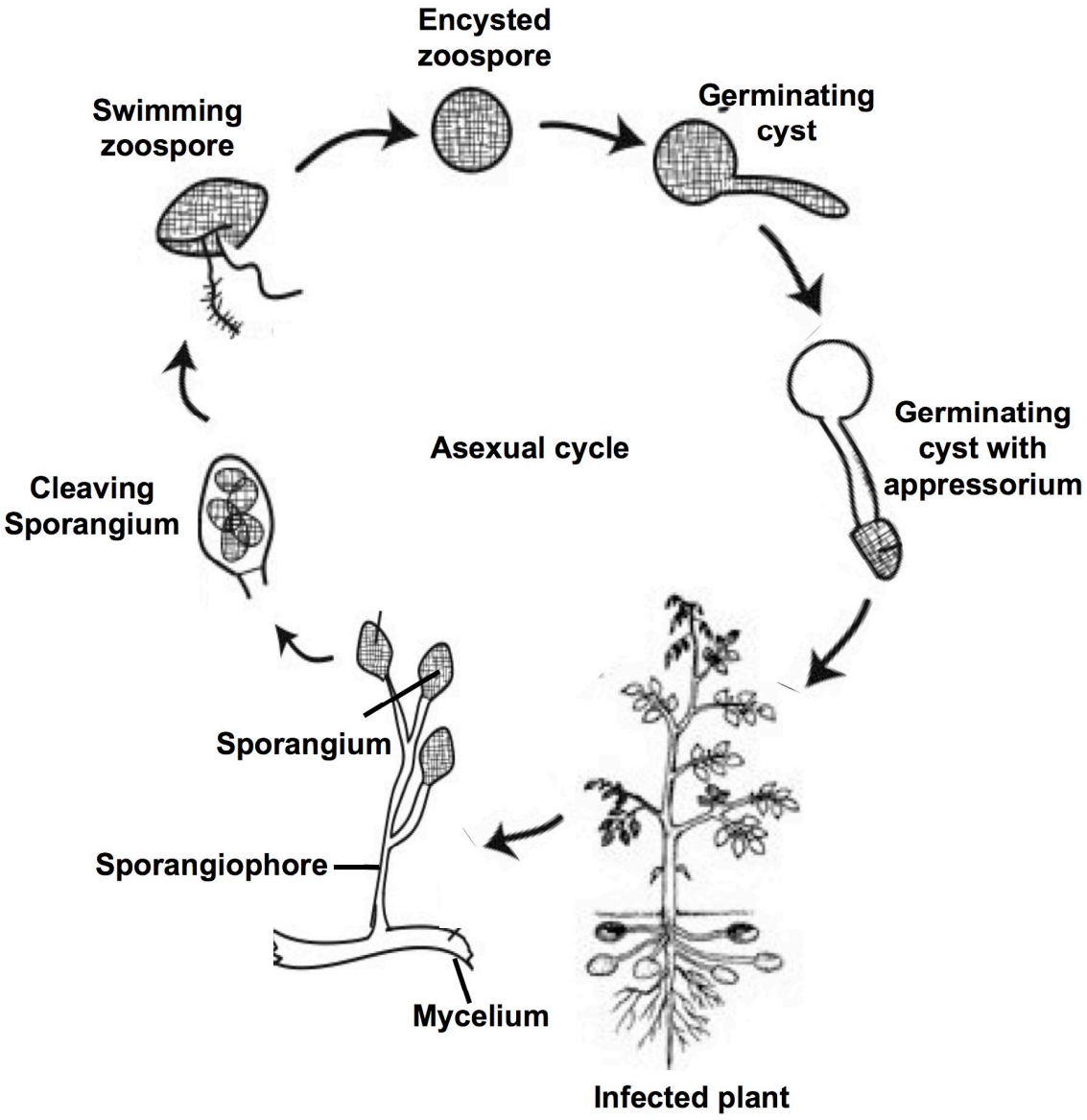
The asexual life cycle in *P. infestans* can broadly be divided into five different stages before infection in plants viz. hyphae, spores, cleaving sporangia, swimming zoospores and germinated cysts (Fig 1). Some hyphae develop into sporangiophores, which go on to bear sporangia. The sporangia can easily detach from the sporangiophores, and can be dispersed aerially to other plant tissues (Aylor et al., 2001). These sporangia under cold (below 15 °C) and moist conditions cleave and release wall-less, biflagellate and motile zoospores. The zoospores are motile for a very shot period; they often encyst within an hour from the time they are released. A germ tube comes out of the germinating cyst under favorable conditions to penetrate leaf or stem tissue (Fry, 2008). There is some knowledge about the conditions required for the transition between the different stages but not much is known about what triggers these transitions. Checking what causes the differential expression of genes during the various stages can provide some clues.

Fig 1 legend:

Asexual cycle in *P. infestans*:

Shown is the asexual cycle for development in *P. infestans*. Thread-like structures called mycelia (hyphae) gives rise to sporangia on the terminus of specialized structures called sporangiophores. Sporangia cleaves and releases biflagellate swimming zoospores, which encysts, forms germtube and infects the plant with the help of infection structure called appressoria.

Fig 1


Encysted zoospore
Germinating cyst
Swimming zoospore
Germinating cyst with appressorium
Asexual cycle
Cleaving Sporangium
Sporangium
Sporangiophore
Mycelium
Infected plant

Gene expression and stage-specificity:

A 2008 research paper by Judelson and his colleagues (Judelson et al. 2008) showed that approximately 60% of the *P. infestans* genes exhibit greater than two-fold changes in mRNA levels over the life cycle and ~15% are stage specific. This number is relatively high in comparison with other plant pathogens and is most likely due to huge structural and physiological differences between zoospores and other life stages (Judelson et al. 2008). What causes these genes to express differentially during a certain stage, most likely depends on the elements present in their non-coding regions.

Non-coding regions:

Non-coding regions are under less evolutionary pressure as compared to the coding regions. As a result, the degree of conservation in the non-coding regions is much more varied, with most of the sequences drifting randomly, and only a few showing conservation due to positive selection pressure (Jareborg et al., 1999). Conserved sequences within the non-coding regions might have important functions related to regulation of gene expression, maintenance of the structural organization of the genome, or other chromosomal functions that are not yet known (Koop and Hood 1994). These conserved sequences might regulate gene expression by binding proteins responsible for transcription, known as the transcription factors (TFs).

Transcription Factor Binding Sites (TFBS):

TFBSs are DNA sequences typically 5-15 base-pairs (bp) long (Bulyk, 2003), close to and upstream of the transcription start site, that control gene expression by activation or inhibition of the transcription machinery (Tompa et al. 2005). There are few TFs for which there are well-characterized TFBSs (Bulyk, 2003). TFBSs are often degenerate and the sequence degeneracy selected through evolution is beneficial (Bulyk, 2003). The degeneracy helps in determining different levels of activity in different gene promoters and thereby varied levels of expression, as per the requirement of the cell (Stromo 2000). Orientation, most of the times does not determine the function of a TFBS (Bulyk, 2003). Identifying TFBSs in higher eukaryotes is challenging as a TFBS can be close to or far from the genes that it regulates and can be found upstream, downstream, or even within the introns of these genes. TFBSs can broadly be divided in two categories based on their distance from the TSS viz. promoter elements (within 1 kb upstream of the TSS) and distal regulatory elements (>1 kb upstream of the TSS).

Core promoter elements:

Core promoter elements, unlike proximal promoter elements or distal regulatory elements, are present within ~50 bases on either side of the TSSs in most protein-coding genes transcribed by RNA polymerase II. The minimal stretch of contiguous DNA required by the RNA polymerase II machinery for accurate transcription initiation can be defined as the core promoter (Butler and

Kadonaga 2002, Struhl 1987, Weis and Reinberg 1992, Smale 1994, 1997, 2001 Smale et. al. 1998, Burke et al. 1998). The core promoter elements interact directly with the components that make up the basal transcription machinery (Hochheimer and Tjian 2003, Woychik and Hampsey 2002, Hampsey 1998) and direct the assembly of pre-initiation complex (McLeod et. al, 2004), general transcription factors and a mediator, thereby orchestrating the initiation of transcription (Burke and Kadonaga, 1997).

Upstream regulatory elements:

Upstream regulatory elements are those TFBSs that are not within the core promoter region. These can be further classified into proximal promoter elements and distal regulatory elements, which can be enhancers, silencers, insulators, or locus control regions (Manston et al., 2006). These are responsible for the determination of the strength of the promoter, by enhancing or repressing transcription (Manston et al., 2006) and in most cases do not have a role in initiation of transcription. Little is known about the binding sites of most transcription factors, and this is an area of great interest. Both molecular biology techniques and computational tools are used for the discovery of novel regulatory elements and there is great hope regarding the computational methods (Tompa et al, 2005). For the *de novo* TFBS identification, usually a collection of regions upstream of the start sites of coregulated genes are used as input, from which the computational tool identifies short, statistically overrepresented DNA sequence 'motifs' (Tompa et al. 2005).

Overrepresented motifs:

It is highly likely that most stage-specific genes with identical expression profile are controlled by TFs that are similar in nature and would in all likelihood bind to identical TFBSs. Therefore, in a set of promoters from co-expressed genes the TFBSs would appear as motifs that are overrepresented. Any motif present in a set of sequences more than the number of times it is supposed to be identified by random chance, based on the background sequences, is an overrepresented motif. There are several softwares that follow different algorithms for the discovery of overrepresented motifs. But, overrepresentation of a motif does not assure that it has relevance in transcription or any other biological process.

Positional Bias:

One of the most challenging aspects of bioinformatics promoter analysis is determining the biological relevance of the predicted TFBSs (overrepresented motifs; Waveren and Moraes, 2008). Many TFBSs identified to date show a bias for a certain position within the gene promoters (Waveren and Moraes, 2008). This is probably because similar TFs binding to the sequence would maintain the distance from the TSS in order to control proper expression. Previous studies have adopted an approach of checking the positional bias along with the conservation of predicted TFBSs, and found that there is high likelihood for TFBSs with a positional bias of being biologically significant (FitzGerald et al., 2004).

Evolutionary conservation and Phylogenetic Footprinting:

As discussed previously, conservation within non-coding regions across species suggest purifying selection pressure and therefore these regions are expected to carry elements that are functional. Islands of highly conserved functional regions surrounded by a background of sequences evolving without any selection pressure are known as Phylogenetic footprints and the process of identifying them has been termed as Phylogenetic footprinting (Tagle et al., 1988). With sequencing getting cheaper by the day and the number of alignable genome sequences increasing rapidly, phylogenetic footprinting is being used widely for the identification of binding sites (Levy and Hannenhalli, 2002; Wasserman and Fickett, 1998; Xie et al., 2005). Checking for evolutionary conservation can be an effective means for reducing the false-positive rate in binding site prediction, but one has to keep in mind that conservation is neither a sufficient (Nobrega et al., 2004) nor a necessary condition for biological functionality (Dermitzakis and Clark, 2002). Also, conserved regions might have a number of functional roles other than binding transcription factors (Hannenhalli, 2008).

Aims of this study:

As discussed earlier, *Phytophthora infestans* has five key stages in the asexual life cycle which are immensely important for this organism to succeed as a pathogen. Not much is known about what is responsible for the transition from one stage to another. But we know that certain genes show high or low levels of

expression during each of these stages (Judelson et al. 2008). Transcription factors recognize specific nucleotide sequences (Babu et al. 2007), called TFBS within the promoter region of a gene. We believe that knowledge about these binding sites would lead us to the transcription factors, which in turn would enhance our understanding of the signaling pathways involved in the transition between the stages. This should then help in the development of new and improved disease control strategies that are economically viable and environment friendly. To identify the TFBSs and to get an idea about the promoter structure in *P. infestans,* I have studied both proximal promoter (ChapterI and II) and core promoter (Chapter III) regions in detail.

Chapter I presents the proof of the concept that bioinformatics tools when combined with molecular biology techniques can increase the robustness of transcription factor binding site (TFBS) prediction, thereby decreasing the time required to prove their functionality. Stage-specific, co-expressed genes, upregulated in sporangia and cleavage were identified from microarray data. The promoter regions (1 kb) upstream of the translation start sites (ATG) were extracted from the *Phytophthora infestans* database, created and maintained by the Broad Institute (http://www.broadinstitute.org/annotation/genome), and were used as inputs for the motif finding programs. 'CGTCCTCG', one of the overrepresented motifs within the promoters of the genes up-regulated in cleaving sporangia, showed bias for a certain position. This motif was also found to be conserved in the promoters of the *P. sojae* and *P. ramorum* orthologs of a

P. infestans gene (PITG_16321). The bioinformatics data thus suggested that this motif might be functional. Therefore, the functionality of this motif within the promoter of the said gene was tested by promoter-reporter fusion assay. Different sized promoters created by serial deletion were tested for function. Position-specific mutations and oligo-chimera assays were also employed to access the function of these promoters. GUS (β-glucuronidase, reporter gene) staining and RNA-blots confirmed the functionality of this motif. Specific binding activity was detected by electromobility shift assays (EMSA). Once it was observed that the bioinformatics analyses were increasing the robustness of the predicted TFBS, another motif, 'CTTCAAC', was chosen for functional analyses. This motif is overrepresented in the gene promoters of hyphae, sporangia and cleaving sporangia genes and showed positional bias within the hyphae and sporangia gene promoters and was conserved in all three stages. The functionality was proven by GUS staining of the sporangia stage tissue. This resulted in considerable reduction of the time required to prove that the motif is functional, in comparison with the time one needs for conventional blind promoter bashing techniques. The aim of this chapter was to show that bioinformatics combined with molecular biology can be a very powerful method for TFBS prediction.

In Chapter II, the upstream regulatory elements that are specific to one or more of the five key asexual stages, viz. hyphe, sporangia, cleavage, swimming zoospore and germinating cysts, were studied. Promoters of genes specific to

each of these developmental stages were searched for overrepresented motifs. These motifs along with those from sporangia and cleavage, were then subjected to positional bias and evolutionary conservation analyses. The idea behind these bioinformatics analyses was to increase the level of confidence for these overrepresented motifs being biologically functional. Three motifs, 'TACATGTA' that is overrepresented in all developmental stages, 'GTCGTCG' that is overrepresented in hyphae, sporangia, and 'TATTAATA' that is overrepresented in hyphae and germinating cyst stages, were checked for their functionality. It was also checked if these motifs showed any binding activity with nuclear proteins using electrophoretic mobility shift assays. The aim of this chapter, was to not only detect overrepresented motifs within the promoters of genes specific to each of the five developmental stages, but also to check how many of these putative TFBSs are shared by the promoters of genes specific to other stages. I have also analyzed if the motifs showed positional bias and evolutionary conservation.

Chapter III describes the results of a computational study to identify core promoter elements in *Phytophthora infestans.* Putative transcription start sites (TSSs) for certain genes were identified based on the EST data available. DNA sequences, 50 bases on either side of these putative TSSs, were extracted and subjected to *in-silico* search for overrepresented motifs, and their bias for a certain position within this region. To include those genes that lacked EST data a genome-wide analysis was done by searching 200 bases upstream of the

predicted translation start sites of all *P. infestans* genes. The effects of the core promoter elements on gene expression were analyzed with the help of microarray data. The conservation of the *P. infestans* core promoter elements among eight other heterokont species was checked. The aim of this chapter was to understand the core promoter structure of *P. infestans* genes, and throw some light on the evolution of the core promoter elements within the phylum Heterokontophyta and gives some idea about the correlation between the elements and gene expression.

This research should lead to the identification of TFs for selected TFBSs with the help of biochemical approaches and the study of pathways that activate these TFs with genetic, biochemical and cell-biological methods. That eventually should lead to a detailed understanding of what triggers the formation of spores and will give some insight on the other developmental stages. This certainly would lead to the signaling pathways regulate development in oomycetes and how oomycetes communicate with their environment. In terms of broader impact this study can be a first step towards new and improved strategies for blocking the disease. Transgenic plants that degrade molecules found to trigger development or chemicals that block the receptors of those molecules can be used to arrest the disease cycle.

REFERENCES:

1. van den HC, Mann DG, Jahns HM (1995) Algae: An Introduction to Phycology. Cambridge: Cambridge University Press. pp. 104, 124, 134, 166. ISBN 0-521-31687-1
2. Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, et al. (2011) Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. PLoS Biol. 9: e1001094. doi:10.1371/journal.pbio.1001094
3. Rossman AY, Palm ME (2006) Why aren't Phytophthora and other Oomycota true Fungi? Outlooks in Pest Manag. 17: 217-219
4. Thines M, Kamoun S (2010) Oomycete–plant coevolution: recent advances and future prospects. Curr. Opin. in Plant Biol. 13: 427-433
5. Mirzaee MR, Abbasi M, Mohammadi M (2009) Albugo candida causing white rust on Erysimum crassicaule in Iran. Australas. Plant Dis. Notes 4: 124–125
6. Bridge PD, Newsham KK, Denton GJ (2008) Snow mould caused by a Pythium sp.: a potential vascular plant pathogen in the maritime Antarctic. Plant Pathol. 57: 1066–1072
7. Hughes KA, Lawley B, Newsham KK (2003) Solar UV-B radiation inhibits the growth of Antarctic terrestrial fungi. Appl. Environ. Microbiol. 69: 1488–1491
8. Gijzen M (2009) Runaway repeats force expansion of the *Phytophthora infestans* genome. Geno. Biol. 10: 241

9. Latijnhouwers M, de Wit PJGM, Govers F (2003) Oomycetes and fungi: similar weaponry to attack plants. Trends in Microbiol. 11: 462-469
10. Zentmyer GA (1980) *Phytophthora cinnamomi* and the diseases it causes. Amer. Phytopathol. Soc. Monogr. 10: 1-96
11. Chee KH (1969) Hosts of *Phytophthora palmivora*. Review of Applied Mycol. 48: 337-344.
12. Forbes GA, Jarvis MC (1994) Host resistance for management of potato late blight. Adv in Potato Pests Biol and Mgmt The A. P. S. (St. Paul, MN. Zehnder GW, Powelson ML, Jansson RK, Raman KV eds) pp. 439-457
13. Fernández-Northcote EN, Navia O, Gandarillas A (2000) Basis of strategies for chemical control of potato late blight developed by PROINPA in Bolivia. Fitopatología 35: 137-149
14. Andrade-Piedra JL, Hijmans RJ, Forbes GA, Fry WE, Nelson RJ (2005) Simulation of potato late blight in the Andes. I: modification and parameterization of the Late blight model. Phytopathology 95: 1191-1199
15. Hakiza JJ (1999) The importance of resistance to late blight in potato breeding in Africa (Abstract). In: Proceedings of the Global Initiative on Late Blight Conference, Quito, Ecuador, March 16–19: 4
16. Braiser CM (1992) Evolutionary biology of Phytophthora Part I. Genetic system sexuality and the generation of variation. Annu. Rev. Phytopathol. 30: 153-171
17. Bourke PM (1964) Emergence of potato blight. Nature 203: 805–808

18. Berkeley MJ (1846) Observations, botanical and physiological on the potato murain, J Hort Soc London 1: 9–34
19. DeBary A (1876) Researches into the nature of the potato-fungus - *Phytophthora infestans*. J. Roy Agr. Soc.12: 239–268
20. 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 Res. 51: 47–57
21. Reader J (2009) Potato: A History of the Propitious Esculent. Yale Univ. Press pp. 336
22. Fry W (2008) *Phytophthora infestans*: the plant (and R gene) destroyer. Mol. Plant Pathol. 9: 385–402
23. Aylor DE, Fry WE, Mayton H, Andrade-Piedra J (2001) Quantifying the rate of release and escape of *Phytophthora infestans* sporangia from a potato canopy. Phytopathol. 91: 1189–1196
24. Judelson HS, Ah-Fong AMV, Aux G, Avrova AO, Bruce C et al. (2008) Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. Mol. Plant-Micro. Inter. 21: 433–447
25. Jareborg N, Birney E, Durbin R (1999) Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs. Genome Res 9: 815–824

26. Koop BF, Hood L (1994) Striking sequence similarity over almost 100 kilobase of human and mouse T-cell receptor DNA. Nat. Gen. 7: 48–53
27. Bulyk M (2003) Computational prediction of transcription-factor binding site locations. Geno. Biol. 5: 201
28. Tompa M, Li N, Bailey TL, Church GM, De Moor B et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23: 137
29. Stormo G (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16–23
30. Butler JEF, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. 16: 2583-2592
31. Struhl K (1987) Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast. Cell 49: 295–297
32. Weis L, Reinberg D (1992) Transcription by RNA polymerase II: Initiator-directed formation of transcription-competent complexes. FASEB J 6: 3300–3309
33. Smale ST (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim. Biophys. Acta. 1351: 73–88
34. Smale ST (2001) Core promoters: Active contributors to combinatorial gene regulation. Genes Dev. 15: 2503–2508

35. Smale ST, Jain A, Kaufmann J, Emami KH, Lo K et al. (1998) The initiator element: A paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harb. Symp. Quant. Biol. 58: 21–31
36. Burke TW, Willy PJ, Kutach AK, Butler JEF, Kadonaga JT (1998) The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. Cold Spring Harb. Symp. Quant. Biol. 63: 75–82
37. Hochheimer A, Tjian R (2003) Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev. 17: 1309–1320
38. Woychik NA, Hampsey M (2002) The RNA polymerase II machinery: structure illuminates function. Cell 108: 453–463
39. Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol. Mol. Biol. Rev. 62: 465–503
40. McLeod A, Smart CD, Fry WE (2004) Core Promoter Structure in the Oomycete *Phytophthora infestans*. Euk. Cell 3: 91-99
41. Burke TW, Kadonaga JT (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by $TAF_{II}60$ of *Drosophila.* Genes Dev. 11: 3020–3031
42. Manston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu. Rev. Gen. Hum. Gen. 7: 29-59

43. van Waveren C, Moraes CT (2007) Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system. BMC Genomics 9: 18
44. Fitz Gerald PC, Shlyakhtenko A, Mir AA, Vinson C (2004) Clustering of DNA sequences in human promoters. Genome Res. 14:1562–1574
45. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. 203: 439–455
46. Levy S, Hannenhalli S (2002) Identification of transcription factor binding sites in the human genome sequence. Mamm. Genome 13: 510–514
47. Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol. 278: 167–181
48. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. Nature 434: 338–345
49. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. Nature 431: 988–993
50. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. 19: 1114–1121

51. Hannenhalli S (2008) Eukaryotic transcription factor binding sites-modeling and integrative search methods. Bioinformatics 24: 1325–1331
52. Babu MM, Balaji S, Aravind L (2007) General trends in the evolution of prokaryotic transcriptional regulatory networks. Genome Dyn. 3: 66–80

Chapter I

Integrating bioinformatics with molecular biology increases robustness and decreases the time required for identification of transcription factor binding sites

ABSTRACT:

In this chapter it is shown how bioinformatics tools and techniques increase the robustness of transcription factor binding site (TFBS) prediction and thus can reduce the time required to prove that a predicted TFBS is functional, when compared to molecular biology techniques. *Phytophthora infestans,* a microbial eukaryote and an oomycete, is one of the most devastating plant pathogens. It causes late blight in potato and tomato, resulting in a loss of around 20% of the annual global yield. The principal inoculum for the disease is the zoospore, which develops from sporangia upon chilling. Therefore, understanding the mechanisms that activate transcription during the formation of spores is important. I believe that better management of the disease will result from improved understanding of what causes spores to cleave and release the motile zoospores. A big step towards achieving that goal is to identify regulatory motifs, such as Transcription Factor Binding Sites (TFBSs), responsible for controlling gene expression in sporangia and cleaving sporangia. Promoters of genes specific to these two stages were searched for over-represented motifs with the help of different algorithms. Five overrepresented motifs for each of

these stages were checked for any positional bias within the promoters. Phylogenetic footprinting involving three sequenced *Phytophthora* genomes, was employed to check for evolutionary conservation and thereby increase the robustness of the putative TFBS prediction. One putative TFBS, specific to the cleaving sporangia stage, was selected for functional analyses. A promoter with this motif was subjected to serial deletion to show that the region carrying this motif was important for the promoter to be functional. Target-specific mutations and an oligo-chimera assay were done to prove that the said motif with the help of core promoter elements was able to drive the expression of the reporter gene. Nuclear extracts from cleaving sporangia tissue were used to prove the binding affinity of nuclear proteins for this motif by electrophoretic mobility shift assay. Once it was proved that this overrepresented, positionally biased and evolutionarily conserved motif was functional, the functionality of a similarly high confidence, sporangia-specific motif, was confirmed in much less time. The results led us to believe that, the approach of integrating multiple bioinformatics techniques for TFBS prediction can reduce the time required for functional analyses considerably, by increasing the robustness of the predictions.

INTRODUCTION:

Many critical biological processes are dependent upon regulation of gene expression, and promoters have an essential role to play in controlling these processes. Therefore, to have a clear understanding of gene expression,

knowing the promoter strength and regulation is absolutely necessary (Bajic et al., 2004). Characterization of eukaryotic promoters is very difficult due to their extreme diversity (Smale and Kadonaga, 2003). In eukaryotes, promoters typically lie upstream of the genes (Tompa et al., 2005). The promoter elements can broadly be divided into core promoter and proximal promoter elements (Manston et al., 2006) based on their distance from the TSS. Core promoter elements in most cases interact with RNA polymerase II and the components that make up the basal transcription machinery (Hochheimer and Tjian, 2003; Woychik and Hampsey, 2002; Hampsey, 1998), and are primarily responsible for initiation of transcription (Butler and Kadonaga, 2002; Struhl, 1987; Weis and Reinberg, 1992, Smale, 1994, 1997, 2001; Smale et. al., 1998, Burke et al., 1998). Proximal promoter elements, unlike the core elements, are mainly responsible for the regulation of transcription and usually do a have stronger influence than the distal regulatory elements that are further upstream (Manston et al., 2006; Weis and Reinberg, 1997; Emami et al., 1995, Martinez et al. 1994), even though exceptions have been observed (Crawford et al., 1999; Yean and Gralla, 1997). Here in this study, a few proximal promoter elements in the sporulation and cleavage-induced genes were analyzed for one of the most destructive phytopathogens, *Phytophthora infestans*.

P. infestans spores play an essential role in plant-to-plant dissemination and in infection structure generation, and only certain genes are activated during the spore cycle. Understanding the regulatory elements like TFs and DNA

binding sites in the promoters of the genes, is therefore a promising approach to identify new control strategies for crop protection. Once these are identified, chemical libraries can be screened for compounds inhibiting TFs or proteins regulating such factors. The oomycetes are a group that has not been studied extensively, as a result, data on oomycete promoters are limited and no systematic genome-wide survey of promoter structure has been reported. Non-oomycete promoters do not work in *Phytophthora*, indicating the presence of unique transcription machinery (Judelson et al., 1991, 1992). Checking *P. infestans* promoters against motif databases like TRANSFAC is not productive due to the taxonomic distance of oomycetes from well-studied organisms. Most *P. infestans* promoters appear to be small as the intergenic regions average only 603 nt (Hass et. al., 2009) with some of the shortest untranslated regions noted for eukaryotes (Pesole et al., 1994). Also, oomycete promoters lack CpG islands, as there is no cytosine methylation (Judelson and Tani, 2007). Motifs for only a few TFs have been identified experimentally by approaches like promoter bashing, such as motifs that induce genes during spore formation (6 nucleotide long spore box; Ah Fong et. al, 2007) and zoospore release (7 nucleotide long cold box; Tani and Judelson, 2006). Identification of TFBSs by molecular biology techniques, even though reliable, is highly laborious and time consuming.

A number of computational approaches have been developed in the post-genomic era to counter the challenge of identifying the short and often degenerate binding sites in DNA for TFs (Bulyk, 2003). This has been a

frustrating problem for standard methods in computational sequence analysis. Simple *cis-regulatory* TFBSs, which usually are short and often degenerate DNA segments known as motifs, do not have enough sequence information on their own for dependable predictions. Thus, *de novo* motif identification has proved to be extremely difficult (Tharakaraman et al., 2008). Recent advances in genome sequence availability and high-throughput gene expression analysis technologies have facilitated the development of computational methods for motif discovery. This has lead to the implementation of a large number of motif discovery algorithms, which have been applied to various motif models over the past decade (Das and Dai, 2007). Methods to find new regulatory motifs such as TFBSs usually start from genes having similar expression patterns or sequences from chromatin immunoprecipitation (Tyler et al., 2006; Wakefield et al., 2005). The idea is to find overrepresented sequences in a dataset, versus control DNA, that in all likelihood binds TFs. To discover putative regulatory motifs in sets of co-regulated genes from the same genome, several pattern discovery algorithms have been developed (Hertz et al., 1990; Lawrence et al., 1993; Neuwald et. al, 1995; Bailey and Elkan, 1995; van Helden et al., 1998; Brazma et al. 1998; Hertz et al. 1999; van Helden et al., 2000; Thijs et al., 2001; Liu et. al. 2001). The detection usually starts from a random motif model, which is represented as a probabilistic weight matrix and is iteratively refined by different algorithms. The main algorithmic strategies used are expectation-maximization, Gibbs sampling and statistical overrepresentation. Each of these approaches has pros and cons.

Gibbs sampling and statistical overrepresentation methods are faster but tend to produce many spurious hits and may miss motifs due to the flexibility of the bases within the TFBSs, whereas the expectation-maximization approach is time consuming and the results may vary with the number of iterations used.

There is a particular position for a TF within the transcriptional complex that is anchored by the TSS. Therefore, it is highly likely that a TFBS, with which the TF interacts, is constrained positionally with respect to the TSS (Tharakaraman et al., 2008). The concept of using positional information with respect to TSS, for the prediction of TFBSs, has been used by multiple studies (Ptashne at al., 1982; Kielbasa et al., 2001; FitzGerald et al., 2004; Xie et al., 2005; Tharakaraman et al. 2005; Zhang et al., 2006; Marino-Ramirez et al. 2006). A major drawback for this method is that one needs proper annotations for robust predictions, and therefore, merely the sequence is not enough (Tharakaraman et al., 2008).

Another common strategy to predict cis-acting regulatory elements is the detection of conserved motifs in promoters of orthologous genes (phylogenetic footprints). Several software tools are routinely used to test hypotheses about regulation (Janky and van Helden, 2008). The premise behind this method is that selective pressure causes functional elements to evolve more slowly than non-functional sequences. Thus, conserved regions within orthologous promoters are candidate TFBSs. Phylogenetic footprinting has been applied with success in bacteria, fungi, plants and animals to identify TFBSs (Dermitzakis et al., 2002,

McCue et al., 2002, Cliften et al., 2003, Guo et al. 2003, Hong et al., 2003, Bowser and Tobe, 2007). It is likely that footprinting of distantly related species would only identify ancient regulatory elements. Utilizing the conservation patterns in multiple closely related species can identify more recently evolved regulatory elements. This technique is known as a phylogenetic shadowing and has been proposed by Boffelli et al. (2003). Like motif discovery methods the phylogenetic footprinting approaches too have pluses and minuses. These approaches rely on having suitably evolved sequences; if the species being used are too closely related, TFBSs may not be more conserved than the bases lacking function, due to the lack of sequence divergence. On the other hand, it is unlikely that a good alignment can be obtained if the species are too distant. Another limitation for this method is the assumption that the orthologs have the same expression pattern and hence the same TFBSs.

Therefore, based on the above discussion it can be concluded that both molecular biology techniques and computational methods for TFBS predictions have their own share of pros and cons. Computational algorithms are fast but not very reliable whereas molecular biology techniques even though reliable are laborious and highly time consuming. This is the reason that I decided to integrate bioinformatics approaches with molecular methods. I decided to use different computational algorithms to predict overrepresented motifs within the promoters of co-regulated genes, increase the robustness of these predictions by computational methods like positional bias analysis and phylogenetic footprinting

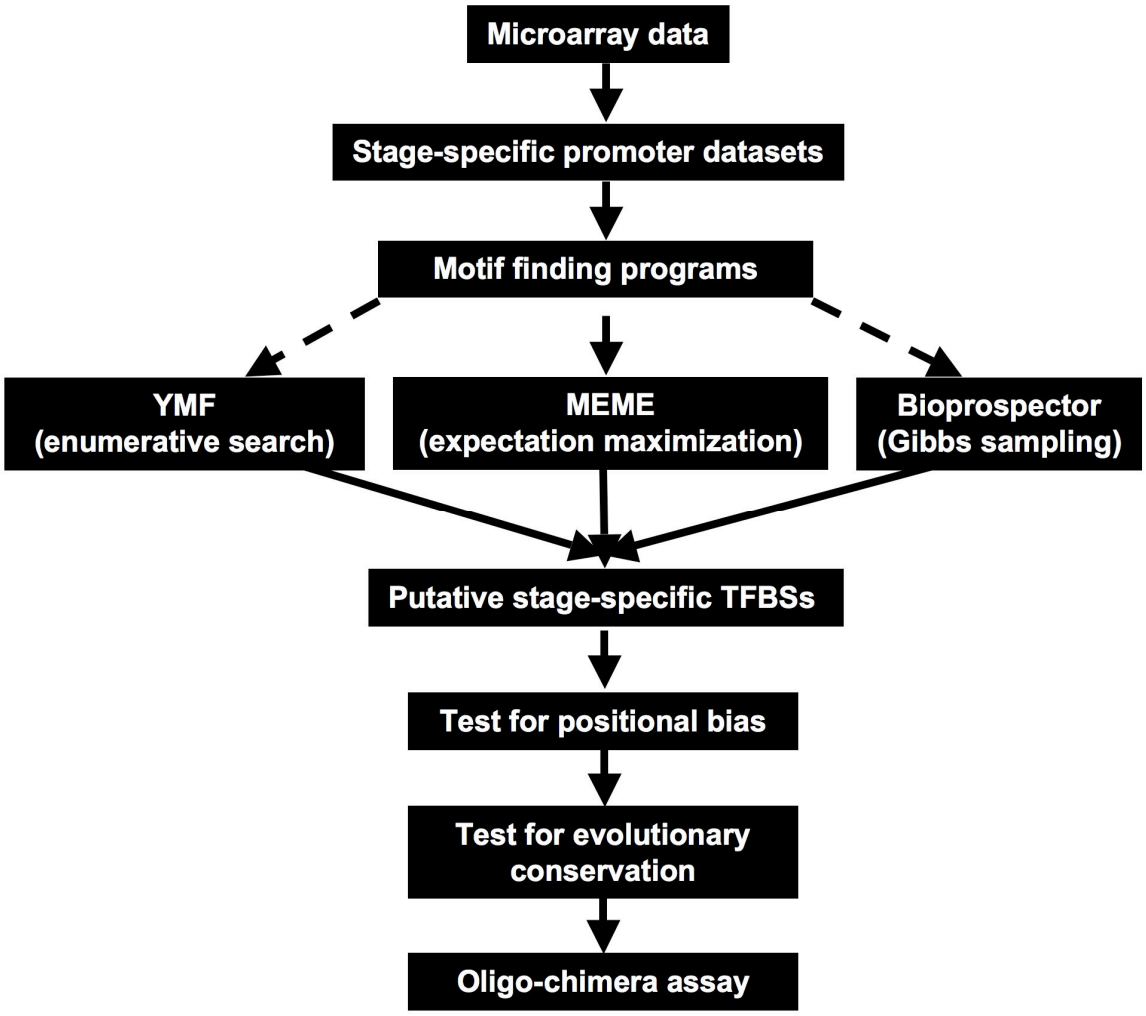
and finally test the functionality of some of these motifs by molecular biology techniques (Fig 1). I wanted to test this approach on a small scale before going ahead and predicting high confidence putative TFBSs for all the five asexual stages. Five overrepresented motifs from the cleaving sporangia stage were selected randomly and analyzed for positional bias. All of these five motifs showed a bias for a certain position within the promoters and were checked for evolutionary conservation by phylogenetic footprinting with *P. sojae* and *P. ramorum* orthologus promoters. Only one out of the five was found to be conserved in both species. The motif that was found to be conserved among more than one orthologus gene promoters in both *P. sojae* and *P. ramorum*, was tested for functionality. Promoter-reporter fusion assays, histochemical staining, RNA-blots showed that this was functional. Analyzing another motif that was associated with sporangia-specific expression, by fusing the motif to a minimal promoter, proved that the approach of using three different algorithms for identifying overrepresented motifs and integrating the three TFBS prediction methods indeed helped in making robust TFBS predictions.

Fig 1 legend:

The approach adopted for identification of robust putative TFBSs:

Shown is the approach adopted for this study, where stage-specific genes are identified from the microarray data, extract their promoters and look for over-represented motifs within the promoter data sets with three different motif-finding programs. The common motifs from the outputs are considered to be putative TFBSs, which are checked for their positional bias and evolutionary conservation. A motif that is overrepresented, shows a bias for a certain position and is conserved evolutionarily is considered to be a high confidence candidate for a TFBS, which is then tested for functionality by oligo-chimera assay.

Fig 1


Microarray data
Stage-specific promoter datasets
Motif finding programs
YMF
(enumerative search)
MEME
(expectation maximization)
Bioprospector
(Gibbs sampling)
Putative stage-specific TFBSs
Test for positional bias
Test for evolutionary conservation
Oligo-chimera assay

MATERIALS AND METHODS:

Selection of genes and development of promoter data sets:

94 genes that were >10 fold up-regulated in cleaving sporangia tissue when compared to sporangia, and 99 genes >10 fold up-regulated in sporangia when compared to hyphae, were selected from microarray data (Judelson et al., 2008).

Gene models created and maintained by the Broad Institute (http://www.broadinstitute.org/annotation/genome/phytophthora_infestans/MultiHome.html) were manually curated using stand-alone java applet, 'argoc-1'. Promoter regions (1 kb upstream of the coding region) for most of these genes were extracted from the same database with the help of an in-house PERL script; some of the promoters were extracted manually. The above-mentioned and all other PERL scripts, used for this study, were developed by myself.

Detection of overrepresented motifs:

Stand-alone versions of three different motif-finding programs were used. Background models created with 1000 base pairs (bp) upstream of the coding region for each of 100 randomly selected *P. infestans* genes. Also, degeneracy was allowed at two positions for two of these programs viz MEME and YMF. MEME (Multiple EM for Motif Elicitation; Baily and Elkan, 1994) version 4.3.0, with a minimum width (minw) of 5 and a maximum width (maxw) of 8, was used. The default gap opening cost (wg) and gap extension cost (ws) for multiple alignments were 11 and 1, respectively. The distribution of motifs (mod) used

was “anr” with a default E-value cut-off (evt) of 1e-05 and the maximum number of EM iterations (maxiter) was set at 5. The minimum sites for each motif (minsites) used were 5 with the rest of the parameters being default.

YMF (Yeast Motif Finder; Sinha and Tompa, 2000) version 3.0 was used specifying lenRegion (the length of the upstream regions in which motif is to be searched) for each set. The lenOligo i.e. the significant length or the number of non-spacer characters of the motifs to find was specified to be 8. The motifs were sorted by z-score using the sort (-sort) command line parameter.

BioProspector (Liu et al., 2001) Release 2 was used with motif width (-w) specified as 8. The number specified for top motifs to be reported (-r) was 100 and rest of the parameters used were default. BioProspector was run for ten times on each set of promoters and a PERL script was used to get rid of the redundant motifs and generate an output file with the non-redundant motifs from all ten runs.

Detection of positional biased and Phylogenetic footprinting:

To analyze the positional bias, the frequency of the motifs within each 50 base windows that the 1 kilobase promoters were divided into, were obtained by a PERL script developed in-house. The p-values for the frequencies within the five 200 base regions (four 50 base windows) were then calculated from their z-scores. For phylogenetic footprinting *P. infestans* gene promoters were aligned with the promoters of their orthologs in *P. sojae* and *P. ramorum* with the help of three different alignment programs on the web. The webtools that were used

were CLUSTALW (Thompson et al., 1994), MultAlin (Corpet, 1988) and DIALIGN (Morgenstern et al., 2004) and the source for the *P. sojae* and *P. ramorum* sequences was Joint Genome Institute (http://genome.jgi-psf.org). For CLUSTALW, the 'gap open' and the 'gap extension' penalties were set at 10 and 0.1 respectively. The 'gap open' and extension penalties for MultAlin were 10 and 1 respectively. All other parameters for both CLUSTALW and MultAlin were set to default. For DIALIGN all parameters including the threshold (T; default value, 0) and the regions of maximum similarity used (default value, 5) were set to default.

***P. infestans* strain, culture and manipulations:**

Isolates from the 1306 strain were cultured on rye-sucrose media at 18°C in the dark. *P. infestans* stable transformants were generated by the protoplast method described previously by Judelson et al. (1993). Non-sporulating mycelia were obtained by inoculating clarified rye-sucrose broth with a sporangial suspension, followed by 48 hours of incubation. Cultures on rye-sucrose agar plates that were 9 to 11 days old, were used for sporulating mycellia. Sporangia were obtained from 7 to 9 day old sporulating mycelia by adding water, rubbing with a glass rod, and passing the fluid through a 50-µm mesh to remove hyphal fragments. To induce cleavage, sporangia were placed in 100 mm petri plates which were kept on ice for ~60 mins. RNA was extracted using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA).

Construction of vectors:

Clone 38.2, also known as pNPGUS, is a modified version of pOGUS (Clone38; Cvitanich C, Judelson HS, 2003) vector which contains a promoterless GUS (β-glucuronidase) reporter gene and a neomycin phosphotransferase (nptII) gene driven by the ham34 promoter for G418 selection. To construct the Clone38.2 vector, I first removed the multiple cloning site between NotI and EcoRI restriction sites downstream of the GUS reporter gene and then re-ligated the vector, after blunting the cut sites with Klenow polymerase. This version was named Clone 38.1. A double stranded oligonucleotide was then made by annealing two synthetically designed single stranded oligonucleotides (Table 1). The double stranded oligonucleotide contained the entire region with multiple cloning sites, which was removed from pOGUS. The oligonucleotide in addition to the multiple cloning sites contained two stop codons in two different frames at the 5' end, downstream of the overhang for the ApaI site. Overhang for the ClaI site was at the 3' end (Table 1). This double stranded oligonucleotide was then inserted between the ApaI and ClaI sites upstream of the GUS reporter gene.

For construction of the *NIFS*+Clone38.2 vector, a 74 bp long minimal promoter of the *P. infestans NIFS* gene (nuclear LIM factor interactor-interacting protein, spore-specific form) was added to the Clone 38.2 vector, upstream of the GUS gene, using the XmaI and EcoRI restriction digestion sites.

Promoter::Reporter plasmid construction and sequencing:

Full length (500 bp upstream of ATG) and different-sized 5' promoter fragments (for serial deletion analysis) from the cleavage specific PITG_16321 gene were obtained by PCR amplification, using 1306 genomic DNA, Taq DNA polymerase, and primers (Table 1) specific to the promoter regions based on the sequence provided by the Broad Institute database. Primers (Table 1) for motif specific mutation analyses and oligos for oligo-chimera and EMSA assays along with those for serial deletion were designed with the primer designing Oligo software version 4.0, developed by "Molecular Biology Insights" (Cascade, CO, USA).

The fragments for serial deletion and motif-specific mutation analysis were cloned into transformation vector Clone38.2 (pNPGUS) in between the XbaI and the EcoRI restriction sites. Oligos for the oligo-chimera assay were cloned into the *NIFS*+Clone38.2 plasmid in between the XbaI and XmaI restriction sites. Chemically competent DH5∝ cells were used for bacterial transformation. Clones found to be positive by restriction digestion were sequenced.

Table 1:

Primers and oligonucleotides used in this study:

Shown below is a list of primers and oligonucleotides used for this study. All of primers/oligonucleotides are in 5'-3' direction. The number 4671 refers to the Pi (cDNA) number of the PITG_16321 gene, from the microarray study of Prakob and Judelson (2007). The last column describes the usage of these primers/oligonucleotides.

Primer/Oligo name	Primer/Oligo 5'-3'	Used for
C38.2U	CTAAAATAGATAAGGCGGCCGCTCTAG AACTAGTGGATCCCCCGGGCTGCAGGA ATTCAATAAAAT	Construction of Clone38.2 vector
C38.2L	CGATTTTATTGAATTCCTGCAGCCCGG GGGATCCACTAGTTCTAGAGCGGCCG CCTTATCTATTTTAGGGCC	Construction of Clone38.2 vector
4671U	GCTCTAGAAGAAACTGAGCCTCG GTATGA	Amplifying 500 bp upstream of 4671 gene
4671L	CGGAATTCAGAAATGCTAAGCGAA GACTG	Amplifying 500 bp upstream of 4671 gene
Del1	GCTCTAGAGCCGTCGGTATCCAAG AGGTA	Amplifying 312 bp upstream of 4671 gene
Del2	GCTCTAGAGCCTCCCTGCTGTCG TCCTC	Amplifying 187 bp upstream of 4671 gene
Del3	GCTCTAGAGCGGGTCCGTCTTCTA GTCCA	Amplifying 104 bp upstream of 4671 gene
Mut1	GCTCTAGAGCAGAAACCTCACCGTCCT CGAACCACA	Amplifying 187 bp upstream of 4671 gene with 1st mutation
Mut2	GCTCTAGAGCCTCCCTGCTGTACAATCC GAACCACATGGCAT	Amplifying 187 bp upstream of 4671 gene with 2nd mutation
Mut2	GCTCTAGAGCCTCCCTGCTGTCGTCCT CGTTGTTGTATTTATGTGCTCCCATCCG	Amplifying 187 bp upstream of 4671 gene with 3rd mutation

Primer/Oligo name	Primer/Oligo 5'-3'	Used for
NifSU	TCCCCCCGGGGGGATTGAAGATTCGACGG	Amplifying *NIFS* minimal promoter
NifSL	GGAATTCCCGTTGTAGCCGTGGT	Amplifying *NIFS* minimal promoter
BLKU	CTAGACTCCCTGCTGTCGTCCTCGAACCACATGGCTTCCCGTCTTCTCGTCTC	Oligo-chimera with all 3 conseved blocks (motifs)
BLKL	CCGGGAGACGAGAAGACGGGAAGCCATGTGGTTCGAGGACGACAGCAGGGAGT	Oligo-chimera with all 3 conseved blocks (motifs)
OC-upper	CTAGACGTCCTCGGGTTGGTGCAATTTCCCGTCTTCTCGTCTC	Oligo-chimera with conseved motif CGTCCTCG
OC-lower	CCGGGAGACGAGAAGACGGGAAATTGCACCAACCCGAGGACGT	Oligo-chimera with conseved motif CGTCCTCG
SP2_OC upper	CTAGACTTCAACGAGTTGGTGCAATTTCCCGTCTTCTCGTCTACGTCCC	Oligo-chimera with conseved motif CTTCAAC
SP2_OC lower	CCGGGGGACGTAGACGAGAAGACGGGAAATTGCACCAACTCGTTGAAGT	Oligo-chimera with conseved motif CTTCAAC
EMSA_CL_SPEC_UP	CTCCTAGACTCCCTGCTGTCGTCCTCGAACCACATGGCTCCCGTCT	Specific probe of CGTCCTC motif
EMSA_CL_SPEC_LO	AGACGGGAGCCATGTGGTTCGAGGACGACAGCAGGGAGTCTAGGAG	Specific probe of CGTCCTCG motif
EMSA_CL_NS_UP	TCGAGTACTTCTACACCATCATGGCACTGTACTCCTCTAGTCTGTA	Non-Specific probe of CGTCCTCG motif
EMSA_CL_NS_LO	TACAGACTAGAGGAGTACAGTGCCATGATGGTGTAGAAGTACTCGA	Non-Specific probe of CGTCCTCG motif
EMSA_CL_MUT_UP	CTCCTAGACTCCCTGCTGTACAATCCGAACCACATGGCTCCCGTCT	Mutated probe of CGTCCTCG motif
EMSA_CL_MUT_LO	AGACGGGAGCCATGTGGTTCGGATTGTACAGCAGGGAGTCTAGGAG	Mutated probe of CGTCCTCG motif
EMSA_SP2_SPEC_UP	CTGCCTCCTCCAATTTGCACTTCAACTTGTGTAGCCATCTGACGACC	Specific probe of CTTCAAC motif
EMSA_SP2_SPEC_LO	GGTCGTCAGATGGCTACACAAGTTGAAGTGCAAATTGGAGGAGGCAG	Specific probe of CTTCAAC motif
EMSA_SP2_Mut_UP	CTGCCTCCTCCAATTTGCAAGGTCAATTGTGTAGCCATCTGACGACC	Mutated probe of CTTCAAC motif
EMSA_SP2_Mut_LO	GGTCGTCAGATGGCTACACAATTGACCTTGCAAATTGGAGGAGGCAG	Mutated probe of CTTCAAC motif

Gene expression analysis:

Gene expression analyses were done by two different methods. The first one was histochemical staining of stage specific tissues for β-glucuronidase (GUS), performed as described by Judelson et al. (1993). The staining solution is made up of 50 mM sodium phosphate (pH 7.0), 0.1% Triton X-100, 0.1% X-Gluc (bromochloroindoyl-b-glucuronide) in dimethyl formamide, 5 mM potassium ferricyanide, 5 mM potassium ferrocyanide. Tissues were stained at 37°C in the dark, overnight. Northern blotting was the second method that was used for gene expression analyses and was performed as described (Judelson & Roberts, 2002). Five micrograms of total RNA was separated on 1.2% agarose/6.6% formaldehyde gels. This was then transferred to nylon membranes by capillary blotting in 20x SSPE (3.6M NaCl, 0.2M sodium phosphate, 0.02M EDTA pH 7.7). The membranes were then fixed by UV crosslinking, and hybridized overnight, at 65°C with ^{32}P-labeled probes made from β-glucuronidase (GUS) DNA. Two rounds of washing were carried out, the membrane was first washed in 1x SSPE, 0.2% sodium dodecyl sulfate (w/v), and 0.1% sodium pyrophosphate at 65°C. The second wash was in 0.2x SSPE, 0.2% sodium dodecyl sulfate (w/v), and 0.1% sodium pyrophosphate at 65°C. The blots were placed under phosphor screens in autoradiography cassettes and left overnight in the dark. Signals were detected by phosphorimager analysis using Quantity One software developed by BIO-RAD (Philadelphia, PA, USA).

Electrophoretic mobility shift assay (EMSA):

Nuclear protein isolation and EMSA were performed as described by Ah Fong et al. (2007), except that heparin agarose was not used for the extractions. In short, EMSA involved mixing 5 μg of nuclear protein with 1 μg poly dI-dC,1.6 ng of γ^{32}P-ATP, 15 mM HEPES pH 7.9, 25 mM $MgCl_2$, 100 mM KCl, 15% glycerol, 1 mM DTT for 15 mins at room temperature followed by 30 min on ice, followed by electrophoresis on a 4.5% acrylamide gel in 0.5x TBE buffer (89mM Trisborate, 2mM EDTA) for 3 h at room temperature. The gel was dried for an hour, placed under the phosphor screens in autoradiography cassettes and left for overnight in the dark. The screen was then analyzed with a phosphorimager. For competition assays, protein was incubated with unlabeled DNA for 15 min and then with the labeled probe for 30 mins in ice. Double-stranded oligonucleotides described in the 'Results' section were used as hot probes and cold competitors.

RESULTS:

Identification of cleavage genes and development of promoter dataset:

Ninety-four genes (Table 2) that were more than 10 fold up-regulated in cleaving sporangia when compared to sporangia were considered for this study. All 94 gene models were manually curated before extraction of their promoters (1 kb upstream of the translation start site). The promoters of the 17 genes for which the translation start site (ATG) had to be altered during the manual

curation were extracted manually (from the Broad Institute database). The rest of the promoters were extracted with the help of a PERL script developed in-house.

Detection of overrepresented motifs in cleavage-induced gene promoter set:

Overrepresented motifs within the promoter dataset were detected by three different motif finding programs as mentioned in the "Materials and Methods' section. MEME detects overrepresented motifs using the expectation maximization algorithm, YMF based on an enumerative approach, and BioProspector uses a Gibbs sampling technique. MEME was set to detect 100 most overrepresented motifs. YMF detected 221 motifs and there were 82 non-redundant motifs from ten runs of BioProspector. The PERL script that was used to detect common motifs found 35 motifs which were detected by at least two out of the three programs. Six of these motifs were merged manually with six other motifs within the set. Two motifs were manually merged only when, no more than two bases among the motifs were different. The motifs were detected as different motifs as in most cases these had different terminal bases.

Table 2:

The following table shows the PITG numbers of the 94 cleavage-induced genes used for this study:

#	PITG #	#	PITG #	#	PITG #
1	PITG_00591	**33**	PITG_03162	**65**	PITG_12903
2	PITG_01266	**34**	PITG_03467	**66**	PITG_13036
3	PITG_02008	**35**	PITG_03525	**67**	PITG_13115
4	PITG_03346	**36**	PITG_03590	**68**	PITG_13419
5	PITG_04322	**37**	PITG_04281	**69**	PITG_13601
6	PITG_04477	**38**	PITG_04701	**70**	PITG_13644
7	PITG_05149	**39**	PITG_04999	**71**	PITG_13755
8	PITG_05203	**40**	PITG_05204	**72**	PITG_13881
9	PITG_05205	**41**	PITG_05296	**73**	PITG_14228
10	PITG_05714	**42**	PITG_05670	**74**	PITG_15282
11	PITG_05738	**43**	PITG_07355	**75**	PITG_16321
12	PITG_06049	**44**	PITG_07444	**76**	PITG_16473
13	PITG_06835	**45**	PITG_07961	**77**	PITG_16727
14	PITG_07345	**46**	PITG_08258	**78**	PITG_16967
15	PITG_11239	**47**	PITG_08404	**79**	PITG_17344
16	PITG_11504	**48**	PITG_08707	**80**	PITG_17420
17	PITG_17675	**49**	PITG_09410	**81**	PITG_17591
18	PITG_17951	**50**	PITG_09899	**82**	PITG_18174
19	PITG_20590	**51**	PITG_09979	**83**	PITG_18240
20	PITG_20710	**52**	PITG_10337	**84**	PITG_18386
21	PITG_03034	**53**	PITG_10507	**85**	PITG_18393
22	PITG_05111	**54**	PITG_10523	**86**	PITG_18428
23	PITG_06236	**55**	PITG_10571	**87**	PITG_18680
24	PITG_06965	**56**	PITG_10630	**88**	PITG_19451
25	PITG_00321	**57**	PITG_10847	**89**	PITG_19483
26	PITG_00539	**58**	PITG_11102	**90**	PITG_20681
27	PITG_00891	**59**	PITG_11238	**91**	PITG_20886
28	PITG_02028	**60**	PITG_11470	**92**	PITG_21207
29	PITG_02029	**61**	PITG_12293	**93**	PITG_21452
30	PITG_02030	**62**	PITG_12352	**94**	PITG_11400
31	PITG_02110	**63**	PITG_12507		
32	PITG_02227	**64**	PITG_12524		

Analysis of positional bias for five motifs overrepresented in cleavage induced gene promoters:

Five overrepresented motifs (Table 3) from the list of 29 motifs were randomly selected for analyses of their positional bias. A PERL script developed in-house, was used to compute the frequencies of each of the motifs within a 50 base (bp) window that the 1 kilobase (kb) promoters were divided into. These frequencies were used to calculate the positional bias of each motif within the five 200 base regions (four 50 base windows). Equality of proportions for the observed and expected values were calculated to get the z-scores, which were then used to calculate the p-values. All five motifs showed a clear bias for one or more 200 base regions. The 'TACATGTA' and the 'AGAGAGAG' motifs showed a bias for 400 bases (two 200 base regions; Table 3), 201 to 600 and 1 to 400 bases upstream of ATG respectively. The 'TCGTC[GT]TC motif showed a bias for the first 600 bases upstream of ATG. The other two motifs, 'GATGCTG' 'CGTCCTCG', showed a clear bias for only one 200 base region (Table 3). The 'CGTCCTCG' motif that showed a bias for a single 200 base region (Fig 2A) that was closest to the translation start site (1-200 bases) and was analyzed further.

Promoter sets, from 99 sporangia-induced genes, and all *P. infestans* genes (18124 when the study was conducted) were searched to find out if the 'CGTCCTCG' motif showed a bias for any positions within these sets (Fig 2A). The bias for its reverse complement was also checked in all the three sets. It was observed the motif 'CGTCCTCG' and its reverse complement 'CGAGGACG' had

a clear bias for the first 200 bases upstream of ATG, within the cleavage-induced genes promoters. The bias of 'CGTCCTCG' was six times that of 'CGAGGACG' within 100 to 200 bases upstream of ATG. No significant bias could be detected within the sporangia or the total gene promoter sets (Fig 2A).

Analysis of evolutionary conservation of 'CGTCCTCG' motif:

Orthologs of four genes that carried this motif, within the first 200 bases upstream of ATG, were identified in *P. sojae* and *P. ramorum,* by BLASTP (Altshul et al., 1997), from the Joint genome Institute (JGI; http://genome.jgi-psf.org/) database. The promoters of these orthologs were extracted manually and aligned along with the *P. infestans* gene promoters by three different alignment programs as mentioned in the "Materials and methods' section. The 'CGTCCTCG' motif was found to be conserved in three out of the four gene promoters aligned, in all three Phytophthora species. One out of the three genes mentioned above, PITG_16321 (Fig 2B), was chosen for functional analyses.

Table 3:

Distribution of five overrepresented motifs within promoters of genes specific to cleaving sporangia:

The table shows the distribution of five of the overrepresented motifs within the cleavage-induced gene promoters. The 1kb regions are divided into five 200 nt windows. The raw frequency of the motifs within the windows is shown along with the total number of hits and the positions for which the motifs show a bias. The 'CGTCCTCG' (in bold) was chosen for further analyses. The numbers in bold specify the region of bias for each motif (5' to 3' direction).

	MOTIF				
Bases from ATG	AGAGAGAG	**CGTCCTCG**	TACATGTA	TCGTC[GT]TC	GATGCTG
1-200	**9**	**7**	1	**6**	3
201-400	**8**	1	**7**	**5**	**12**
401-600	3	0	**4**	**5**	5
601-800	0	0	2	1	2
801-1000	0	2	**5**	1	2
Total hits	20	12	19	18	24

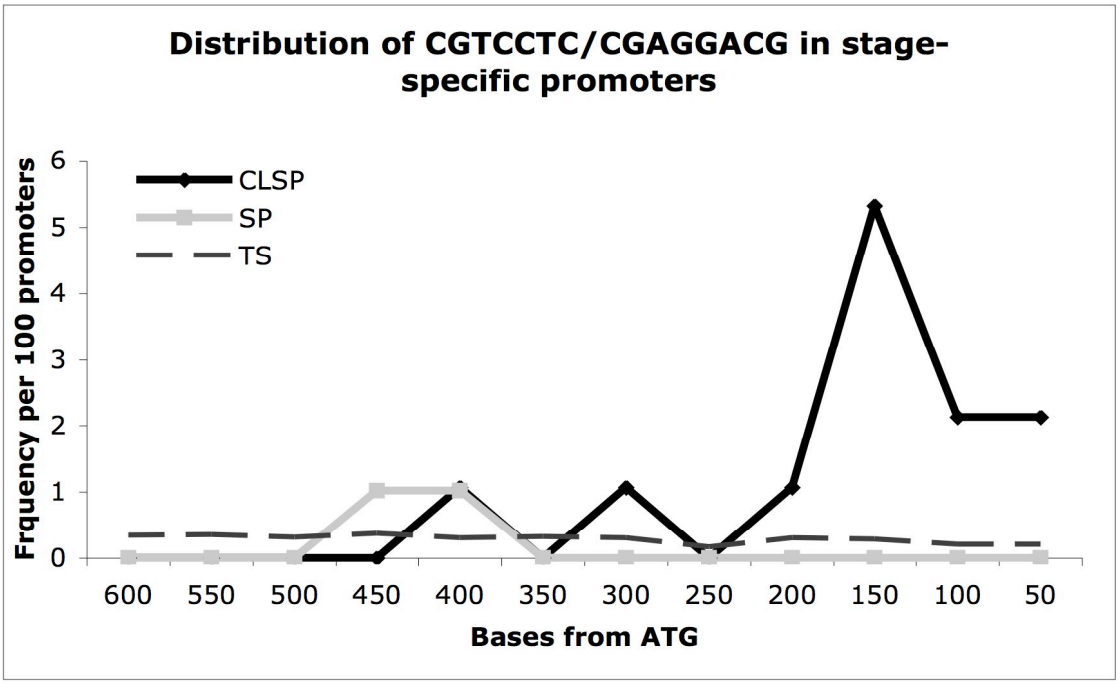
Fig 2 legend:

Positional bias and evolutionary conservation of 'CGTCCTCG' motif:

A) The figure shows the positional bias of the motif 'CGTCCTCG', for the first 200 bases upstream of ATG, within the cleavage-induced gene promoters (CLSP) set. Also shown is the distribution of this motif within the sporangia-induced gene promoters (SP) and the promoters of all *P. infestans* genes (TS). The frequencies for the 'CGTCCTCG' motif and its reverse complement 'CGAGGACG' within 'the first 100 bases windows were similar but the frequency of 'CGTCCTCG' was 6 times that of 'CGAGGACG' within '100-200' window. B) Shown is the evolutionary conservation of the motif 'CGTCCTCG' in the promoters of *P. sojae* and *P. ramorum* orthologs of the PITG_16321 gene in *P. infestans*. The three positions indicated with numbers starting with the minus signs show where the deletions were made.

Fig 2

A


Distribution of CGTCCTC/CGAGGACG in stage-specific promoters
Frquency per 100 promoters
0
1
2
3
4
5
6
CLSP
SP
TS
600
550
500
450
400
350
300
250
200
150
100
50
Bases from ATG


B
P.infestans
P.sojae
P.ramorum
consensus
-312
-187
Motif
-104

Table 4:

Transformants used for the functional analyses of 'CGTCCTCG' and 'CTTCAAC' motifs:

Shown below is the list of transformants used for the functional analyses of the cleaving sporangia-specific 'CGTCCTCG' motif and the sporangia-specific 'CTTCAAC' motif.

Cleaving sporangia specific motif: CGTCCTCG					
Construct name	**Transformant #**				
Full length	1.2	1.3	1.4		
Del 1	188-8	188-48	188-50		
Del2	313-2	313-27	313-52		
Del3	396-6.1	396-30	313-36		
Mut1	4.2	1.3			
Mut2	17.1	17.2			
Mut3	5.1	2.3			
OC	6.3	8.4			
Block 2	3.1	9.1	11.1	2.3	
-ve control	H2				
Sporangia specific motif: CTTCAAC (SP2)					
Construct name	**Transformant #**				
SP2:	3.8	3.16	4.28	4.29	2.1

Expression analyses of full length (500 bases) PITG_16321 promoter:

The full length promoter fragment was amplified by PCR using '4671U' and '4671L' upper and lower primers (Table 1), with isolate 1306 genomic DNA as template. This 500 bp fragment was then inserted into Clone 38.2 vector to create a promoter::reporter plasmid (Fig 3A). *P. infestans* was then transformed with the plasmid DNA. The transformants were subcultured in rye-sucrose agar plates with G418 for selection. Sporangia and cleaving sporangia were obtained and stained for GUS expression as described previously. Staining showed that GUS was expressed (Fig 3B, 3D) in the cleaving sporangia of nine out of 62 transformants. The rest of the transformants did not show expression in any of the tissues presumably due to position effects. Hyphae and sporangia (Fig 3C) from the same transformants did not show any GUS expression (not shown), suggesting that the promoter, which is from a gene induced in cleaving sporangia, was driving the expression of the reporter gene only during that stage. RNA was extracted for blotting from the sporangia and cleaving (chilled) sporangia tissues of three of the transformants (Table 4) that had shown GUS expression. RNA was transferred into a membrane, crosslinked and hybridized with ^{32}P-labeled, randomly primed probes made with DNA from the GUS gene. Signals for GUS expression were visible in the cleaving sporangia RNA of all three transformants, but no signal could be detected from any of the sporangia RNA. This showed that the 500 bp promoter fragment of the PITG_16321 gene, which had the 'CGTCCTCG' motif, was able to drive GUS expression only in

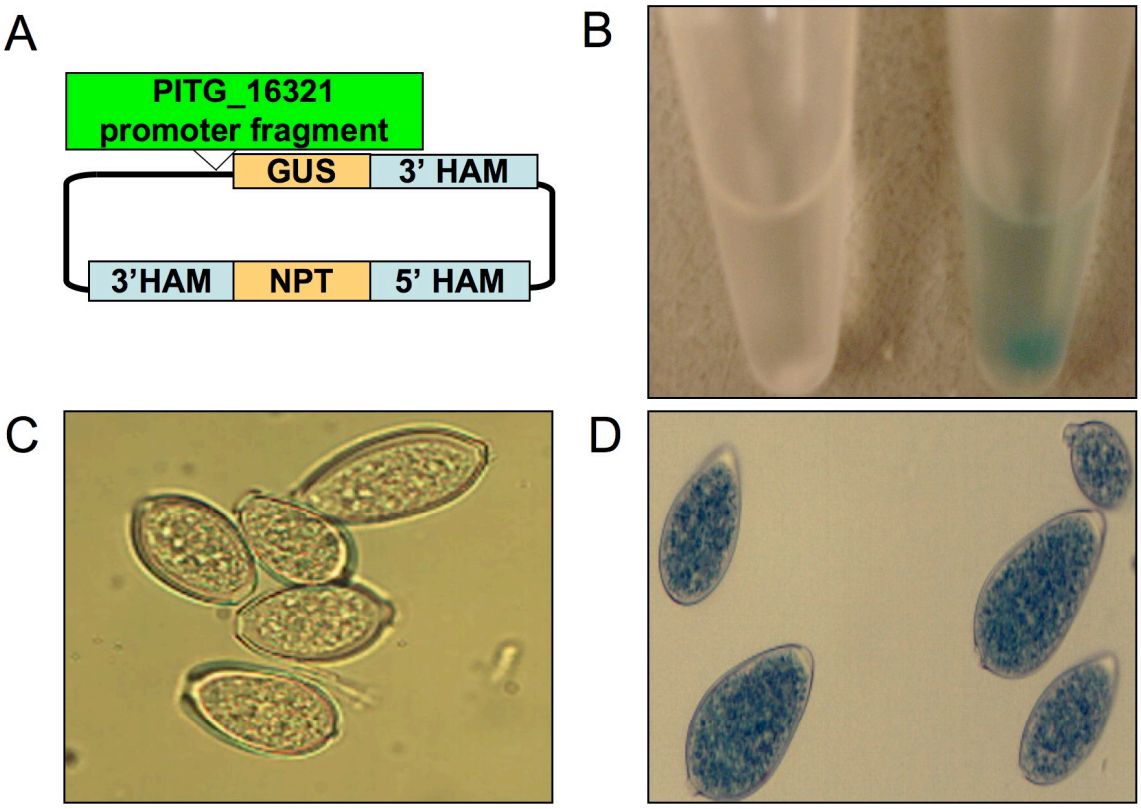
cleaving sporangia suggesting that the motif has a role to play in the expression of genes during cleavage.

Fig 3 legend:

Conformation of functionality for full length (500 bp) PITG_16321 promoter:

A) Shown is a diagram of the promoter::reporter plasmid where a promoter fragment is inserted in front of the GUS reporter gene. The NPT gene driven by the HAM 34 promoter is used as a selection marker. B) Shown are the unstained sporangia (left) and stained cleaving (chilled) sporangia (right) from 1306 strain transformed with the plasmid where the PITG_16321 promoter fragment (500 bp) is inserted in front of the GUS reporter gene, after overnight incubation with the GUS staining solution. C) Shown are unstained sporangia, after overnight incubation with the GUS staining solution, under the microscope. D) Shown are stained cleaving sporangia, after overnight incubation with the GUS staining solution, under the microscope.

Fig 3



Serial deletion of PITG_16321 promoter:

Three 5' to 3' deletion fragments were generated from 1306 genomic DNA by PCR as described previously. The deletions were based on the conservation shown in Fig 2B; '4671L' was used as the lower primer for all three fragments. 'Del1', 'Del2' and 'Del3' (Table 1) were used as upper primers for 312 bp long 'del1', 187 bp long 'del2', and 104 bp long 'del3' fragments, respectively. As evident from Fig 2B, 'del1' and 'del2' contained the 'CGTCCTCG' motif, but 'del3' did not. Each of the three fragments was inserted into Clone38.2 to get three different plasmids. For the 'del1' fragment 49 transformants were analyzed, 58 and 65 transformants each were analyzed for 'del2' and 'del3' fragments respectively. The results from GUS staining and northern blots of *P. infestans,* transformed with the three above mentioned plasmid DNAs, showed that GUS expression was driven by the 'del1' (12 transformants showed staining, all three tested by northern were positive) and 'del2' (11 transformants showed staining, all three tested by northern were positive) promoter fragments in cleaving sporangia (Figs 4A, 4B), but not by the 'del3' fragment (6 transformants showed light GUS staining but none of the three tested by northern were positive; Fig 4C). No signal was detected from any of the sporangial samples taken from the same transformants. Ribosomal RNA was used as a loading control. This proved that the region between 187 bases (-187) and 104 bases (-104), upstream of PITG_16321 translation start site, was responsible for the expression from the GUS gene.

Oligo-chimera assay with a block containing the 'CGTCCTCG' motif:

The 'CGTCCTCG' motif is present within the -187 to -104 region (Fig 2B), but the alignment shows that there are two other conserved motifs, viz. 'CCCTGCTG' and 'ACCACATGGC', on either side of this motif. There is also another conserved motif 'ACTCTGCC', 51 bp downstream of 'CGTCCTCG' within this region. Therefore, to check if the 'CGTCCTCG' functioned on its own, or was influenced by the other sequences, a 47 bp double stranded DNA fragment ('BLKU' 'BLKL' oligos annealed; Table 1) carrying the 'CGTCCTCG' motif along with the two other conserved motifs next to it was made. This was then inserted in front of the *NIFS* minimal promoter (which by itself cannot drive GUS expression; Ah-Fong et al. 2007) within the *NIFS*+Clone38.2 vector, using the XbaI and XmaI restriction sites. The results (Fig 4D) showed that this fragment was able to drive GUS expression. It should be mentioned that only 'CGTCCTCG' motif was found to be overrepresented within the cleaving sporangia promoter set and not the other two. Therefore, it was most likely that the 'CGTCCTCG' motif was driving the expression of the GUS reporter gene on its own with the help of core promoter elements (required for initiation of transcription), within the *NIFS* minimal promoter.

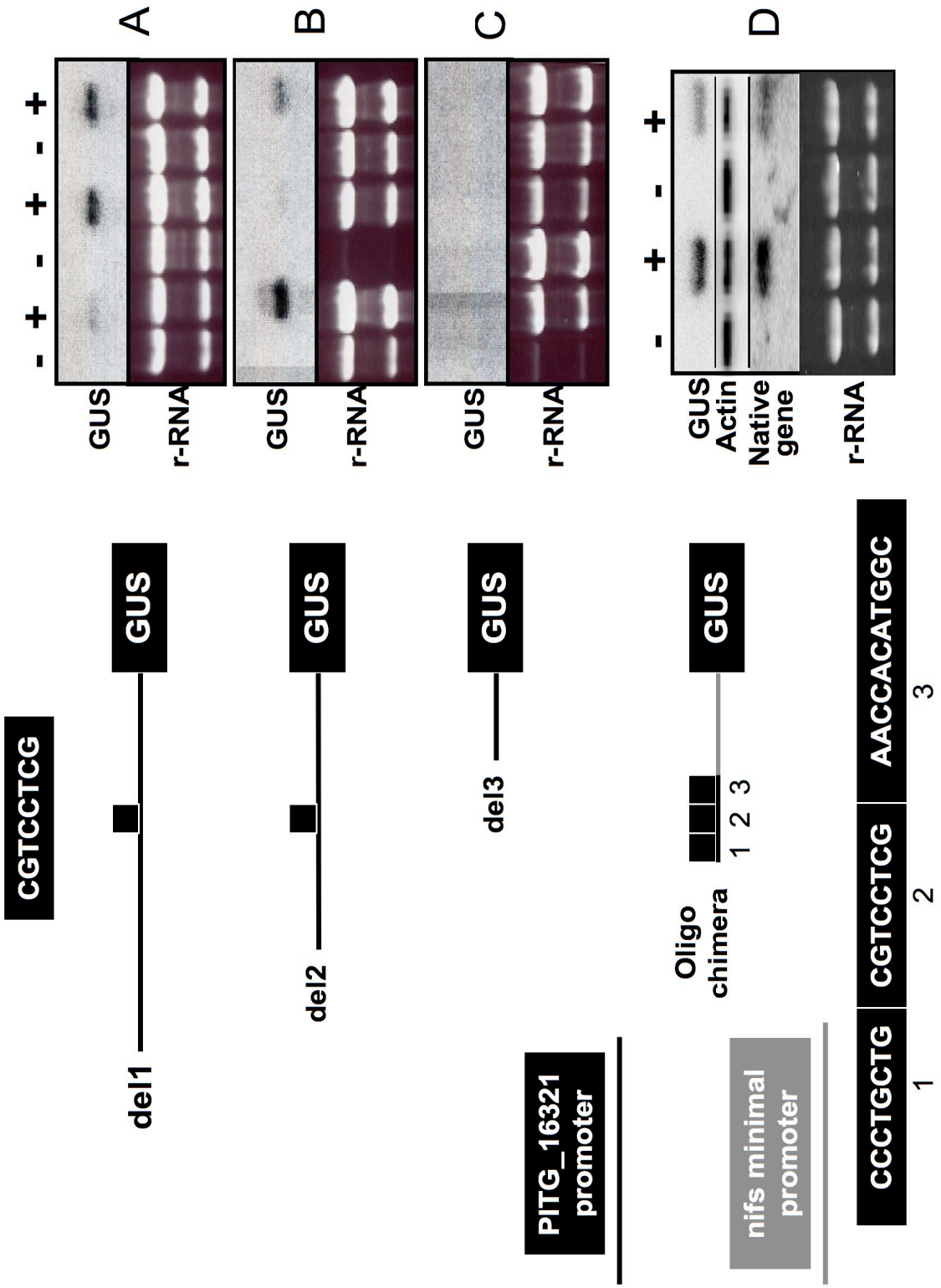
Fig 4 legend:

Deletion analyses for PITG_16321 promoter and oligo-chimera assay with a block containing all 3 conserved motifs:

'CGTCCTCG' (2) is the cleavage-specific motif, 'CCCTGCTG' (1) and 'ACCACATGGC' (3) are the two other conserved motifs.

A) Shown is the result of GUS expression analysis by RNA blot for the 312 bp long 'del1' promoter fragment. The (+) sign denotes RNA from cleaving (chilled) sporangia as a result of cold treatment. RNA from sporangia without cold treatment is shown by the (-) sign. ribosomal-RNA was used as loading control. RNA from three stable transformants viz. '188-8', '188-48' and '188-50' were loaded from left to right respectively. B) Shown is the result for the 187 bp long 'del2' promoter fragment. RNA from three stable transformants viz. '313-2', '313-27' and '313-52' were loaded from left to right respectively. C) The results for 104 bp long promoter fragment is shown here. RNA from three stable transformants viz. '396-6.1', '396-30' and '396-36' were loaded from left to right respectively. D) Shown is the result from the oligo-chimera assay with a 47 bp long DNA fragment carrying all thee conserved motifs mentioned above. RNA from two stable transformants 'OC 6.3' and 'OC 8.4' was loaded from left to right, ribosomal-RNA and actinA were used as loading controls. Native gene was used as a control for stage specific samples.

Fig 4


A
B
C
D
- + - + - +
GUS
r-RNA
GUS
Actin
Native gene
CGTCCTCG
del1
del2
del3
GUS
Oligo chimera
1 2 3
PITG_16321 promoter
nifs minimal promoter
CCCTGCTG
CGTCCTCG
AACCACATGGC
1
2
3

Motif-specific mutations:

To check if the 'CGTCCTCG' motif could function on its own, motif-specific mutations and another oligo-chimera assay with the oligo carrying the 'CGTCCTCG' motif only, were performed. Promoter fragments for motif-specific mutations were designed in such a way that each fragment had all bases of one motif mutated (all bases were changed), but that of the other two remained unchanged. Using 'Mut1', 'Mut2' and 'Mut3' (Table 1) as the upper primers and '4671L' as the lower primer all three fragments were generated by PCR, using the previously made 'del1' promoter fragment as template DNA. The results from the analyses of 'Mut1' transformants showed that GUS expression in cleaving sporangia was not affected by mutating the 'CCCTGCTG' motif, as the 'Mut1' promoter fragment was able to drive GUS expression based on RNA blot analysis (Fig 5A). No signal could be detected in any of the sporangia samples; ribosomal RNA along with actinA (PITG_ 15117), were used as loading controls. The native gene was used to make sure that the samples were actually sporangia and cleavage as the native gene being cleavage specific would express only in cleaving sporangia tissue. I wanted to confirm that the sporangia samples did not start to cleave. The 'Mut2' promoter fragment which had a mutated 'CGTCCTCG' motif, was unable to drive GUS expression in cleaving sporangia, showing that this is the functional motif (Fig 5B) was responsible for driving the GUS gene and acted on its own. For 'Mut3' ('AACCACATGGC' mutated), expression in cleaving sporangia was not eliminated (Fig 5C). But,

some signals could be detected in sporangia samples too. The quantification of signals, from the native gene when compared to that from the GUS reporter in case of different samples, were ambiguous, suggesting that the signals in sporangia might have been due to the samples getting chilled and starting to cleave. Another probable explanation for the signals in sporangia is that the 'AACCACATGGC' motif that was mutated in the 'Mut3, fragment acts as a repressor for some other motif in sporangia.

Oligo-chimera assay with 'CGTCCTCG' motif:

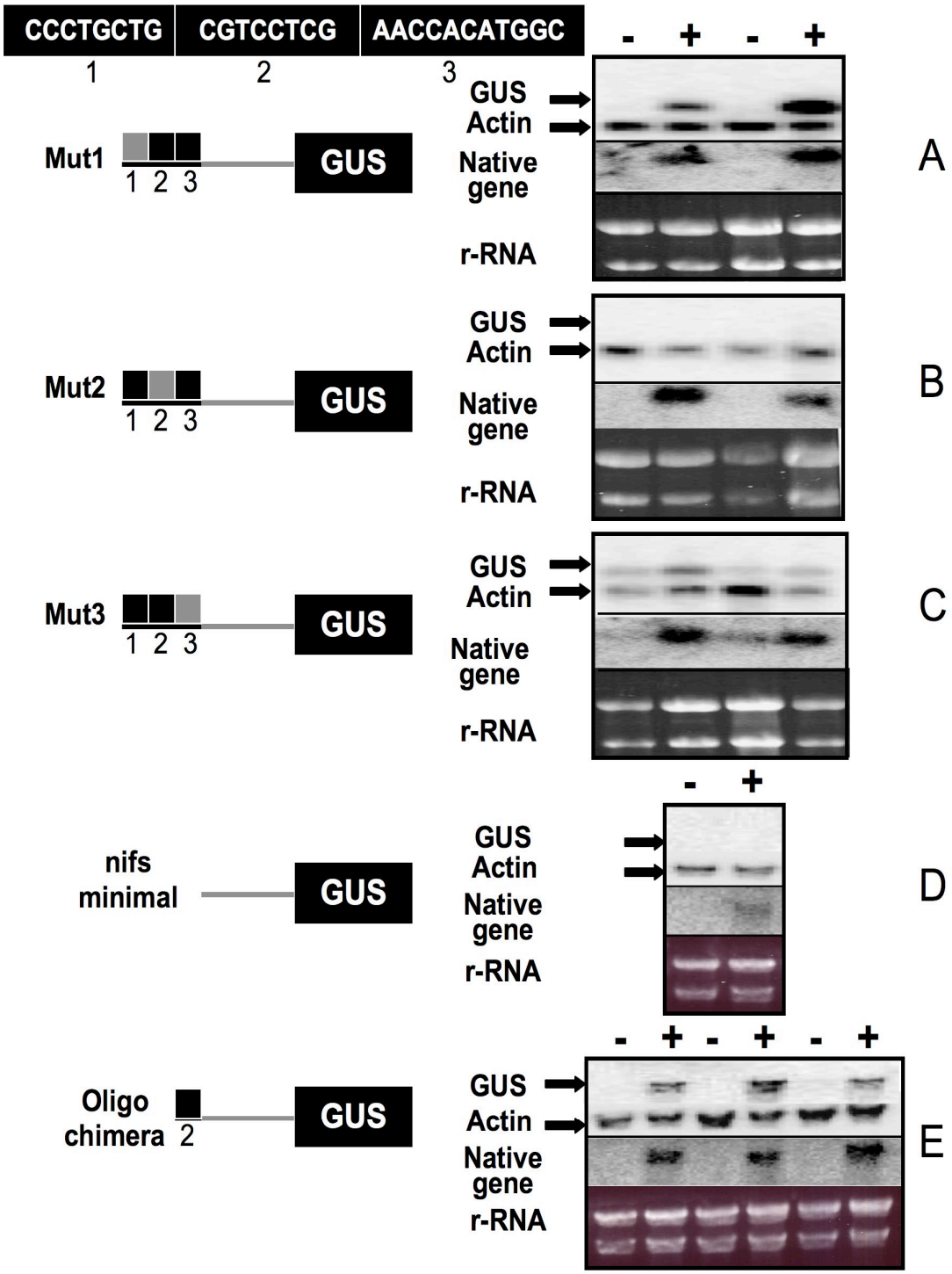
As a confirmatory test another oligo-chimera assay was done with a 37 bp long oligo ('OC_upper' and 'OC_lower' oligos annealed; Table 1) containing just the 'CGTCCTCG' motif, along with 25 random bases. The results (Fig 5E) confirmed that this motif by itself was sufficient to drive GUS expression in cleaving sporangia. It was also confirmed that the *NIFS* minimal promoter used for this study could not drive GUS expression on its own (Fig 5D).

Fig 5 legend:

Mutation analyses for PITG_16321 promoter and oligo-chimera assay with an oligo containing the 'CGTCCTCG' motifs:

A) Shown is the result of GUS expression analysis by RNA blot for 1306 transformants carrying the 'Mut1' promoter fragment, which had a mutated 'CCCTGCTG' motif. The (+) sign denotes RNA from cleaving (chilled) sporangia as a result of cold treatment. RNA from sporangia without cold treatment is shown by the (-) sign. RNA from two stable transformants '1.3' and '4.2' were loaded from left to right. Staining of r-RNA with Ethidium Bromide and hybridization with actinA probe served as loading controls. Native gene was used as a control for stage-specific samples. B) Shown is the result for mutating the 'CGTCCTCG' motif within the promoter fragment. RNA from two stable transformants '17.1' and '17.2' were loaded from left to right. C) Shown is the result for mutating the 'AACCACATGGC' motif within the promoter fragment. RNA from two stable transformants '2.3' and '5.1' were loaded from left to right. D) Shown is the result from GUS expression by the *NIFS* minimal promoter. E) Shown is the result from the oligo-chimera assay with a 37 bp long oligo carrying the 'CGTCCTCG' motif by itself. RNA from three stable transformants '2.3' and '9.1' and 11.1 were loaded from left to right.

Fig 5


CCCTGCTG
CGTCCTCG
AACCACATGGC
1
2
3
- + - +
Mut1
1 2 3
GUS
GUS
Actin
Native gene
r-RNA
A
Mut2
1 2 3
GUS
GUS
Actin
Native gene
r-RNA
B
Mut3
1 2 3
GUS
GUS
Actin
Native gene
r-RNA
C
- +
nifs minimal
GUS
GUS
Actin
Native gene
r-RNA
D
- + - + - +
Oligo chimera
2
GUS
GUS
Actin
Native gene
r-RNA
E

Electrophoretic mobility shift assay to test the binding affinity of 'CGTCCTG' motif:

To test if the 'CGTCCTCG' motif has any binding affinity for any of the proteins within a nucleus-enriched extract a 46 bp double-stranded oligonucleotide was designed by annealing single-stranded 'EMSA_CL_SPEC_UP' and 'EMSA_CL_SPEC_LO' (Table 1) oligonucleotides. This was run on a 10% polyacrylamide gel, purified, and then radio-labeled. Binding reactions as mentioned in the 'Materials and methods' section were done with nuclear extracts from sporulating mycelia, sporangia and cleaving sporangia tissues. Two different bands (Fig 6A), an upper band 'a' and a lower band 'c' could be detected for the sporangia and cleaving sporangia nuclear extracts. A different band 'b', at a position slightly lower than 'a', could be seen for the nuclear extracts from sporulating mycelia, and this was tested by repeating the experiment three times.

To test the specificity of these bands, two separate competition analyses were done, one with nuclear extracts from cleaving sporangia (Fig 6A) and the other with that from sporulating mycelia (Fig 6B). The results showed that both band 'a' and 'c' in cleaving sporangia were due to some specific binding activities between the 'CGTCCTCG' motif and some proteins. This could be deduced as higher concentrations (5x, 25x and 125x) of unlabeled specific probe were able to out-compete the labeled probe. As a result the signals from the radioactive probe became weaker and was lost ultimately (Fig 6A). In contrast, the signals

from the labeled specific probe was not affected by increasing the concentrations (5x, 25x and 125x) of either the non-specific ('EMSA_CL_NS_UP' and 'EMSA_CL_NS_LO' oligos annealed; Table 1) or the mutated ('EMSA_CL_Mut_UP' and 'EMSA_CL_Mut_LO' oligos annealed; Table 1) cold probes. It is mention-worthy that the mutated probe had only the 'CGTCCTCG' motif mutated with the other bases remaining unchanged, while all bases for the non-specific probe were randomly changed.

Similar results from competition assay with the sporulating mycelia nuclear extracts (Fig 6B) proved that the band 'b' was specifically due to the binding activity between the 'CGTCCTCG' motif and some sporulating mycelia nuclear protein/proteins.

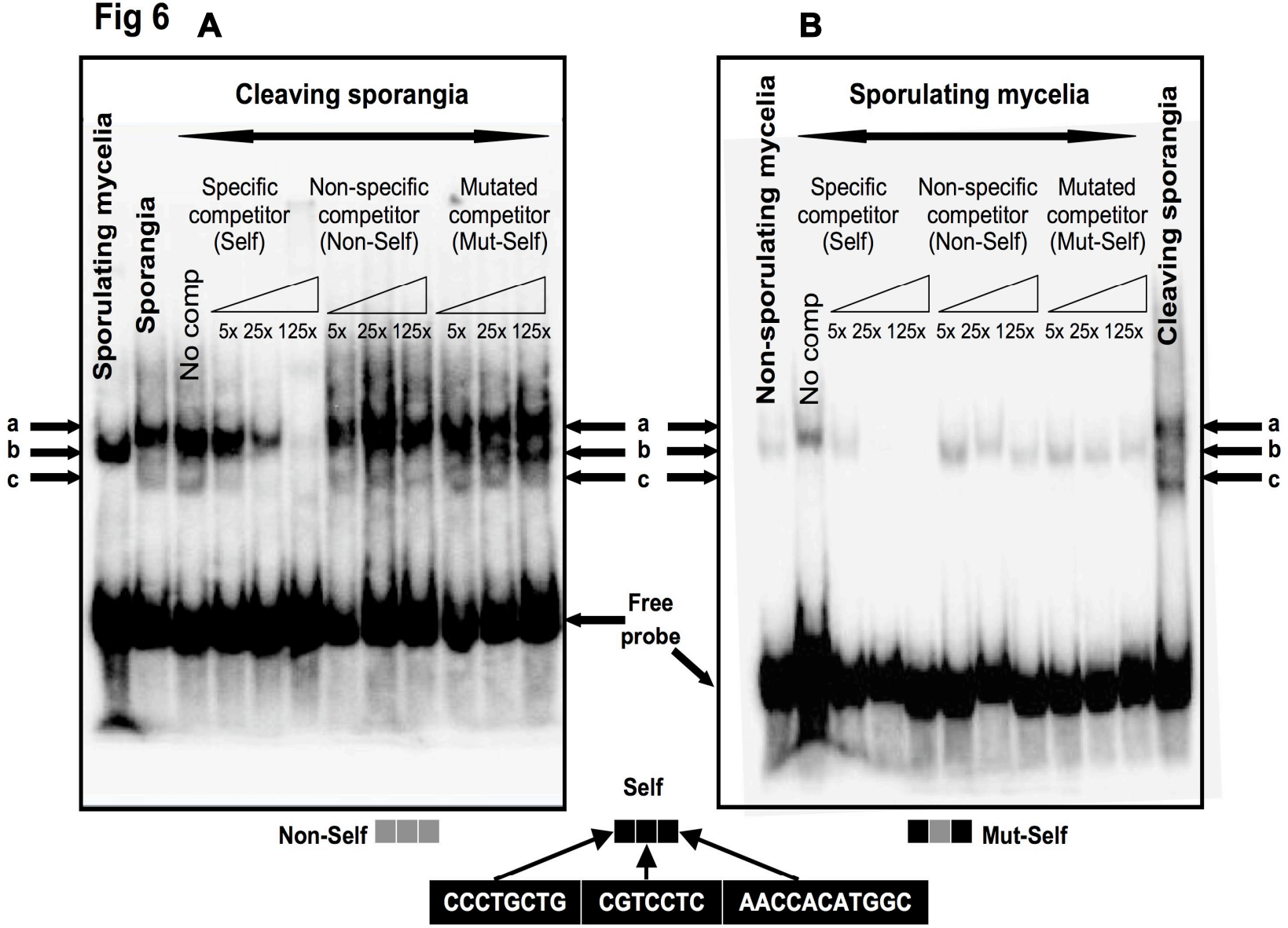
It is worth mentioning that there was an apparent shift in mobility of the proteins in cleavage when compared to that of sporangia. This can be said as the band 'a', which could be seen when the binding reactions were done with nuclear extracts from cleaving sporangia, was not seen for the reaction that had extracts from sporangia. But a different band 'b', at a position slightly lower than band 'a', could be detected.

Fig 6 legend:

EMSA competition assays for motif 'CGTCCTCG' within a 46 bp oligonucleotide, to show specific binding affinities for nuclear proteins:

A) Shown are the binding activities of the 'CGTCCTCG' motif with nuclear extracts from cleaving sporangia. The first three lanes (from left) were loaded with reactions having sporulating mycelia, sporangia and cleaving sporangia nuclear extracts respectively, along with the labeled specific probes. The next three lanes were loaded with reactions having cleaving sporangia nuclear extracts, along with the labeled specific probes and increasing concentrations (5x, 25x and 125x of the labeled probe) of cold specific probe. The three lanes further to the right were loaded with reactions similar to the previous three lanes, only the cold probes used, were non-specific. The last three lanes were loaded with reactions similar to the previous six lanes, but the mutated probe was used as the cold probe. 'a', 'b' and 'c' are the three different sized bands. The cartoons for structures of the specific (self), non-specific (non-self) and mutated (mutself) DNA fragments are shown at the bottom. B) Shows binding activities of the 'CGTCCTCG' motif with nuclear extracts from sporulating mycellia. The first and the last lanes (from left) were loaded with reactions having non-sporulating mycellia and cleaving sporangia nuclear extracts along with labeled specific probes. The nine lanes in the middle were loaded with reactions similar to the nine lanes panel 'A'. These had reactions with both hot and cold probes, the only difference was that the nuclear extracts were from sporulating mycelia tissues.

Fig 6


A
Cleaving sporangia
Sporulating mycelia
Sporangia
Specific competitor (Self)
Non-specific competitor (Non-Self)
Mutated competitor (Mut-Self)
No comp
5x 25x 125x
5x 25x 125x
5x 25x 125x
a
b
c
Free probe
Non-Self
B
Sporulating mycelia
Non-sporulating mycelia
Cleaving sporangia
Specific competitor (Self)
Non-specific competitor (Non-Self)
Mutated competitor (Mut-Self)
No comp
5x 25x 125x
5x 25x 125x
5x 25x 125x
a
b
c
Mut-Self
Self
CCCTGCTG
CGTCCTC
AACCACATGGC

Overrepresentation and positional bias towards selection of a sporangia specific motif:

From a set of 99 promoters of genes >10 fold induced in sporangia, compared to hyphae, 26 overrepresented motifs (detected by at least 2 motif finding programs) were detected. After looking for the common motifs and merging 4 of them, based on the criteria (no more than 2 bases different) used in case of the cleaving sporangia set (all data shown in Chapter II) I had a set of 22 candidate TFBSs. Five of the overrepresented motifs were chosen randomly and looked for their positional bias within the 1 kb promoter regions of these 99 sporangia-specific genes.

The results showed that all of these motifs have biases for one or more regions within their promoters (Table 5). Two of the motifs, 'CTTCAAC' and 'AGC[AG]CAAG' showed bias for two 200 bp regions. 'CTGCAAG' and 'GATCGAG' and 'GTGC[AT]GCA' motifs had a bias for one 200 bp region each. The 'CTTCAAC' motif, which has a bias for a region closest to the ATG (1-200 bases) when compared to the other motifs, was picked for further analyses. It was observed the reverse complement of the 'CTTCAAC' motif' 'GTTGAAG' did not have much of a bias for the first 200 bp. The cleaving sporangia and the set with all *P. infestans* gene promoters were searched, to find out if the 'CTTCAAC' motif and its reverse complement 'GTTGAAG' had any bias for certain positions within those sets. No clear bias could be detected in any of the two above mentioned promoter sets (Fig 7A).

Analysis of evolutionary conservation for the CTTCAAC motif:

Orthologs in *P. sojae* and *P. ramorum* of five genes carrying this motif within the first 200 bp upstream of ATG were identified from the Joint Genome Institute (JGI; http://genome.jgi-psf.org/) database using BLASTP. These orthologus gene promoters were then aligned with the respective *P. infestans* promoters as described previously. The 'CTTCAAC' motif was found to be conserved in two out of five gene promoters, in all three species. The alignment of one such gene, PITG_03886, is shown in Fig 7B.

Table 5:

Distribution of five overrepresented motifs within promoters of genes specific to cleaving sporangia:

The table shows the distribution of the five overrepresented motifs within the sporangia induced gene promoters. The 1 kb regions are divided into five 200 nt windows. The raw frequency of the motifs (in 5' to 3' direction) within the windows is shown along with the total number of hits and the positions for which the motifs show a bias. The regions for positional bias are in bold. The motif 'CTTCAAC', and its bias for the first 200 bps, is shown in bold. This motif was chosen for further analyses.

	MOTIF				
Bases from ATG	AGC[AG]CAAG	CTGCAAG	**CTTCAAC**	GATCGAG	GTGC[AT]GCA
1-200	2	4	**15**	2	2
201-400	**5**	**8**	4	**8**	4
401-600	**8**	3	0	5	**7**
601-800	0	0	2	2	4
801-1000	2	0	**8**	1	0
Total hits	17	18	29	18	16

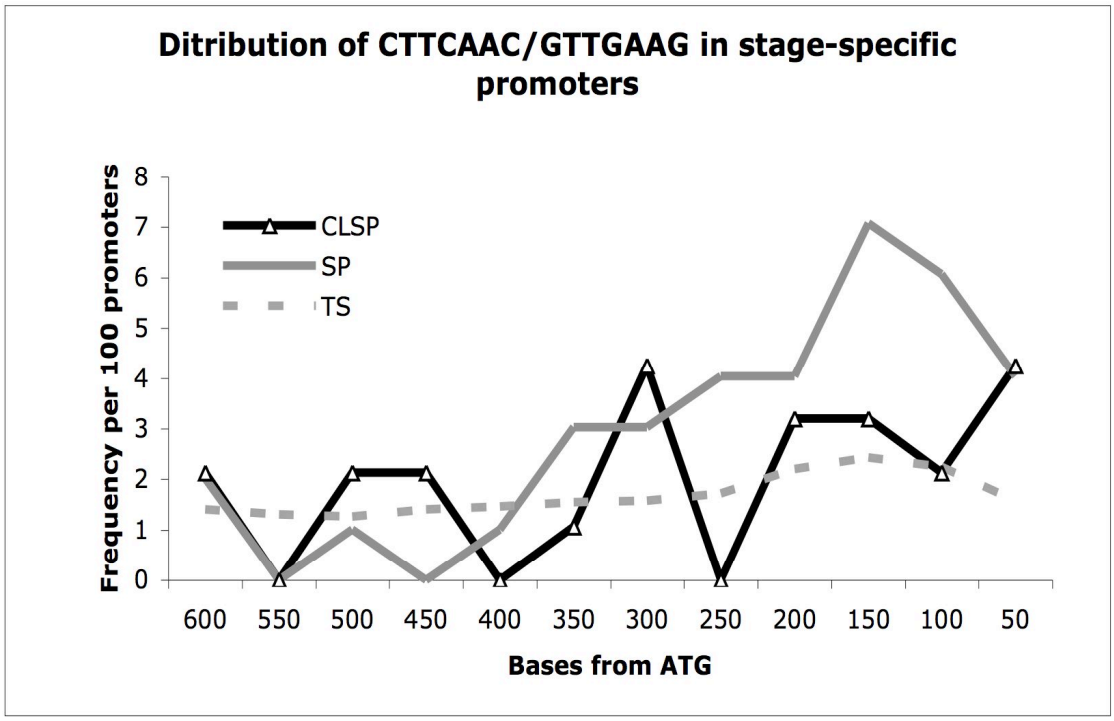
Fig 7 legend:

Positional bias and evolutionary conservation of 'CTTCAAC' motif:

A) The figure shows the positional bias for the first 200 bases upstream of ATG for the motif 'CTTCAAC' within the sporangia-induced gene promoter set. The frequency of 'CTTCAAC' is 3 times, 2 times and 6 times of its reverse complement 'GTTGAAG' in the first three windows upstream of ATG, respectively. Also shown is the distribution of this motif within the cleavage-induced gene promoters and the promoters of all *P. infestans* genes. B) Shown is the evolutionary conservation of the motif 'CTTCAAC' in the promoters of *P. sojae* and *P. ramorum* orthologs of the PITG_03886 gene of *P. infestans* with the help of multiple sequence alignment using ClustalW webtool. The star symbol shows conserved bases and the shaded region shows the conservation for the 'CTTCAAC' motif. 'CLSP', 'SP' and 'TS' refers to cleaving sporangia, sporangia and total set of promoters respectively.

Fig 7

A


Ditribution of CTTCAAC/GTTGAAG in stage-specific promoters
Frequency per 100 promoters
0
1
2
3
4
5
6
7
8
CLSP
SP
TS
600
550
500
450
400
350
300
250
200
150
100
50
Bases from ATG

B

```
P.sojae      CCATAGAGC-----CCCGA---AGCTTTCCACTTCAACCCCGCCA-ATAACATACAACC 50
P.ramorum    TTGGAAAGCAATCAATACAATAGAATCCCGCACTTCAACTTCAGTA-CTAT--TGCGACC 57
P.infestans  ----------------AA------TTTGCACTTCAACTTGTGTAGCCATCTGACGACC 36
                               *          **********     *    *    * ***

P.sojae      AGCATCA-GTTTTTTATAGCCATCGCAGCCTGT-----TTATAGC--CCGTAG---AG 97
P.ramorum    ACCGAAA-ATCC---AAAGATACC-CAGCCAATCA----TCGTACC--CTCTATTTTTAG 106
P.infestans  GCTGAAACGGTCAAAACAGAAAAC--AGCCAATCAGGCACCGTCTCGACCTTATTTT-AG 93
                  *      * **  * *  ****  *          *  *  *  **    **

P.sojae      CCAACTGGAGCCCCTCGCGTTGAAG----------CCATTCGTCGAGAGGAGAC------- 141
P.ramorum    CTCTCTGACGCCCC---CGTTGAAGTGTGTAGCCCCATTC-TCAAGA------------- 149
P.infestans  CTCTTCAACGACCG--GTGGTGAG----------CCATTCGCCGACAAGCAGCTGCTGGA 141
             *      * **    * ***              ******  * * *

P.sojae      GCGGACGGAATG 153
P.ramorum    --------AATG 153
P.infestans  GTTGAAGCGATG 153
                    ***
```

Oligo-chimera assay with 'CTTCAAC' motif:

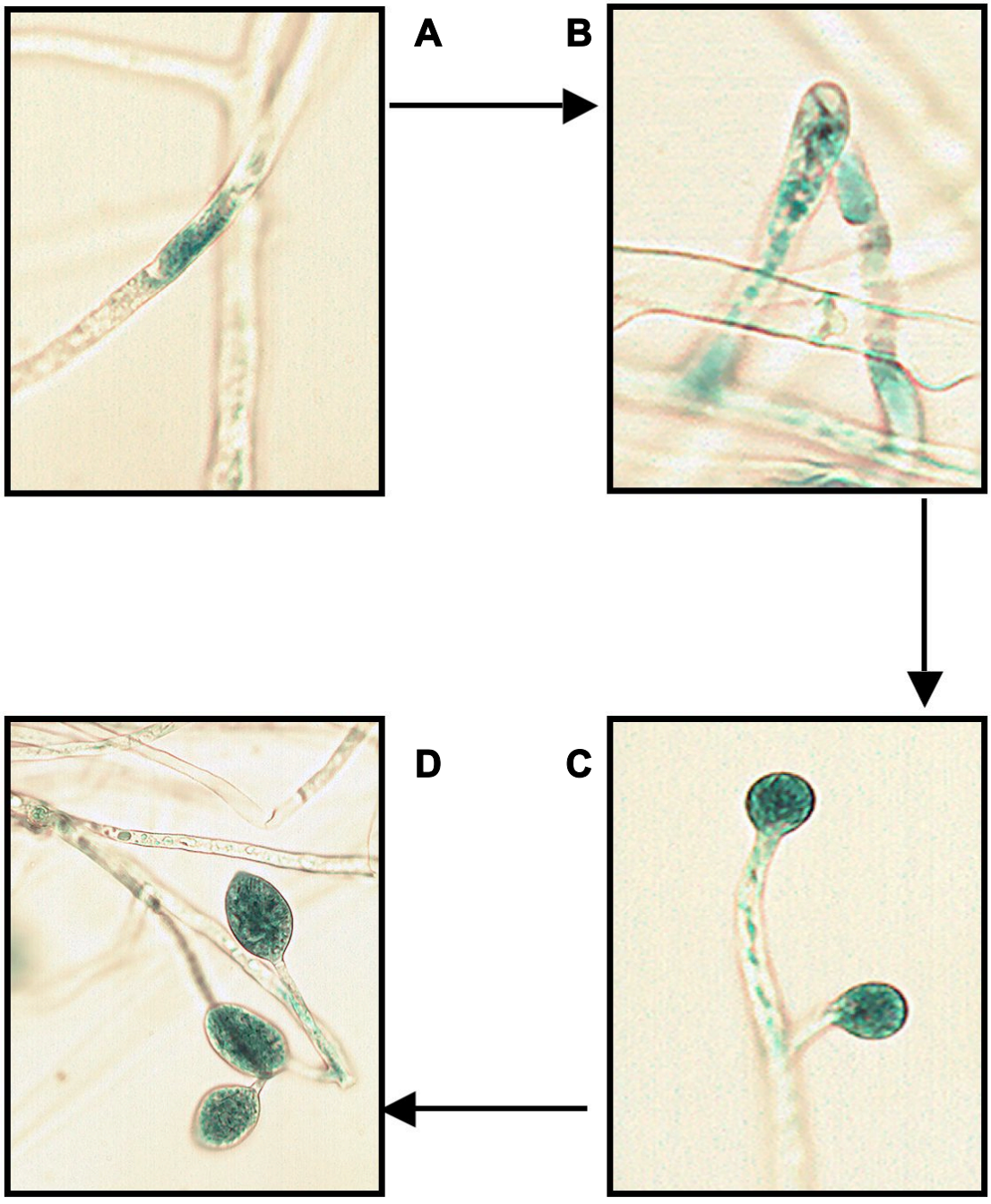
To prove the functionality of the overrepresented, positionally biased and evolutionarily conserved 'CTTCAAC' motif, an oligo-chimera assay was done directly, without deletion or mutation analyses. A 43 bp long double stranded oligonucleotide carrying the 'CTTCAAC' motif was designed by annealing the 'SP2_OC_upper' and 'SP2_OC_lower' oligos. This was then inserted in front of the *NIFS* minimal promoter within the *NIFS*+ Clone38.2 vector. The plasmid DNA was used to transform *P. infestans*. Sporangia and sporulating mycelia from the 70 transformants were incubated with GUS staining solution as described in the 'Materials and methods' section. 10 of these transformants showed GUS staining in sporulating mycelia and sporangia tissues, 5 of which were studied in detail (Table 4). Evidence for GUS expression could be detected as early as the start of sporangiophore development from the sporangiophore initials (Fig 8A). GUS staining could be seen in some sporulating mycelia (Fig 8B), in sporangiophores (Fig 8C) and in early sporangia (Fig 8D). This signal faded in mature sporangia (not shown). 48 hr old non-sporulating mycelia were stained as a control, but no staining could be seen, confirming that this motif is able to drive the expression of a reporter gene with the help of a minimal promoter in sporangia, more specifically in early sporangial development.

Fig 8 legend:

Histochemical staining of *P. infestans* transformed with a plasmid carrying the 'CTTCAAC' motif and the GUS reporter gene:

'CTTAAC' is the motif in front of the *NIFS* minimal promoter driving GUS expression. A) GUS stained sporangiophore initials (stained greenish blue) that gives rise to a sporangiophore. B) GUS stained sporulating mycelia. C) Shown is GUS stained sporangiophore. D) GUS stained sporangia.

Fig 8



Electrophoretic mobility shift assay to test the binding affinity of 'CTTCAAC' motif:

To test the binding affinity of the 'CTTCAAC' motif for nuclear proteins, sets of double stranded oligonucleotides carrying this motif was made with single standed synthetic oligonucleotides (mentioned in Table 1). This was then purified, radio-labeled and incubated with nuclear extracts from sporulating mycelia, non-sporulating mycelia, sporangia and cleaving sporangia tissues. The reactions were run in a 4.5% polyacrylamide gel, dried and put under a phosphorimager screen. A single band 'a' could be detected in reactions with sporangial nuclear extracts. As a result a competition assay was done to check the binding specificity of this motif as described in the 'Materials and methods' section with nuclear extracts from sporangia tissues.

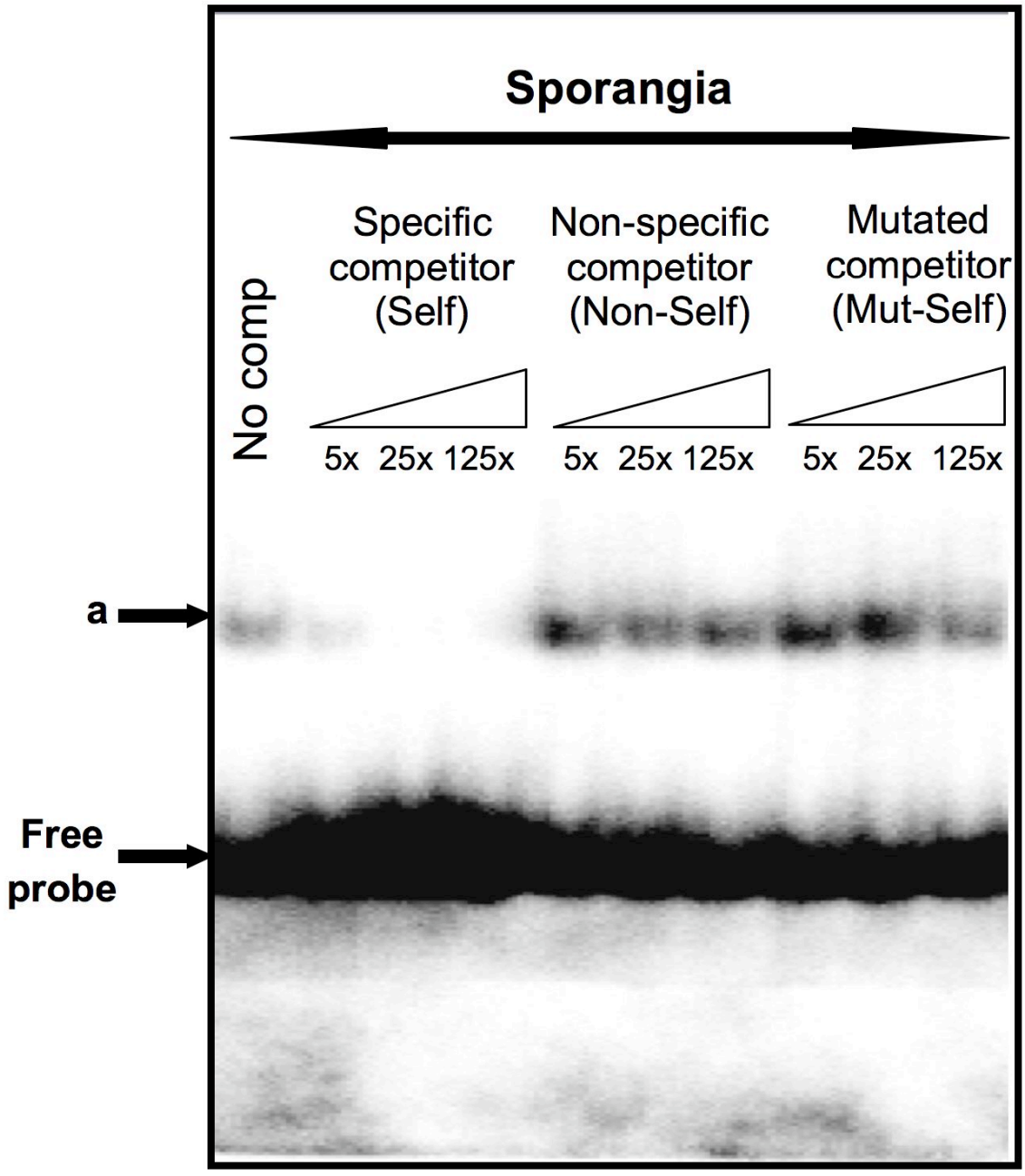
The results showed that the band 'a' was a result of specific binding affinity of the 'CTTCAAC' motif for some nuclear proteins. This was evident as the band faded and finally disappeared with competition from increasing concentrations (5x, 25x and 125x of labeled probe) of unlabeled specific probe (Fig 9). The non-specific and mutated probes did not have any effect on the signal. This proved that the signal was a result of specific to the 'CTTCAAC' motif.

Fig 9 legend:

EMSA competition assays for motif 'CTTCAAC' to show specific binding affinity with nuclear proteins:

Shown is the binding activity of the 'CTTCAAC' motif with nuclear extracts from sporangia. The first lane (from left) was loaded with reactions having sporangial nuclear extracts along with labeled specific probe. The next three lanes were loaded with reactions having nuclear extracts from sporangia, along with the labeled specific probes and increasing concentrations (5x, 25x and 125x of the labeled probe) of cold specific probe. The three lanes further to the right were loaded with reactions having nuclear extracts from sporangia, along with the labeled specific probes and increasing concentrations (5x, 25x and 125x) of cold non-specific probe. The last three lanes were loaded with reactions having nuclear extracts from sporangia, along with the labeled specific probes and increasing concentrations (5x, 25x and 125x) of cold mutated probe.

Fig 9


Sporangia
No comp
Specific competitor (Self)
Non-specific competitor (Non-Self)
Mutated competitor (Mut-Self)
5x 25x 125x
5x 25x 125x
5x 25x 125x
a
Free probe

DISCUSSION:

One of the key issues in understanding regulatory networks is the identification of transcription factor binding sites. Most studies try to solve this problem by either bioinformatics approaches or by molecular biology techniques. But, the reliability of bioinformatics approaches has always been questioned whereas molecular biology approaches remained laborious and time-consuming. High-throughput experimental methods, like ChIP-chip and ChIP-sequencing, are not used widely as these involve high cost and are strongly dependent on cellular type (Vallania et al. 2009). As a result, in recent times, there has been a growing tendency for combining bioinformatics and molecular biology techniques (Vallania et al. 2009). The combined approach seems to be the way forward in solving the puzzle of fast and reliable TFBS prediction. With bioinformatics tools providing fast and reliable predictions, molecular techniques can validate the functionality and specific binding affinity for the candidate TFBSs coming out of the bioinformatics tools.

A key step towards robust TFBS prediction is to search the correct region for these elements and for that one should have a clear idea about where the coding region actually starts. Keeping this in mind, we manually curated all genes before extracting their promoters. It was found that ~15% of the gene models in the Broad Institute database did not have a correct 5' start, when the manual curation was done. Therefore, I believe that doing manual curation was worth the effort.

Identification of overrepresented motifs in a set of co-expressed genes is a well established and much used approach for TFBS predictions but the false discovery rates for this approach is pretty high (Tompa et al., 2005) There are numerous software that are used for identification of overrepresented motifs, but each of these have their own share of pros and cons. Therefore, three different motif finding programs using three different algorithms were picked and then looked for the common motifs. The goal was not to discover the maximum number of overrepresented motifs but to come up with robust predictions for putative TFBSs. To achieve that goal three softwares each of which uses different algorithms were employed and then only those motifs that were detected by at least two of the programs were considered. I believe that by considering only the common motifs detected by two or by all three programs I started with a robust set of overrepresented motifs.

The specific position of a motif within the promter (Tharakamaran et al., 2008) and its evolutionary conservation (Janky and Helden, 2008) has been the focus of attention in the recent years. We found that most overrepresented motifs, that were detected by at least two of the programs, did indeed have a bias for a certain position within the gene promoters. Evolutionary conservation was comparatively more difficult to detect, the most likely reason behind this is probably the incorrect gene models for *P. sojae and P. ramorum*. It is worth mentioning that *P. sojae and P. ramorum* gene models, unlike the *Phytophthora infestans* models, were not curated manually. Therefore, some of these models

might not have their translation start properly annotated. As a result I decided to consider a motif to be evolutionarily conserved only if it was conserved in all three species, in case of two gene promoters.

Multistep functional analyses like serial deletion or mutation are highly labor intensive and time consuming. This is especially true for an organism like *P. infestans*, for which one has to wait ~15 days after transformation and subculture to collect tissue samples for staining. For the PITG_16321 gene carrying the 'CGTCCTCG' motif, even though I did not have to perform promoter bashing blindly, it took several months to complete the analyses. The oligo-chimera assay, where transforming *P. infestans* with just one plasmid can confirm if a motif is functional, is therefore a much faster approach.

It is worth mentioning that in all cases only 15% to 20% of the transformants showed GUS expression. This might be a result of heterokaryosis. But, I also wanted to confirm that the minimal promoter (already shown to be non functional; Ah-Fong et al., 2007) was not driving the expression of the reporter gene, as that would influence the conclusion drawn from the assay. That was the reason behind transforming *P. infestans* with just the minimal promoter by itself. No expression could be detected in sporulating mycelia (by histochemical staining) or in sporangia and chilled (cleaving) sporangia (by histochemical staining and northern blot) as shown in 'Results', which confirmed that the minimal promoter was not functional. *NIFS* is a gene up-regulated in the sporangia, therefore it was important to check if the promoter was able to drive

GUS expression in sporangia tissue samples. It should be mentioned that the staining results from the oligo-chimera assay were comparable to that of the deletion and mutation analysis. In case of the serial deletion and mutation analyses, where unlike the oligo-chimera assays I had the functional PITG_16321 promoter fragments, I could not detect GUS expression in more than 20% of the transformants.

The results from EMSA competition assays for the 'CGTCCTCG' motif suggested that a single TFBS could be responsible for binding different proteins during the course of various life stages. The presence of two different sized bands, with the nuclear extracts from sporulating mycelia and cleaving sporangia also suggests that a single motif might be performing different functional roles by binding to different proteins. It might also be a result of the formation of different complexes due to multiple proteins interacting. The shift in mobility might also be a result of a change in charge of the complex due to phosphorylation. This assay, therefore, apart from showing that a motif has specific binding affinities for certain proteins, also throws light on the potentially complex stage-specific interactions that might be going on during the various life stages.

It was observed that the 'CGTCCTCG' motif that was overrepresented showed a bias for a certain position and was evolutionarily conserved within the promoters of genes up-regulated in cleaving sporangia, was actually functional. Also, it showed binding activities with nuclear proteins from the cleaving sporangia tissue. This led us to hypothesize that the combination of

bioinformatics approaches was adopted for TFBS predictions can give us robust candidate TFBSs. Also, these candidate TFBSs can be tested for their functionality fairly quickly without labor intensive and time consuming deletion or mutation analyses. To confirm this hypothesis the sporangia specific 'CTTCAAC' motif was tested for its function by oligo-chimera assay, and its functionality was confirmed.

The method that was developed and tested in the study combines three different bioinformatics approaches for predicting robust candidate TFBS and then validating the functionality of these candidates by relatively fast experiments like oligo chimera assay and EMSA analyses. I combined gene expression (microarray), regulatory genomics (overrepresentation), positional regulomics (positional bias), comparative genomics (phylogenetic shadowing), functional genomics (oligo-chimera) and protein-DNA binding affinity (EMSA) data for TFBS identification. To the best of my knowledge, this is the only study to date where information from so many different sources have been used for TFBS discovery. The results suggest that this approach on one hand is pretty robust and inexpensive, on the other hand it is not very laborious or time-consuming. Therefore, I believe that with the cost of sequencing going down, and the number of genomes sequenced going up every day, this approach can be applied in case of any organism to great effect.

REFERENCES:

1. Bajic VB, Tan SL, Suzuki Y, Sugano S (2004) Promoter prediction analysis on the whole human genome. Nat. Biotech. 22: 1467-73

2. Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. Ann. Rev. Biochem. 72: 449-79

3. Tompa M, Li N, Bailey TL, Church GM, De Moor B et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23: 137

4. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu. Rev. Genom. Human Gen. 7: 29-59

5. Hochheimer A, Tjian R (2003) Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev. 17: 1309–1320

6. Woychik NA, Hampsey M (2002) The RNA polymerase II machinery: structure illuminates function. Cell 108: 453–463

7. Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Micro. Mol. Biol. Rev. 62: 465–503

8. Butler JEF, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. 16: 2583-2592

9. Struhl K (1987) Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast. Cell 49: 295–297

10. Weis L, Reinberg D (1992) Transcription by RNA polymerase II: Initiator-directed formation of transcription-competent complexes. FASEB J. 6: 3300–3309

11. Smale ST (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim. Biophys. Acta. 1351: 73–88.

12. Smale ST (2001) Core promoters: Active contributors to combinatorial gene regulation. Genes Dev. 15: 2503–2508

13. Smale ST, Jain A, Kaufmann J, Emami KH, Lo K et al. (1998) The initiator element: A paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harb. Symp. Quant. Biol. 58: 21–31

14. Burke TW, Willy PJ, Kutach AK, Butler JEF, Kadonaga JT (1998) The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. Cold Spring Harb. Symp. Quant. Biol. 63: 75–82

15. Weis L, Reinberg D (1997) Accurate positioning of RNA polymerase II on a natural TATA-less promoter is independent of TATA-binding-protein-associated factors and initiator-binding proteins. Mol. Cell. Biol. 7: 2973–984

16. Emami KH, Navarre WW, Smale ST (1995) Core promoter specificities of the Sp1 and VP16 transcriptional activation domains. Mol. Cell. Biol. 15: 5906–5916

17. Martinez E, Chiang CM, Ge H, Roeder RG (1994) TATA-binding protein-associated factor(s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. EMBO J. 13: 3115–3126

18. Crawford DL, Segal JA, Barnett JL (1999) Evolutionary analysis of TATA-less proximal promoter function. Mol. Biol. Evol. 16: 194–207

19. Yean D, Gralla J (1997) Transcription reinitiation rate: a special role for the TATA box. Mol. Cell. Biol. 17: 3809–3816

20. Judelson HS, Michelmore RW (1991) Transient expression of genes in the oomycete Phytophthora infestans using *Bremia lactucae* regulatory sequences. Curr. Gen. 19: 453–459

21. Judelson HS, Tyler BM, Michelmore RW (1992) Regulatory sequences for expressing genes in oomycete fungi. Mol. Gen. Genet. 234: 138–146

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

23. Pesole G, Liuni S, Grillo G, Licciulli F, Larizza A, Makalowski W, Saccone C (2000) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. 28: 193–196

24. Judelson HS, Tani S (2007) Transgene-induced silencing of the zoosporogenesis specific PiNIFC gene cluster of *Phytophthora infestans* involves chromatin alterations. Euk. Cell 6: 1200-1209

25. Ah Fong A, Xiang Q, Judelson HS (2007) Motifs regulating sporulation specific expression and transcription start site preference in the promoter of *Phytophthora infestans* Cdc14 gene. Euk. Cell 6: 2222-30

26. Tani S, Judelson HS (2006) Activation of zoosporogenesis-specific genes in *Phytophthora infestans* involves a 7-nucleotide promoter motif and cold induced membrane rigidity. Euk. Cell 5: 745-752

27. Bulyk M (2003) Computational prediction of transcription-factor binding site locations. Genome Biol. 5: 201

28. Tharakaraman K, Bodenreider O, Landsman D, Spouge JL, Mariño-Ramírez L (2008) The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. Nucleic Acids Res. 36: 2777–2786

29. Das MK, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinfo. 8: S21

30. 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: 1261-6

31. Wakefield MJ, Maxwell P, Huttley GA (2005) Vestige: maximum likelihood phylogenetic footprinting. BMC Bioinfo. 6: 130

32. Hertz GZ, Hartzell GW, Stormo GD (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput. Appl. Biosci. 6: 81-92

33. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262: 208-214

34. Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci. 4: 1618-1632

35. Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3: 21-29

36. van Helden J, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. 281: 827-842

37. Brazma A, Jonassen I, Eidhammer I, Gilbert D (1998) Approaches to the automatic discovery of patterns in biosequences. J. Comput. Biol. 5: 279-305

38. Brazma A., Jonassen I, Vilo J, Ukkonen E (1998) Predicting gene regulatory elements in silico on a genomic scale. Genome Res. 8: 1202-1215

39. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-577

40. van Helden J, Rios AF, Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. 28: 1808-1818.

41. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17: 1113-1122

42. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput. 6: 127-138

44. Ptashne M, Johnson AD, Pabo CO (1982) A genetic switch in a bacterial virus. Sci. Am. 247: 128–140

45. Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H (2001) Combining frequency and positional information to predict transcription factor binding sites. Bioinformatics 17: 1019–1026

46. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C (2004) Clustering of DNA sequences in human promoters. Genome Res. 14: 1562–1574

47. Xie XH, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 30UTRs by comparison of several mammals. Nature 434: 338–345

48. Tharakaraman K, Marino-Ramirez L, Sheetlin S, Landsman D, Spouge JL (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. Bioinformatics 21: 1440–1448

49. Zhang C, Xuan Z, Otto S, Hover JR, McCorkle SR, Mandel G, Zhang MQ (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res. 34: 2238–2246

50. Marino-Ramirez L, Jordan IK, Landsman D (2006) Multiple independent evolutionary solutions to core histone gene regulation. Genome Biol. 7: R122

51.Tompa M, Li N, Bailey TL, Church GM, De Moor B et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23: 137

52. Janky R, van Helden J (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. BMC Bioinfo. 9: 37

53. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. Mol. Biol. Evol. 19: 1114-1121

54. McCue LA, Thompson W, Carmack CS, Lawrence CE (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res. 12: 1523-1532

55. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71-76

56. Guo H, Moose SP (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell 15: 1143-58

57. Hong RL, Hamaguchi L, Busch MA, Weigel D (2003) Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell 15: 1296-1309

58. Bowser PRF, Tobe SS (2007) Comparative genomic analysis of allatostatin-encoding (Ast) genes in Drosophila species and prediction of regulatory elements by phylogenetic footprinting. Peptides 28: 83-93

59. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391–1394

60. Judelson HS, Ah-Fong AM, Aux G, Avrova AO, Bruce C et al. (2008) Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. Mol. Plant-Microbe Inter. 21: 433–447

61. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Sec. Int. Con. Intelli. Sys. Mol. Biol. pp. 28-36

62. Sinha S, Tompa M (2000) A statistical method for finding transcription factor binding sites. Proc. Int. Conf. Intel. Syst. Mol. Biol. 8: 344-54

63. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput. pp: 127-138

64. Thompson JD, Higgins GD, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680

65. Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res.16: 10881-10890

66. Morgenstern B (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res. 32: 33-36

67. Judelson HS, Coffey MD, Arredondo FR, Tyler BM (1993) Transformation of the oomycete pathogen *Phytophthora megasperma* f. sp. *glycinea* occurs by DNA integration into single or multiple chromosomes. Curr. Gen. 23: 211-218

68. Cvitanich C, Judelson HS (2003) A gene expressed during sexual and asexual sporulation in *Phytophthora infestans* is a member of the Puf family of translational regulators. Euk. Cell 2: 465-73.69.

69. Prakob W, Judelson HS (2007) Gene expression during oosporogenesis in heterothallic and homothallic Phytophthora. Fungal Genet. Biol. 44: 726-739

70. Judelson HS, Roberts S (2002) Novel protein kinase induced during sporangial cleavage in the oomycete *Phytophthora infestans.* Euk. Cell 1: 687-695

71. Altschul SF, Madden TL, Schäffer 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: 3389-402

72. Vallania F, Schiavonea D, Dewildea S, Pupoa E, Garbay S (2009) Genome-wide discovery of functional transcription factor binding sites by comparative genomics: The case of Stat3. Proc. Natl. Acad. Sci. USA 106: 5117-22

Chapter II

Identification of putative transcription factor binding sites from five key asexual stages in the Irish potato famine pathogen, *Phytophthora infestans*.

ABSTRACT:

In this chapter I present a systematic study of the proximal promoter elements that might be involved in the regulation of gene expression, during five key asexual stages in *Phytophthora infestans*. I have employed the approach of integrating different bioinformatics tools with molecular techniques, which was proposed and tested in Chapter I, for the identification of these promoter elements. Promoters of genes specific to each of the five stages viz. hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst, were searched for overrepresented motifs. Motifs which were found to be overrepresented within these sets were then subjected to positional bias and evolutionary conservation analyses. These were done to increase the level of confidence that these overrepresented motifs are true transcription factor binding sites. Forty one overrepresented motifs that were not only positionally biased but also showed evolutionary conservation, were detected. It was also checked if any of these putative TFBSs were shared between two or more stages to get an idea about their role in the developmental biology of *P. infestans*. The functionality of three overrepresented motifs, namely 'TACATGTA', overrepresented in promoters of genes expressed all developmental stages, 'GCTGCTG',

overrepresented hyphae and sporangia and ‘TATTAATA’, overrepresented in hyphae and germinating cysts, were checked. I was able to show that the ‘GCTGCTG’ and the ‘TATTAATA’ motifs are able to drive the expression of the reporter gene in sporangia and germinating cysts respectively and demonstrated specific binding affinity in EMSA when incubated with nuclear extracts from the same stages. For the ‘TACATGTA’ motif, which was overrepresented in all stages, no significant staining or binding activity could be detected in any of the five stages. This suggested that the ‘TACATGTA’ motif might not be functional on its own. The hypothesis that ‘TACATGTA’ probably binds to a transcription factor that may not be able to drive the expression of the reporter gene on its own is supported by the presence of the ‘TATTAATA’ motif in the promoters of the bZIP-like transcription factors that are up-regulated in germinating cysts, at a specific distance upstream of the ‘TACATGTA’ motif.

INRODUCTION:

P. infestans has two separate reproductive cycles, sexual and asexual. Here in this chapter, the focus is specifically on five key asexual stages in *P. infestans,* before plant infection. The thread-like vegetative stage called hyphae that gives rise to sporangia. The sporangia stage forms upon the termini of specialized hyphae called sporangiophores, can easily detach (Aylor et al., 2001) and can be transported by wind. These sporangia releases zoospores, (Judelson, 1997) by cleaving (cleaving sporangia stage), under wet and cool

conditions (Schumann and D'Arcy, 2000). During the swimming zoospore stage the organism is biflagellate, can swim to and encyst in the host. The germinating cyst is a stage when the germ tubes (with appressoria) develop from the cysts and invade the host tissue allowing *P. infestans* to draw nutrients from its host (Schumann and D'Arcy, 2000). Various environmental and physiological factors are known to favor sporulation (Ribeiro, 1983) and other *Phytophthora infestans* stages. However, very little is known about the molecular mechanisms that result in the transition from one stage to another. I believe that a better understanding of spore development and the other stages, at the molecular level, might unravel the mechanisms behind these transitions. A key step towards this goal is to find an answer for what causes differential expression of genes during the different asexual stages. Analyses of the promoters of these differentially regulated genes and the TFs responsible for their stage-specific expression can give us that answer, and is therefore essential. Only a few genes have been identified which seem to have importance in the function of sporangia, zoospores or appressoria (Ah-Fong et al., 2007; Latijnhouwers, 2003, 2004; Blanco et al., 2005). There is also evidence that *de novo* transcription is required to complete sporulation (Griffin et al.,1969), direct germination (Clark et al., 1978; Penington et al., 1989) and appressoria formation (Penington et al., 1989; Hardham, 2001; Tyler, 2002; Deacon et al., 1993). A great deal about the biology related to the asexual cycle, at the molecular level, is yet to be known.

I believe that performing studies of promoters of differentially expressed genes, activated during the transition from one stage to another, should help us understand the molecular events occurring during these transitions. To study the gene promoters and identify the elements that most likely bind the transcription factors responsible for the expression of these genes, a method was developed and tested in Chapter I. In brief, three different bioinformatics approaches to find over-represented motifs, were combined, and positional bias and phylogenetic conservation were checked, to robustly predict putative transcription factor binding sites (TFBS). Functional analyses of three motifs were done by histochemical staining of stage-specific tissues and RNA blots. Binding affinity was tested by EMSA.

The goal of this chapter (Chapter II) was not only to detect overrepresented motifs within the proximal promoter regions of genes specific to each of the five developmental stages, but also to come up with robust predictions for putative TFBSs by checking which show positional bias and evolutionary conservation. Testing a few motifs for their functionality within specific stages, was also within the scope of this chapter.

MATERIALS AND METHODS:

Identification of genes and development of promoter data set:

Aprroximately 100 genes from each of the three stages viz. sporangia, cleaving sporangia and germinating cysts and 47 genes from swimming zoospores, that were >10 fold up-regulated (Table 1) when compared to the preceding stages were selected from microarray data (Judelson et al., 2008). For the 100 hyphal genes (Table 1), those that were >10 induced in hyphae when compared to sporangia were selected. For a better understanding of the gene sets we refer the reader to Fig 4 in page 437 of the 'Molecular Plant-Microbe Interaction' paper by Judelson et al. (2008). The right panel of the referred figure shows the expression profiles of the *P. infestans* genes in the five asexual stages. Subsets of the genes induced in during these stages were used for this analysis. Gene sets for hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cysts comprised of subsets of genes shown in panels 'e', 'g', 'i', 'j' and 'b' of the referred figure, respectively.

Models for the genes selected were manually curated, by accessing the Broad Institute database using stand-alone java applet, 'argoc-1'. Promoter regions (1 kb upstream of the coding region) for most of these genes were extracted from the same database with the help of an in-house PERL script; some of the promoters for which the 5' start of the coding region had to be changed were extracted manually.

Table 1:

Shown is the list of genes up-regulated in different asexual stages which were used for this study

#	Hyphae	Sporangia	Germcyst	Swimzoo	#	Hyphae	Sporangia	Germcyst	Swimzoo
1	PITG_00127	PITG_05746	PITG_00339	PITG_14462	**36**	PITG_12158	PITG_01416	PITG_08510	PITG_07866
2	PITG_00287	PITG_16466	PITG_00375	PITG_00201	**37**	PITG_12551	PITG_05447	PITG_08577	PITG_14172
3	PITG_00366	PITG_20763	PITG_00394	PITG_00971	**38**	PITG_12562	PITG_03278	PITG_08892	PITG_14267
4	PITG_00693	PITG_12827	PITG_00698	PITG_06370	**39**	PITG_12808	PITG_20527	PITG_09160	PITG_11594
5	PITG_00754	PITG_02784	PITG_01101	PITG_10035	**40**	PITG_13448	PITG_10184	PITG_09173	PITG_02485
6	PITG_01218	PITG_17002	PITG_02210	PITG_10702	**41**	PITG_14055	PITG_20221	PITG_09218	PITG_10415
7	PITG_01329	PITG_02460	PITG_02332	PITG_12832	**42**	PITG_14083	PITG_20042	PITG_09224	PITG_15909
8	PITG_01397	PITG_03484	PITG_02972	PITG_15838	**43**	PITG_14124	PITG_13895	PITG_09316	PITG_01526
9	PITG_01398	PITG_15867	PITG_03192	PITG_10824	**44**	PITG_14357	PITG_02249	PITG_09355	PITG_13055
10	PITG_01752	PITG_13323	PITG_03415	PITG_09378	**45**	PITG_14436	PITG_09086	PITG_09824	PITG_15640
11	PITG_02047	PITG_12123	PITG_03640	PITG_09528	**46**	PITG_14439	PITG_10162	PITG_09831	PITG_02485
12	PITG_02277	PITG_16089	PITG_04158	PITG_18999	**47**	PITG_14721	PITG_03634	PITG_09995	PITG_15909
13	PITG_03639	PITG_20749	PITG_04314	PITG_17070	**48**	PITG_15398	PITG_09249	PITG_10037	
14	PITG_03719	PITG_14588	PITG_04325	PITG_11231	**49**	PITG_15731	PITG_10352	PITG_10237	
15	PITG_04010	PITG_17247	PITG_04447	PITG_13041	**50**	PITG_16760	PITG_18578	PITG_10722	
16	PITG_04321	PITG_15414	PITG_04949	PITG_00979	**51**	PITG_16866	PITG_09667	PITG_10995	
17	PITG_04717	PITG_06689	PITG_05339	PITG_09124	**52**	PITG_17956	PITG_13461	PITG_11021	
18	PITG_05067	PITG_00121	PITG_05603	PITG_15640	**53**	PITG_18083	PITG_00672	PITG_11278	
19	PITG_05117	PITG_00123	PITG_05881	PITG_08306	**54**	PITG_18299	PITG_00046	PITG_11598	
20	PITG_05616	PITG_12812	PITG_06087	PITG_17040	**55**	PITG_20102	PITG_15875	PITG_11891	
21	PITG_05902	PITG_08419	PITG_06099	PITG_06369	**56**	PITG_20670	PITG_07288	PITG_11944	
22	PITG_06357	PITG_17315	PITG_06202	PITG_02757	**57**	PITG_21127	PITG_15021	PITG_11947	
23	PITG_06478	PITG_02952	PITG_06333	PITG_06438	**58**	PITG_21293	PITG_14635	PITG_12143	
24	PITG_06632	PITG_09287	PITG_06371	PITG_07121	**59**	PITG_21410	PITG_04335	PITG_12361	
25	PITG_07068	PITG_14582	PITG_06402	PITG_04486	**60**	PITG_00833	PITG_09847	PITG_12497	
26	PITG_07077	PITG_14281	PITG_06476	PITG_09954	**61**	PITG_01124	PITG_13011	PITG_12529	
27	PITG_07340	PITG_20800	PITG_06990	PITG_19794	**62**	PITG_02598	PITG_06741	PITG_12872	
28	PITG_07661	PITG_12577	PITG_07100	PITG_01360	**63**	PITG_04570	PITG_00447	PITG_12959	
29	PITG_08599	PITG_08038	PITG_07143	PITG_16978	**64**	PITG_05007	PITG_10127	PITG_13093	
30	PITG_08704	PITG_08750	PITG_07383	PITG_18473	**65**	PITG_06710	PITG_04678	PITG_13196	
31	PITG_10924	PITG_16664	PITG_07387	PITG_03267	**66**	PITG_07069	PITG_05835	PITG_13314	
32	PITG_11294	PITG_02414	PITG_07452	PITG_13283	**67**	PITG_07521	PITG_03886	PITG_13374	
33	PITG_11450	PITG_21504	PITG_07530	PITG_22624	**68**	PITG_09309	PITG_17342	PITG_13507	
34	PITG_11676	PITG_01027	PITG_07725	PITG_03015	**69**	PITG_09798	PITG_10746	PITG_13636	
35	PITG_11883	PITG_05525	PITG_08358	PITG_18452	**70**	PITG_10107	PITG_03018	PITG_13680	

Table 1 contd.

#	Hyphae	Sporangia	Germcyst
71	PITG_10643	PITG_12186	PITG_14003
72	PITG_11573	PITG_09833	PITG_14113
73	PITG_12833	PITG_01317	PITG_14237
74	PITG_13762	PITG_13769	PITG_14238
75	PITG_14583	PITG_05606	PITG_14360
76	PITG_14695	PITG_14087	PITG_14390
77	PITG_14950	PITG_18238	PITG_14685
78	PITG_15067	PITG_15042	PITG_14720
79	PITG_15807	PITG_20518	PITG_15235
80	PITG_16648	PITG_15858	PITG_15239
81	PITG_16926	PITG_16002	PITG_15301
82	PITG_17161	PITG_00937	PITG_15675
83	PITG_17507	PITG_09960	PITG_15745
84	PITG_18312	PITG_12568	PITG_15847
85	PITG_18683	PITG_04454	PITG_15930
86	PITG_21523	PITG_14222	PITG_16084
87	PITG_08002	PITG_03760	PITG_16705
88	PITG_09354	PITG_13183	PITG_16737
89	PITG_09760	PITG_04075	PITG_17063
90	PITG_09793	PITG_04121	PITG_17251
91	PITG_10250	PITG_08625	PITG_17252
92	PITG_13660	PITG_03875	PITG_17546
93	PITG_20139	PITG_02342	PITG_18521
94	PITG_02335	PITG_12236	PITG_19942
95	PITG_04131	PITG_17975	PITG_19957
96	PITG_06454	PITG_12222	PITG_05912
97	PITG_06570	PITG_16262	PITG_06093
98	PITG_07160	PITG_12567	PITG_06433
99	PITG_10674	PITG_16338	PITG_07550
100	PITG_21759		PITG_07761
101			PITG_08193
102			PITG_08879
103			PITG_11671

Detection of overrepresented motifs:

Stand-alone versions of three different motif-finding programs viz. MEME, YMF and BioProspector, were used with background models that were created with 1000 base pairs (bp) upstream of the coding region for each of 100 randomly selected *P. infestans* genes. Also, degeneracy was allowed at two positions by two of these programs.

MEME (Multiple EM for Motif Elicitation; Baily and Elkan, 1994) version 4.3.0, with a minimum width (minw) of 5 and a maximum width (maxw) of 8, was used. The minimum sites for each motif (minsites) used was 5, and the maximum number of iterations was 5 with the rest of the parameters were as described in Chapter I.

YMF (Yeast Motif Finder; Sinha and Tompa, 2000) version 3.0 was used specifying lenRegion (the length of the upstream regions in which motif is to be searched) for each set. The lenOligo i.e. the significant length or the number of non-spacer characters of the motifs to find was specified to be 8. The motifs were sorted by z-score using the sort (-sort) command line parameter.

BioProspector (Liu et al., 2001) Release 2 was used with motif width (-w) specified as 8. The number specified for top motifs to be reported (-r) was 100 and rest of the parameters used were as described in Chapter I. BioProspector was run for 10 times on each set of promoters and a PERL script was used to get rid of the redundant motifs coming out of the ten runs, and generate an output file with the non-redundant motifs only.

A PERL script was used to detect motifs which were detected by at least two out of the three programs. Two motifs coming out of a particular set of co-expressed genes were manually merged, only when those had no more than two different bases. The motifs were originally considered to be different by the programs, as in most cases these had differed only at terminal bases.

Detection of Positional bias:

A PERL script developed in-house was used to get the frequencies of each motif within a 50 base (bp) window that the 1 kilobase (kb) promoters were divided into. The frequencies within each of the five 200 base windows (sum of 4 50 base windows) were used to calculate the positional bias for each motif. The z-score for the equality of proportions was used to calculate the p-value. A p-value cut-off of 0.1 (90% level of confidence) was used. The frequencies of these motifs were calculated by using their 5' to 3' orientation only

Phylogenetic footprinting:

For phylogenetic footprinting, *P. infestans* gene promoters were aligned with the promoters of their orthologs in *P. sojae* and *P. ramorum,* which were extracted manually from the Joint Genome Institute (JGI) database (http://genome.jgi-psf.org/). The promoters were aligned with the help of two alignment programs on the web. The webtools that were used were CLUSTALW (Thompson et al., 1994) and DIALIGN (Morgenstern et al., 2004). For CLUSTALW, the 'gap open' and the 'gap extension penalties' were set at 10 and

0.1 respectively. For DIALIGN the threshold (T) and the regions of maximum similarity used were the default values of zero and five. While checking for conservation, if the motif in *P. sojae* or *P. ramorum* aligned with the one in *P. infestans*, or was within 20 bases of the *P. infestans* motif, a score of '1' was assigned. A two base degeneracy was tolerated in these cases. A motif got a score of '0.5' if it was detected at a site more than 20 bp away (excluding gaps) from the *P. infestans* site, and only one base degeneracy was tolerated in these cases. For most motifs five promoters carrying those, within the established region of positional bias, were looked at.

***P. infestans* strain, culture and manipulations:**

The 1306 strain of *P. infestans* was cultured on rye-sucrose media at 18°C in the dark. *P. infestans* stable transformants were generated by the protoplast method described previously (Judelson et al., 1993). Non-sporulating mycelia were obtained by inoculating clarified rye-sucrose broth with a sporangial suspension, followed by 48 hours of incubation. Cultures on rye-sucrose agar plates that were 9 to 11 days old were used for sporulating mycellia. Sporangia were obtained from 7 to 9 day old sporulating mycelia by adding water, rubbing with a glass rod, and passing the fluid through a 50-µm mesh to remove hyphal fragments. To induce cleavage, sporangia were placed in 100 mm petri plates which were kept on ice for ~60 mins. Germinating cysts were obtained by encystment of motile zoospores (coming out of cleaved sporangia) by adding 0.5

mM calcium chloride, vortexing at medium speed for 1 min and then incubating them at 18°C.

Plasmid construction for oligo-chimera assay and sequencing:

Double stranded oligonucleotides were obtained by annealing two single stranded synthetically designed oligonucleotides (Table 2). These were then cloned into the *NIFS*+Clone38.2 (Chapter I) vector in front of the *NIFS* minimal promoter using the XbaI and XmaI restriction sites. Chemically competent DH5∝ cells were used for bacterial transformation. Clones found to be positive by restriction digestion were sequenced.

Gene expression analysis:

Gene expression analyses were done by histochemical staining for β-glucuronidase (GUS) in tissues from various stages that were stained by the method described by Judelson et al. (1993). The recipe of the staining solution was similar to that described in Chapter I. Tissues were stained at 37°C in the dark, overnight.

Electrophoretic mobility shift assay (EMSA):

Nuclear protein isolation and EMSA were performed as described in Chapter I. In brief, I used the same ingredients like 5 μg of nuclear protein, 1 μg poly dI-dC, 1.6 ng of γ^{32}P-labeled probe and 1 mM DTT, and incubated these with the binding buffer (recipe described in Chapter I) for 15 mins at room temperature followed by 30 min on ice, followed by electrophoresis on a 4.5%

acrylamide gel in 0.5x TBE. The electrophoresis was for 2½ hr at room temperature, unlike 3 hr as used in Chapter I. The gel was dried for an hour, placed under the phosphor screens in autoradiography cassettes and left overnight in the dark. The screen was then analyzed with a phosphorimager. For competition assays, protein was incubated with unlabeled DNA for 15 min and then with the labeled probe for 30 mins in ice. Double-stranded oligonucleotides described in the 'Results and Discussion' section were used as hot probes and cold competitors.

Table 2:

Oligonucleotides used for this study:

Shown below is a list oligonucleotides used for this study. All of the primers/oligonucleotides are in 5'-3' direction. The prefixes used in the names stand for asexual developmental stages; SP and GC are for sporangia and germinating cyst, respectively.

Oligonucleotide name	Oligonucleotides 5'-3'	Used for
SP1_OC_ upper	CTAGAGCTGCTGCGGTTGGTGA TTTCCCGTCTTCTCGTCTACGTC	Oligo-chimera assay with conseved motif GCTGCTG
SP1_OC_ lower	CCGGGGGACGTAGACGAGAAGAG GGAAATTGCACCAACCGCAGCAGT	Oligo-chimera assay with conseved motif GCTGCTG
GC1_OC_ upper	CTAGAGTACATGTAGTTGGTGCAAT TTCCCGTCTTCTCGTCTACGTCCC	Oligo-chimera with conseved motif TACATGTA
GC1_OC_ lower	CCGGGGGACGTAGACGAGAAGAG GGAAATTGCACCAACTACATGTACT	Oligo-chimera with conseved motif TACATGTA
GC2_OC_ upper	CTAGACTATTAATAGTTGGTGCAAT TTCCCGTCTTCTCGTCTACGTCCC	Oligo-chimera with conseved motif TATTAATA
GC2_OC_ lower	CCGGGGGACGTAGACGAGAAGAG GGAAATTGCACCAACTATTAATAGT	Oligo-chimera with conseved motif TATTAATA
EMSA_SP1_ SPEC_UP	AGCGAGTCGATGTCTCCGCCGCTG CTGTGCACGTCACCAACCAGGGA	Specific probe of GCTGCTG motif
EMSA_SP1_ SPEC_LO	TCCCTGGTTGGTGACGTGCACAGC AGCGGCGGAGACATCGACTCGCT	Specific probe of GCTGCTG motif
EMSA_SP1_MUT _UP	AGCGAGTCGATGTCTCCGCCAAGC TATTGCACGTCACCAACCAGGGA	Mutated probe of GCTGCTG motif
EMSA_SP1_MUT _LO	TCCCTGGTTGGTGACGTGCAATAG CTTGGCGGAGACATCGACTCGCT	Mutated probe of GCTGCTG motif
EMSA_GC1_ SPEC_UP	CTGCCTCCTCCAATTTGCATACATG TATTGTGTAGCCATCTGACGACC	Specific probe of TACATGTA motif

Oligonucleotide name	Oligonucleotides 5'-3'	Used for
EMSA_GC2_ SPEC_UP	CTGCCTCCTCCAATTTGCATATTAA TATTGTGTAGCCATCTGACGACC	Specific probe of TATTAATA motif
EMSA_GC2_ SPEC_LO	GGTCGTCAGATGGCTACACAATATT AATATGCAAATTGGAGGAGGCAG	Specific probe of TATTAATA motif
EMSA_GC2_ Mut_UP	CTGCCTCCTCCAATTTGCAACTACG CGTTGTGTAGCCATCTGACGACC	Mutated probe of TATTAATA motif
EMSA_GC2_ Mut_LO	GGTCGTCAGATGGCTACACAACGC GTAGTTGCAAATTGGAGGAGGCAG	Mutated probe of TATTAATA motif

RESULTS AND DISCUSSION:

Motifs overrepresented in the five asexual stages:

Gene sets for each of the five stages were identified using the information on their expression profile from the microarray data, as described in the 'Materials and methods' section. The promoter sets of these genes were assembled and subjected to search for overrepresented motifs by the three programs as described earlier ('Materials and methods'). 125 overrepresented motifs (Table 3A) which were detected by two or more programs were identified. I manually merged 18 of these into other motifs using the criteria of '≤ 2 different bases' described in the 'Materials and methods' section. That left us with 107 overrepresented motifs within the promoter sets of the five different stages. To check how significantly these motifs are overrepresented, the p-value, from the z-score obtained by calculating the equality of proportions between the observed and the expected values of each motif, was calculated. It was observed that 100 out of the 107 motifs were significantly overrepresented at the 90% level of confidence. Out of the 7 motifs that had a p-value of >0.1, 3 motifs ACCGGAA (Table 4), GCGCTC (Table 6) and CC[CT][TG]CACG (Table 7), had a p-value of <0.12. Three motifs, viz. ACGCCGG (Table 4), AGAGACGC (Table 5) and CTTTTG had a p-value of <0.18. One motif, CCGTTG had a p-value of <0.3. Maximum (29) and minimum (13) numbers of overrepresented motifs were detected within the cleavage and swimming zoospore promoter sets, respectively (Table 3A).

Overrepresentation does not guarantee the biological functionality of these motifs; two landmark studies (Tompa et al., 2005; Hu et al. 2005) showed that the prediction accuracy of the motif finding programs are pretty low and the false discovery rates are high. Therefore, even though three motif finding programs were used, only those motifs which were detected by at least two or all of the three programs were selected, I wanted to strengthen the predictions further and increase my confidence in the motifs being biologically functional as TFBSs. As a first step towards increasing the confidence the positional bias of all 107 overrepresented motifs were analyzed.

Detection of positional bias:

To analyze the positional bias a script built in-house was used. The script detected the frequency of each motif within 50 base windows that the 1 kb stage-specific promoter datasets were divided into. The bias was calculated as described in the 'Materials and methods section'. In brief, it was checked if a motif had a bias for any 200 bp region (four 50 base windows) by calculating the p-value (from its z-score) for the number of hits within the said regions. The cut-off used for determination of bias was 0.1(90% level of confidence). The average intergenic region in *P. infestans* is 603 bases (Haas et al., 2009) therefore, I also checked which of the motifs had a bias only for a region that was more than 600 bp upstream of ATG.

It was observed that 68 motifs had a bias for one or more regions and 39 motifs did not show any bias (Tables 3A, 4, 5, 6, 7, 8; 3A is a summary for all

motifs and 4-8 are tables showing motifs overrepresented in the five stages, respectively). Cleavage promoter set carried the maximum number of motifs (12) that did not show any bias (Table 6). Whereas, hyphae and sporangia sets had the maximum number of motifs with positional bias (Tables 4 and 5).

It was also observed that 63 out of 68 motifs, with positional bias, had a bias for at least one region that was less than 600 bases upstream of the translation start site. Unlike in human, there has been no evidence of distal regulators in *P. infestans*, as most functional promoters identified are small, in whatever little data that is available related to the promoters. Also, the average intergenic region in *P. infestans* is 603 bp, therefore, it is highly likely that these promoter elements would work in close co-ordination with the core promoter elements, from a short distance to control gene expression.

Recent studies (Bellora et al., 2007; Tharakaraman et al, 2008 et al.) have also shown that positional bias could be used as an important tool for identification of TFBSs. The premise behind using positional bias is that the TFBSs would show a bias for a particular position within the promoter, as the TFs that bind to these are positionally constrained with respect to the TSS (Tharakamaran et al. 2008). Once I was able to assess the positional bias for all motifs, I went ahead to check if the motifs were evolutionarily conserved.

Analysis of evolutionary conservation by phylogenetic footprinting:

The evolutionary conservation of the motifs was analyzed by phylogenetic footprinting as described in the 'Materials and and methods' section. In brief, for each biased motif, five gene promoters carrying the motif within the region of its bias, were aligned, with their orthologus promoters from *P. sojae* and *P. ramorum*, the only two sequenced *Phytophthora* genomes at the time when I started this analysis. For the motifs within the hyphae, cleavage and sporangia sets that did not show a bias, promoters that carried the motif within the first 600 bases were aligned. Evolutionary conservation was not analyzed for six and seven unbiased motifs within the swimming zoospore and germinating cyst promoter sets, respectively. A scoring system was devised for the analysis ('Materials and methods') where '1' point was awarded if the *P. sojae* and/or *P. ramorum* motif aligned perfectly with the *P. infestans* motif (Fig 1A), or was within 20 bases of the *P. infestans* motif (Fig 1C), a two base degeneracy (Fig 1B) was tolerated in these cases. '0.5' points was awarded if the motif was found in *P. sojae* and/or *P. ramorum* promoters, but at a position more than 20 bases (Fig 1D) away from the *P. infestans,* only one base degeneracy was allowed in these cases. '0' points were awarded for each gene that did not show any conservation (Fig 1E). Tables A,B,C,D,E in the 'Appendices' shows the gene wise score for each of the motifs. These points were then added up and a motif was said to be conserved in one or both species only when the score for one or both species added up to '2' points, i.e. the motif was found to be conserved in at least 2

genes (Tables 4, 5, 6, 7, 8). It should be mentioned that a motif was not considered to be conserved if a total score of '2' was a result of the sum of '1' and two '0.5' points e.g. 'GATGCTG' motif (Table 6, Appendices Table C).

A recent study (Elemento and Tavazoie, 2005) found that motifs can be conserved within orthologus promoters independent of their specific positions within the respective promoters. I wanted to check if such motifs could be detected and therefore searched the orthologus promoters in case no conservation was observed for a motif in its alignment. Some of the motifs were found to be present in *P. sojae* and/or *P. ramorum* promoters at a distance greater than 20 bases away from the sites of the *P. infestans* motif, these were scored (Appendices Tables A, B, C, D, E) but as mentioned previously were not counted when the final conservation score (Tables 4, 5, 6, 7, 8) was determined. I did not count them towards the final score as these being far off from the site of *P. infestans* motif, might be interacting with a different transcription factor and cannot be regarded as a conserved putative TFBS.

The premise behind checking evolutionary conservation is that selective pressure causes functional elements to evolve more slowly than non-functional sequences. Thus, conserved regions within orthologous promoters are candidate TFBSs. But one has to be careful in the selection of species while checking conservation. If the species selected are too closely related TFBSs may not be much different from the bases lacking function. On the other hand it is hard to come up with a good alignment if the species are too distant. I believe that in *P.*

sojae and *P. ramorum* I had two species that were phylogenetically neither too distant nor too close for this analysis. However, it is worth mentioning that promoters evolve at different rates within species.

Five *P. infestans* gene promoters were aligned with their orthologs from *P. soaje* and *P. ramorum* in most cases, as checking one or two genes may not give a clear idea about the conservation. A motif was considered to be conserved only if it showed conservation in two or more (40% or greater) promoters. Twenty-three motifs were found to be conserved in both *P. sojae* and *P. ramorum,* while eleven motifs each were found to be conserved only in *P. sojae* or only in *P. ramorum,* respectively. Multinucleate sporangia cleave to release uninucleate and biflagellate zoospores that infect the plant. During zoosporogenesis, 15% of genes show greater than two-fold induction (Judelson et al., 2008) therefore, one would expect to find more putative TFBSs within the sporangia and cleaving sporangia due to the complexity of these stages. This was true, as we were able to detect 16 putative TFBSs, each for sporangia and cleaving sporangia respectively respectively, that were evolutionarily conserved (Table 3A). One motif 'CTTCAAC' was found to be conserved in hyphae, sporangia and cleavage. Two motifs 'CCGTTG' and 'CTCCTTC' were found to be conserved in *P. sojae* and *P. ramorum,* but in different genes.

After the analyses of overrepresentation, positional bias and evolutionary conservation I can say that 41 of these motifs are highly-likely candidates for being putative TFBSs. Seven of these motifs, as mentioned previously, were not

overrepresented at the 90% level of confidence as mentioned previously. The motif finding programs yielded these as these had more hits than the number of specified minimum sites to be found. I have included these motifs in the analysis as the p-values for most cases were close to 0.1. All of these motifs were found to be show evolutionarily conserved. I believe that the conservation shows that these are elements might have important biological functions related to transcription I must also acknowledge the fact that not all TFBSs are overrepresented.

Fig 1 legend:

Alignments to show the scoring scheme for evolutionary analysis:

Shown is the alignments of motifs for which different scores were assigned. The score of '1' was assigned to 'conservation in both organisms' in case of (A) and (B). Score of '1' was assigned due to 'conservation in *P. sojae*' only in case of (C). '0.5' was assigned to 'conservation in both organisms' in case of (D). (E) got a score of '0'. The motifs are highlighted.

Fig 1


A
P.sojae_131162 344 CTCGTCCTCGCTCGGCCATCGAGGACGCGAGCGAGATAACTCAATCCGGCACTTGCCTGA
P.ramorum_74238 375 CTCGTCCTCGCTCGGCAATCGAGGACG-------GATAACTCAATCCAACCCAAGGCAGG
PITG_18680 403 CGCGTCCTCGTTGAGCTATCGAG--------------AACTCAATCCGACACTCG-----
B
P.sojae_131855 496 AGAAGCGCACCGCCGACTTCGGAGACGCCGTGGAGTTCCTGCTACGCAAGCACGGCAAGA
P.ramorum_76916 529 AGAAGCGCACAGGTGATTTCGGCGATGCAGTGGAGTTCCTGCTGCGCAAGCACGGAAAGA
PITG_09954 493 AGAAACGAACGGCTGATTTTGGTGACGCTGTGGAGTTTCTTCTTCGTAAACACGGCAAGA
C
PITG_10630 465 CCTTATTGATCTCCCTCCACCACTCACCCAAGTGGCTTCACATTTCATTCCACCGGCACG
P.sojae_135657 470 CCGCGTT-----------ACCGGCCACCAAAGTGACTGCA-------TTCCCCCTGCAT-
P.ramorum_79450 464 GAAGAT-------------CCGGTT-CCTGCGCTGCTCCA-------TCTCGCCATCGC-
D
PITG_20590 353 CAAGTCAAGCCCCCTGCTTCATCTTGGTGCAGTA-----CCCAA--CGGCCCTCTTTGCG
P.sojae_140954 346 CAAGGAAAGCTTGCCTCCC-GTC--GGCACATCCTTTCTCCCAAGGCTGCGCGCACTGCG
P.ramorum_72860 359 CAAGCAAACCTTCCGTCACAGAC--AGCACATCGCTTCTCAGCTTCCAGCTTGCT---CG
PITG_20590 406 ACATCG-GGCTCATTCGTCAAAAGCCTTCGCGGCGAGCAAGCACGAC-----CCCCGCTG
P.sojae_140954 403 ACAGCG-AGCTCATTCGGCAACCCACCGCCCAGCGCTCAAGCCATCT-----GTCCGCCC
P.ramorum_72860 417 ACAGCGaGGCTCATTCGCCTATCTGCCTCGACGCGAACAAGTCCGCCGGTCTCCTCCTAG
PITG_20590 465 ACAAGCAAATCCAAGCAAAGTACTCCGGA-
P.sojae_140954 462 TAAGGCCAAAGCAACTTCA--ACTCCGACT
P.ramorum_72860 474 CAAATCCCGAGCAACTTCA--ACGCCGCCA
E
PITG_02972 264 CTGGACAGTGCAATGTTCTCCGCAACTATTAATA------ACTCCATCAAGGCTGCCCTC
P.ramorum_84624 269 TT--ACGGTTCACCCCCATCAAGAGCAGTT------CGTT-CTCCGCCAAGGCTGTCCTC
P.sojae_136264 260 CTGAGCTCTCCGGCTCCTCCTTCACGTATTTTCTTTCGCGGCGACACCAGGGC-GCTATC

Some other features related to the motifs:

I also checked some other features related to the motifs such as how many of these are palindromic sequences, how many were overrepresented in both orientations and how many are overrepresented in more than one stage. It was found that two motifs each in sporangia, cleavage and swimming zoospore to be plaindromic (Table 3A). Three motifs in hyphae and four in germinating cysts were palindromic (Fig 3A). Out of these the 'TACATGTA' motif was found in all stages. Six motifs each in hyphae and cleaving sporangia were found to be overrepresented in both orientations. Seven sporangia motifs and two and four swimming zoospore and germinating cyst motifs, respectively, were found to be overrepresented in both orientations (Fig 3A).

The possible overrepresentation of a motif in multiple stages was checked (Table 3B, 3C) to get an idea about their role in developmental biology. The results were interesting as most of the motifs that were found to be in more than one stage were actually in two sequential stages, e.g. '[AT]GAAGCT' motif which was overrepresented in hyphae, sporangia or 'TATTAATA' in germinating cyst and hyphae, but were either not conserved or conserved in later stage, suggesting that these are present in the preceding stage as these are required at the later stage. One motif 'TACATGTA' was found to be overrepresented in all the stages and I have done functional analyses of this motif (results shown later in the chapter).

Table 3:

A) Shown is a summary of the entire analysis related to the promoter sets of stage-specific genes. The table is divided into five parts, the 'Genes in dataset' part gives information about the genes that were analyzed. The 'Over-represented motif discovery' summarizes the results from all three programs used. The 'Positional bias' section shows the overall results of the positional bias analysis that was conducted for each overrepresented motif. The 'Evolutionary conservation' section gives consolidated data for conservation of motifs in the two other species *P. sojae* (shown as P. soj) and *P. ramorum* (shown as P. ram). The last section 'Other features' shows data related palindromic sequences, motifs overrepresented in both orientations and motifs present in more than one stage. B) Shown is the stage-wise distribution of motifs present in more than one stage. The stages are in bold. C) Shows the list of motifs overrepresented in multiple stages, the stages these are overrepresented in and their conservation data.

Table 3A Summary for motifs in five asexual stages

			Asexual stages		
Genes in dataset:	**Hyphae**	**Sporangia**	**Cleavage**	**Swimzoo**	**Germcyst**
# of genes checked	100	99	94	47	103
# of gene models for which 5' start was changed	8	14	17	3	9
Over-represented motif discovery:					
# of over-rep. motifs detected by MEME	100	100	100	50	100
# of over-rep. motifs detected by YMF	288	343	221	123	286
#of over-rep. motifs detected by BioProspector	80	90	82	52	87
# of common motifs detected by 2 programs	29	26	35	16	19
# of motifs merged manually	2	4	6	3	3
Final # of motifs	27	22	29	13	16
# of motifs over-rep at 90% level of confidence	25	20	27	12	16
Positional bias:					
# of motifs with pos. bias	21	17	17	4	9
# of motifs with pos. bias at a region < 600 bp	17	17	16	4	9
# of motifs with no bias	6	5	12	9	7
Evolutionary conservation:					
# of of motifs checked for evol. consv.	27	22	29	7	9
# of motifs showing evol. consv. in P. soj	3	4	4	0	0
# of motifs showing evol. consv. in P. ram	1	2	5	1	2
# of motifs showing evol. consv. in both	3	10	7	3	0
Other features:					
# of motifs that were palindromes	3	2	2	2	4
# of motifs over-represneted in both orientation	6	7	6	2	4
# of motifs present in more than one stage	6	6	6	2	5

Table 3B Motifs in multiple stages:

Hyphae	Sporangia	Cleaving sporangia	Swimming zoospores	Germinating cyst
T[AT]TTAATA - **GC**	CA[AG]CAACA - **CL**	AAAAAT[GA][AT] - **GC**	A[CG]GAAGA[ACG] - **SP**	TATTAATA[GA] - **HY**
[AT]GAAGCTG - **SP**	CTTC[AG]AC - **HY,CL**	CAACA[GA]CA - **SP**	TACATGTA - **HY, SP, CL, GC**	TTTAAAAA - **CL**
CTTCAAC - **SP, CL**	GAAGC[GT][AG]C - **HY**	CTTCAAC - **HY, SP**		T[AG]CCGG[TC]A - **HY, CL**
TACCGGTA - **CL, GC**	GCTGC[TA]GCA - **HY**	TACATGTA - **HY, SP, ZO, GC**		TACATGTA - **HY, SP, CL, ZO**
TACATGTA - **SP, CL, ZO, GC**	TACATGTA - **HY, CL, ZO, GC**	TACCGGTA - **HY, GC**		AAAAATAT - **CL**
GCTGC[TA]G - **SP**	[CG]AAGAAG - **ZO**	TTTAAAAA - **GC**		

Table 3C Conservation for motifs overrepresented in multiple stages:

Motifs	Overrep in stages	Conservation
TACATGTA	HY, SP, CL, ZO, GC	Not conserved
TACCGGTA	HY, CL, GC	Not conserved
T[AT]TTAATA	HY, GC	Not conserved
[AT]GAAGCTG	HY, SP	Con in spores
GCTGC[TA]G	HY, SP	Con in spores
CTTCAAC	HY, SP, CL	Con in all 3 stages
CA[AG]CAACA	SP, CL	Con in CL
[CG]AAGAAG	SP,ZO	Con in spores
AAAAAT[GA][AT]	CL, GC	Not conserved
TTTAAAAA	CL, GC	Not conserved

Table 4:

Shown is the frequency of the overrepresented motifs within each of the 50 base window that the 1 kilobase hyphal gene promoter set were divided into. The region for its bias, as per the criteria set in the 'Materials and methods' section are shown in bold. 'bps' refers to bases and 'P.s' and 'P.r' stands for *Phytophthora sojae* and *Phytophthora ramorum* respectively. 'Consv' stands for positionally conserved. HY, SP, CL, ZO and GC stands for hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst respectively. 'NA' refers to not applicable and 'NC' denotes not checked. The motifs that are not significantly overrepresented at the 90% level of confidence are denoted by '*****'.

Table 4: Positional bias and evolutionary conservation of over-represented motifs within hyphal gene promoters

			Motifs			
	[AC][AT]GCAGCA	T[AT]TTAATA	[CG]ATTTTG	AAAAAAAT	AAATAAA	AAGTGGT
Over-rep data						
Total hits	32	24	39	16	29	14
p-value	0.0001	0.0001	0.0002	0.0005	0.0001	0.0782
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	2 (0.664)	1 (0.759)	**13 (0.007)**	3 (0.138)	4 (0.224)	3 (0.386)
201-400 bps	**9 (0.015)**	**8 (0.012)**	**11 (0.02)**	2 (0.264)	**7 (0.044)**	1 (0.882)
401-600 bps	4 (0.224)	**6 (0.036)**	**8 (0.085)**	3 (0.138)	**7 (0.044)**	5 (0.13)
601-800 bps	**11 (0.005)**	2 (0.39)	4 (0.539)	3 (0.138)	**6 (0.075)**	2 (0.664)
801-1K bps	**6 (0.075)**	**7 (0.021)**	5 (0.348)	**5 (0.042)**	5 (0.13)	3 (0.386)
Evol. Consv. Data						
# hits in P. s.	4/5	7/7	5/5	5/5	5/5	4/5
# hits in P. r.	4/5	7/7	5/5	5/5	5/5	4/5
Consv. in P. s. only	0/5	0/7	1.5/5	0.5/5	0/5	1.5/5
Consv. in P.r. only	0.5/5	0/7	1.5/5	0/5	0/5	1/5
Consv. in 3 species	1.5/5	0/7	0/5	0/5	1/5	0/5
Also present in	NA	GC	NA	NA	NA	NA

Table 4: Positional bias and evolutionary conservation of over-represented motifs within hyphal gene promoters

			Motifs			
	ACCGGAA	**ACGCCGG**	**ACGGACG**	**AGAAAAA**	**[AT]GAAGCTG**	**CAATCAG**
Over-rep data						
Total hits	13	12	14	16	16	14
p-value	0.1146*	0.1658*	0.0782	0.0353	0.003	0.0782
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	0 (0.269)	1 (0.882)	1 (0.882)	2 (0.664)	3 (0.208)	**6 (0.075)**
201-400 bps	3 (0.386)	1 (0.882)	4 (0.223)	2 (0.664)	**7 (0.021)**	2 (0.664)
401-600 bps	3 (0.386)	4 (0.224)	2 (0.664)	3 (0.386)	0 (0.435)	1 (0.882)
601-800 bps	**6 (0.075)**	1 (0.882)	5 (0.129)	5 (0.13)	3 (0.209)	2 (0.664)
801-1K bps	1 (0.882)	5 (0.13)	2 (0.664)	3 (0.386)	3 (0.209)	3 (0.386)
Evol. Consv. Data						
# hits in P. s.	3/3	4/5	3/4	5/5	5/5	5/5
# hits in P. r.	2/3	3/5	3/4	5/5	5/5	5/5
Consv. in P. s. only	0/3	**1/5**	0/4	0.5/5	1/5	**3.5/5**
Consv. in P.r. only	**1/3**	0/5	0/4	0/5	1/5	1/5
Consv. in 3 species	**1/3**	**1/5**	0/4	1/5	0/5	0/5
Also present in	NA	NA	NA	NA	SP	NA

Table 4: Positional bias and evolutionary conservation of over-represented motifs within hyphal gene promoters

			Motifs			
	CCTCCAGC	**CGACGCC**	**CGCTGGT**	**CTGGAAA**	**CTTCAAC**	**GCA[AG]TGC**
Over-rep data						
Total hits	9	23	16	14	29	37
p-value	0.0212	0.0017	0.0353	0.0782	0.0001	0.0004
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**5 (0.042)**	3 (0.386)	5 (0.13)	1 (0.882)	**8 (0.026)**	6 (0.221)
201-400 bps	0 (0.581)	**6 (0.075)**	3 (0.386)	5 (0.13)	**6 (0.075)**	6 (0.221)
401-600 bps	0 (0.581)	**7 (0.044)**	3 (0.386)	3 (0.386)	5 (0.13)	**12 (0.012)**
601-800 bps	0 (0.581)	2 (0.664)	2 (0.664)	4 (0.224)	4 (0.224)	**8 (0.043)**
801-1K bps	**4 (0.075)**	5 (0.13)	3 (0.386)	1 (0.882)	**6 (0.075)**	5 (0.348)
Evol. Consv. Data						
# hits in P. s.	5/5	5/5	2/2	5/5	3/3	5/5
# hits in P. r.	5/5	4/5	2/2	5/5	3/3	5/5
Consv. in P. s. only	0.5/5	1.5/5	0/2	1/5	0/3	**1.5/5**
Consv. in P.r. only	1/5	1/5	0/2	0.5/5	0/3	0.5/5
Consv. in 3 species	0/5	0/5	**2/2**	1/5	**3/3**	**1.5/5**
Also present in	NA	NA	NA	NA	SP, CL	NA

Table 4: Positional bias and evolutionary conservation of over-represented motifs within hyphal gene promoters

				Motifs		
	GCTGC[TA]GT	GTACTAC	GTTGAAG	T[AG]C[AT]GTAC	TACATGTA	TACCGGTA
Over-rep data						
Total hits	14	24	18	54	67	16
p-value	0.008	0.0011	0.0154	0.0008	0.0001	0.0005
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	4 (0.114)	3 (0.386)	**6 (0.075)**	**15 (0.023)**	**12 (0.001)**	**4 (0.075)**
201-400 bps	**5 (0.064)**	2 (0.664)	1 (0.882)	**13 (0.055)**	**23 (0.0001)**	**4 (0.0745)**
401-600 bps	2 (0.39)	4 (0.223)	1 (0.882)	**14 (0.036)**	**14 (0.0003)**	**4 (0.0745)**
601-800 bps	4 (0.114)	**7 (0.044)**	**7 (0.044)**	8 (0.385)	**12 (0.001)**	3 (0.138)
801-1K bps	0 (0.435)	**8 (0.026)**	3 (0.386)	4 (0.767)	**6 (0.023)**	1 (0.608)
Evol. Consv. Data						
# hits in P. s.	5/5	5/5	3/3	5/5	8/8	5/5
# hits in P. r.	5/5	5/5	3/3	5/5	8/8	5/5
Consv. in P. s. only	1/5	0/5	0/3	1.5/5	0/8	0/5
Consv. in P.r. only	0.5/5	1/5	0/3	0.5/5	1.5/8	1/5
Consv. in 3 species	1/5	0/5	**2/3**	0/5	0/8	0/5
Also present in	SP	NA	NA	NA	SP, CL, ZOO, GC	CL,GC

Table 4: Positional bias and evolutionary conservation of over-represented motifs within hyphal gene promoters

		Motifs	
	TATTTTT	**TGCCCGGA**	**TGCTG[TG]C**
Over-rep data			
Total hits	25	14	45
p-value	0.0007	0.0015	0.0001
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	5 (0.13)	0 (0.581)	6 (0.222)
201-400 bps	2 (0.664)	3 (0.138)	**15 (0.003)**
401-600 bps	4 (0.224)	**6 (0.023)**	**8 (0.043)**
601-800 bps	**7 (0.044)**	**4 (0.075)**	**8 (0.043)**
801-1K bps	**7 (0.044)**	1 (0.608)	**8 (0.043)**
Evol. Consv. Data			
# hits in P. s.	5/5	1/1	5/5
# hits in P. r.	5/5	0/1	5/5
Consv. in P. s. only	2/5	0/0	0/5
Consv. in P.r. only	1/5	0/0	1.5/5
Consv. in 3 species	0/5	0/0	0/5
Also present in	NA	NA	NA

Table 5:

A) Shown is the frequency of the overrepresented motifs within each of the 50 base window that the 1 kilobase sporangia specific gene promoter set were divided into. The region for its bias, as per the criteria set in the 'Materials and methods' section are shown in bold. 'bps' refers to bases and 'P.s' and 'P.r' stands for *Phytophthora sojae* and *Phytophthora ramorum* respectively. 'Consv' stands for conserved. HY, SP, CL, ZO and GC stands for hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst respectively. 'NA' refers to not applicable and 'NC' denotes not checked, The motifs that are not significantly overrepresented at the 90% level of confidence are denoted by '*****' .

Table 5: Positional bias and evolutionary conservation of over-represented motifs within sporangia gene promoters

			Motifs			
	[CG]AAGAAG	**AGAGACGC**	**AGC[AG]CAAG**	**ATGGCTAC**	**[CT]GCTCGAG**	**CTGCAAG**
Over-rep data						
Total hits	67	5	17	7	15	18
p-value	0.0001	0.1715*	0.0018	0.0599	0.0048	0.0147
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**10 (0.031)**	3 (0.138)	2 (0.387)	3 (0.138)	3 (0.207)	4 (0.221)
201-400 bps	**22 (0.0001)**	0 (0.583)	**5 (0.063)**	1 (0.541)	3 (0.207)	**8 (0.025)**
401-600 bps	**18 (0.0006)**	2 (0.263)	**8 (0.012)**	1 (0.541)	**5 (0.063)**	3 (0.382)
601-800 bps	5 (0.343)	0 (0.583)	0 (0.437)	1 (0.541)	2 (0.387)	0 (0.272)
801-1K bps	12 (0.012)	0 (0.583)	2 (0.387)	1 (0.541)	2 (0.387)	3 (0.383)
Evol. Consv. Data						
# hits in P. s.	5/5	5/5	3/5	4/4	6/6	6/6
# hits in P. r.	5/5	4/5	3/5	4/4	6/6	6/6
Consv. in P. s. only	0/5	0/5	0/5	**1/4**	**3/6**	1/6
Consv. in P.r. only	0/5	**2/5**	0/5	0/4	1.5/6	0/6
Consv. in 3 species	**3/5**	1/5	0/5	**1/4**	0/6	0/6
Also present in	ZO	NA	NA	NA	NA	NA

Table 5: Positional bias and evolutionary conservation of over-represented motifs within sporangia gene promoters

			Motifs			
	CA[AG]CAAC	**CCGTTG**	**CTTC[AG]AC**	**CTTCT[TC]C**	**GAA[AG]AGA**	**GAAGC[GT][AG]C**
Over-rep data						
Total hits	43	32	47	42	43	18
p-value	0.0001	0.2961*	0.0001	0.0001	0.0001	0.0147
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**13 (0.007)**	7 (0.529)	**21 (0.0001)**	**8 (0.084)**	**15 (0.003)**	5 (0.128)
201-400 bps	**15 (0.003)**	7 (0.529)	**9 (0.051)**	**8 (0.084)**	**10 (0.031)**	3 (0.383)
401-600 bps	7 (0.135)	5 (0.958)	4 (0.532)	7 (0.135)	7 (0.1353)	1 (0.888)
601-800 bps	4 (0.532)	6 (0.723)	4 (0.532)	**10 (0.031)**	**8 (0.083)**	3 (0.382)
801-1K bps	4 (0.532)	7 (0.529)	**9 (0.051)**	**9 (0.051)**	3 (0.802)	6 (0.074)
Evol. Consv. Data						
# hits in P. s.	6/6	7/7	7/7	3/3	5/5	5/5
# hits in P. r.	6/6	6/7	7/7	3/3	5/5	5/5
Consv. in P. s. only	1/6	**4.5/7**	1/7	**1/3**	0/5	0.5/5
Consv. in P.r. only	1/6	**2.5/7**	3/7	**1/3**	0/5	0.5/5
Consv. in 3 species	0/6	1/7	**3/7**	**1/3**	**3/5**	**2/5**
Also present in	CL	NA	HY, CL	NA	NA	HY

Table 5: Positional bias and evolutionary conservation of over-represented motifs within sporangia gene promoters

			Motifs			
	GATCGAG	**GTCGTTT**	**GTGC[AT]GCA**	**GCTGC[AT]G**	**GTGGGGTG**	**TACATGTA**
Over-rep data						
Total hits	18	18	16	54	9	17
p-value	0.0147	0.0147	0.0029	0.0001	0.0209	0.0003
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	2 (0.659)	**8 (0.025)**	2 (0.387)	**9 (0.051)**	**5 (0.041)**	3 (0.138)
201-400 bps	**8 (0.025)**	5 (0.128)	4 (0.114)	**14 (0.004)**	1 (0.541)	**4 (0.075)**
401-600 bps	5 (0.128)	1 (0.888)	**7 (0.020)**	**13 (0.007)**	2 (0.263)	2 (0.263)
601-800 bps	2 (0.659)	1 (0.888)	4 (0.114)	**11 (0.019)**	0 (0.583)	2 (0.263)
801-1K bps	1 (0.888)	3 (0.3825)	0 (0.437)	7 (0.135)	1 (0.541)	**6 (0.023)**
Evol. Consv. Data						
# hits in P. s.	5/5	6/6	3/3	5/5	4/4	5/6
# hits in P. r.	5/5	6/6	3/3	5/5	4/4	4/6
Consv. in P. s. only	0.5/5	1/6	0/3	0/5	0/4	1/6
Consv. in P.r. only	0/5	0.5/6	0/3	0/5	0/4	0/6
Consv. in 3 species	**3/5**	**2/6**	**2/3**	**2/5**	**2/4**	0/6
Also present in	NA	NA	NA	HY	NA	HY, CL, ZOO, GC

Table 5: Positional bias and evolutionary conservation of over-represented motifs within sporangia gene promoters

		Motifs		
	T[GC]GAGTTT	TG[CG]CTG[CG]C	TT[CT][AT]TTTT	TTGAAGT
Over-rep data				
Total hits	19	14	32	21
p-value	0.0007	0.0755	0.0001	0.004
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**6 (0.036)**	5 (0.128)	**12 (0.003)**	5 (0.128)
201-400 bps	**7 (0.020)**	4 (0.221)	**8 (0.025)**	**6 (0.025)**
401-600 bps	3 (0.207)	3 (0.383)	3 (0.383)	5 (0.128)
601-800 bps	2 (0.387)	1 (0.888)	4 (0.221)	2 (0.659)
801-1K bps	1 (0.755)	1 (0.888)	5 (0.128)	3 (0.383)
Evol. Consv. Data				
# hits in P. s.	5/7	5/5	5/6	7/7
# hits in P. r.	5/7	5/5	5/6	7/7
Consv. in P. s. only	1/7	1/5	0/6	**3.5/7**
Consv. in P.r. only	0/7	0.5/5	0/6	1/7
Consv. in 3 species	0/7	0/5	0/6	1.5/7
Also present in	NA	NA	NA	NA

Table 6:

A) Shown is the frequency of the overrepresented motifs within each of the 50 base window that the 1 kilobase clavage specific gene promoter set were divided into. The region for its bias, as per the criteria set in the 'Materials and methods' section are shown in bold. 'bps' refers to bases and 'P.s' and 'P.r' stands for *Phytophthora sojae* and *Phytophthora ramorum* respectively. 'Consv' stands for conserved. HY, SP, CL, ZO and GC stands for hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst respectively. 'NA' refers to not applicable and 'NC' denotes not checked. The motifs that are not significantly overrepresented at the 90% level of confidence are denoted by '*****'.

Table 6: Positional bias and evolutionary conservation of over-represented motifs within cleavage gene promoters

			Motifs			
	[AT]CCACCA	AAAAAT[GA][AT]	AAGCA[AG]A	[ACG]AAGT[ACG]G	ACT[CT]G[AGC]AG	AGAGAGAG
Over-rep data						
Total hits	26	32	32	81	24	20
p-value	0.0176	0.0001	0.0018	0.0115	0.0071	0.0001
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**12 (0.010)**	3 (0.363)	**11 (0.017)**	9 (0.754)	5 (0.207)	**9 (0.004)**
201-400 bps	2 (0.887)	**7 (0.040)**	1 (0.476)	14 (0.463)	**8 (0.044)**	**8 (0.007)**
401-600 bps	4 (0.497)	**9 (0.014)**	5 (0.317)	19 (0.112)	2 (0.887)	3 (0.1345)
601-800 bps	3 (0.759)	**6 (0.070)**	**9 (0.046)**	**20 (0.081)**	**9 (0.026)**	0 (0.592)
801-1K bps	5 (0.317)	**7 (0.040)**	6 (0.198)	19 (0.112)	0 (0.189)	0 (0.592)
Evol. Consv. Data						
# hits in P. s.	5/5	4/5	4/4	2/2	5/5	2/3
# hits in P. r.	5/5	4/5	4/4	2/2	5/5	3/3
Consv. in P. s. only	2/5	0/5	0/4	0/2	0/5	0/3
Consv. in P.r. only	1/5	0.5/5	0/4	0/2	**1/5**	**1/3**
Consv. in 3 species	**2/5**	1/5	**3/4**	**2/2**	**1/5**	**1/3**
Also present in	NA	GC	NA	NA	NA	NA

Table 6: Positional bias and evolutionary conservation of over-represented motifs within cleavage gene promoters

			Motifs			
	AGCATC	**AGTTG[AC]A**	**[AT]GGAGGAG**	**ATT[CT]TTA**	**CAACA[GA]CA**	**CATC[AC]A[CT]G**
Over-rep data						
Total hits	49	29	14	32	19	16
p-value	0.0021	0.0059	0.0067	0.0018	0.0006	0.0277
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**14 (0.029)**	**9 (0.046)**	3 (0.199)	**10 (0.028)**	**10 (0.004)**	5 (0.207)
201-400 bps	7 (0.478)	2 (0.887)	3 (0.199)	**10 (0.028)**	1 (0.734)	1 (0.660)
401-600 bps	7 (0.478)	6 (0.198)	4 (0.109)	4 (0.497)	4 (0.109)	1 (0.660)
601-800 bps	7 (0.478)	7(0.123)	3 (0.199)	3 (0.759)	0 (0.449)	4 (0.342)
801-1K bps	**13 (0.045)**	5 (0.317)	1 (0.734)	5 (0.317)	4 (0.109)	5 (0.207)
Evol. Consv. Data						
# hits in P. s.	5/5	4/5	5/5	4/4	5/5	4/5
# hits in P. r.	4/5	4/5	5/5	4/4	4/5	4/5
Consv. in P. s. only	0.5/5	2/5	0/5	1/4	**4.5/5**	0/5
Consv. in P.r. only	**1/5**	0/5	**2/5**	1/4	1/5	0/5
Consv. in 3 species	**1/5**	1/5	0/5	**2/4**	0/5	1/5
Also present in	NA	NA	NA	NA	SP	NA

Table 6: Positional bias and evolutionary conservation of over-represented motifs within cleavage gene promoters

			Motifs			
	CGCCACC	CGTCCTCG	CTCCTTC	CTTCAAC	CTTCGAG	CTTTTG
Over-rep data						
Total hits	15	12	16	19	23	33
p-value	0.0419	0.0039	0.0277	0.0077	0.0013	0.179*
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	3 (0.363)	**7 (0.013)**	4 (0.209)	5 (0.120)	0 (0.284)	10 (0.157)
201-400 bps	5 (0.120)	1 (0.530)	4 (0.209)	3 (0.363)	**7 (0.040)**	3 (0.564)
401-600 bps	3 (0.363)	0 (0.592)	4 (0.209)	3 (0.363)	**6 (0.070)**	5 (0.895)
601-800 bps	2 (0.631)	2 (0.257)	1 (0.920)	5 (0.120)	**6 (0.070)**	4 (0.841)
801-1K bps	2 (0.631)	2 (0.257)	3 (0.363)	2 (0.631)	4 (0.287)	11 (0.104)
Evol. Consv. Data						
# hits in P. s.	4/5	4/4	5/5	5/5	5/5	5/5
# hits in P. r.	4/5	4/4	5/5	5/5	4/5	5/5
Consv. in P. s. only	**2/5**	0/4	**1/5**	1/5	1/5	1/5
Consv. in P.r. only	1/5	0/4	**1/5**	1/5	0/5	0.5/5
Consv. in 3 species	0/5	**3/4**	**1/5**	**2/5**	0/5	**2/5**
Also present in	NA	NA	NA	HY, SP	NA	NA

Table 6: Positional bias and evolutionary conservation of over-represented motifs within cleavage gene promoters

			Motifs			
	G[AC]AGCC[AG]	GAGCT[CG]C	GATGCTG	GCGCTC	GCAC[CG]AC	GCTC[AC]AA
Over-rep data						
Total hits	22	33	24	35	22	26
p-value	0.0689	0.0012	0.0008	0.1134*	0.0689	0.0176
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**8 (0.075)**	6 (0.198)	3 (0.363)	7 (0.479)	6 (0.198)	5 (0.317)
201-400 bps	3 (0.759)	5 (0.317)	**12 (0.003)**	4 (0.841)	2 (0.887)	5 (0.317)
401-600 bps	3 (0.759)	**10 (0.028)**	5 (0.120)	6 (0.665)	7 (0.123)	4 (0.497)
601-800 bps	**8 (0.075)**	7 (0.123)	2 (0.631)	11 (0.104)	4 (0.497)	6 (0.198)
801-1K bps	1 (0.476)	5 (0.317)	2 (0.631)	7 (0.479)	3 (0.759)	6 (0.198)
Evol. Consv. Data						
# hits in P. s.	5/5	5/5	5/5	4/5	4/5	5/5
# hits in P. r.	5/5	5/5	4/5	4/5	4/5	5/5
Consv. in P. s. only	1.5/5	0/5	2/5	1.5/5	0.5/5	0/5
Consv. in P.r. only	0.5/5	1/5	0/5	0.5/5	0/5	1/5
Consv. in 3 species	0/5	1/5	0/5	1/5	0/5	0.5/4
Also present in	NA	NA	NA	NA	NA	NA

Table 6: Positional bias and evolutionary conservation of over-represented motifs within cleavage gene promoters

			Motifs		
	TAAATAA	**TACATGTA**	**TACCGGTA**	**TCGTC[GT]TC**	**TTTAAAAA**
Over-rep data					
Total hits	15	19	11	18	13
p-value	0.0419	0.0001	0.0067	0.0009	0.0023
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	1 (0.920)	1 (0.530)	2 (0.257)	6 (0.0003)	2 (0.257)
201-400 bps	4 (0.209)	**7 (0.013)**	2 (0.257)	**5 (0.061)**	3 (0.135)
401-600 bps	4 (0.209)	**4 (0.073)**	1 (0.530)	**5 (0.061)**	3 (0.135)
601-800 bps	5 (0.120)	2 (0.257)	3 (0.135)	1 (0.734)	3 (0.135)
801-1K bps	1 (0.920)	**5 (0.040)**	3 (0.135)	1 (0.734)	2 (0.257)
Evol. Consv. Data					
# hits in P. s.	5/5	4/7	2/4	4/4	5/5
# hits in P. r.	5/5	4/7	2/4	4/4	5/5
Consv. in P. s. only	0/5	0/7	0/4	**2/4**	0/5
Consv. in P.r. only	0/5	0/7	0/4	0/4	0/5
Consv. in 3 species	1/5	0/7	0/4	0/4	1/5
Also present in	NA	HY, SP, ZO, GC	HY, GC	NA	GC

Table 7:

A) Shown is the frequency of the overrepresented motifs within each of the 50 base window that the 1 kilobase promoter set of genes specific to swimming zoospore were divided into. The region for its bias, as per the criteria set in the 'Materials and methods' section are shown in bold. 'bps' refers to bases and 'P.s' and 'P.r' stands for *Phytophthora sojae* and *Phytophthora ramorum* respectively. 'Consv' stands for conserved. HY, SP, CL, ZO and GC stands for hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst respectively. 'NA' refers to not applicable and 'NC' denotes not checked.

Table 7: Positional bias and evolutionary conservation of over-represented motifs within swimming zoospore gene promoters

			Motifs		
	[AC]C[CGT]TCTTC	A[CG]GAAGA[ACG]G	ACCGCGAG	AGAAACGA	AGAGCCTG
Over-rep data					
Total hits	14	8	5	5	6
p-value	0.0236	0.0216	0.0733	0.0733	0.0415
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	3 (0.277)	1 (0.477)	1 (0.423)	0 (0.705)	1 (0.423)
201-400 bps	1 (0.920)	2 (0.231)	0 (0.705)	0 (0.705)	2 (0.205)
401-600 bps	**6 (0.050)**	**3 (0.121)**	1 (0.423)	**4 (0.058)**	1 (0.423)
601-800 bps	0 (0.353)	1 (0.477)	1 (0.423)	1 (0.423)	2 (0.205)
801-1K bps	4 (0.155)	1 (0.477)	2 (0.205)	0 (0.705)	0 (0.705)
Evol. Consv. Data					
# hits in P. s.	3/5	5/6	NC	3/3	4/5
# hits in P. r.	4/5	6/6	NC	3/3	5/5
Consv. in P. s. only	1/5	0/6	NC	0/3	0/5
Consv. in P.r. only	1.5/5	0.5/6	NC	0/3	**2/5**
Consv. in 3 species	0/5	1.5/6	NC	**2/3**	0/5
Also present in	NA	SP	NA	NA	NA

Table 7: Positional bias and evolutionary conservation of over-represented motifs within swimming zoospore gene promoters

			Motifs		
	AGCTGAAG	CC[CT][TG]CAGC	CCAGCACG	CG[GC]TGGAG	CTGTAGCC
Over-rep data					
Total hits	5	8	5	8	5
p-value	0.0733	0.1196 *	0.0733	0.0325	0.0733
Positional bias data	hits (p-value)	hits (p-value)	hits (p-value)	hits (p-value)	hits (p-value)
1-200 bps	0 (0.705)	2 (0.374)	0 (0.705)	0 (0.592)	1 (0.423)
201-400 bps	2 (0.205)	1 (0.734)	1 (0.423)	2 (0.257)	2 (0.205)
401-600 bps	1 (0.423)	1 (0.734)	3 (0.107)	2 (0.257)	1 (0.423)
601-800 bps	2 (0.205)	2 (0.374)	0 (0.705)	2 (0.257)	1 (0.423)
801-1K bps	0 (0.705)	2 (0.374)	1 (0.423)	2 (0.257)	0 (0.705)
Evol. Consv. Data					
# hits in P. s.	NC	5/5	NC	NC	NC
# hits in P. r.	NC	5/5	NC	NC	NC
Consv. in P. s. only	NC	2/5	NC	NC	NC
Consv. in P.r. only	NC	0/5	NC	NC	NC
Consv. in 3 species	NC	**2/5**	NC	NC	NC
Also present in	NA	NA	NA	NA	NA

Table 7: Positional bias and evolutionary conservation of over-represented motifs within swimming zoospore gene promoters

		Motifs	
	GCTGCGGG	**GTCACTGA**	**TACATGTA**
Over-rep data			
Total hits	5	5	13
p-value	0.0733	0.0733	0.0009
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	2 (0.205)	1 (0.423)	2 (0.205)
201-400 bps	0 (0.705)	0 (0.705)	**6 (0.018)**
401-600 bps	0 (0.705)	3 (0.107)	2 (0.205)
601-800 bps	2 (0.205)	0 (0.705)	2 (0.205)
801-1K bps	1 (0.423)	1 (0.423)	1 (0.423)
Evol. Consv. Data			
# hits in P. s.	NC	2/2	5/5
# hits in P. r.	NC	2/2	5/5
Consv. in P. s. only	NC	0/2	0/5
Consv. in P.r. only	NC	0/2	1/5
Consv. in 3 species	NC	**2/2**	0/5
Also present in	NA	NA	HY, SP, CL, GC

Table 8:

A) Shown is the frequency of the overrepresented motifs within each of the 50 base window that the 1 kilobase germinatin cyst specific gene promoter set were divided into. The region for its bias, as per the criteria set in the 'Materials and methods' section are shown in bold. 'bps' refers to bases and 'P.s' and 'P.r' stands for *Phytophthora sojae* and *Phytophthora ramorum* respectively. 'Consv' stands for conserved. HY, SP, CL, ZO and GC stands for hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst respectively. 'NA' refers to not applicable and 'NC' denotes not checked.

Table 8: Positional bias and evolutionary conservation of over-represented motifs within germinating cyst gene promoters

			Motifs			
	[ACG]TAC[ACT]GTA	AAAAATAT	AAAAGAAG	AAAGCTTT	AACGGGGT	ACTCGAGC
Over-rep data						
Total hits	81	11	12	11	7	9
p-value	0.0001	0.0078	0.0046	0.0078	0.0637	0.0223
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**11 (0.028)**	1 (0.550)	3 (0.140)	1 (0.550)	1 (0.550)	0 (0.576)
201-400 bps	**25 (0)**	3 (0.140)	1 (0.550)	3 (0.140)	3 (0.140)	3 (0.140)
401-600 bps	**19 (0.0005)**	2 (0.268)	2 (0.268)	**4 (0.076)**	1 (0.550)	1 (0.550)
601-800 bps	**18 (0.001)**	2 (0.268)	3 (0.140)	0 (0.576)	1 (0.550)	3 (0.140)
801-1K bps	8 (0.117)	3 (0.140)	3 (0.140)	3 (0.140)	1 (0.550)	2 (0.268)
Evol. Consv. Data						
# hits in P. s.	5/5	NC	NC	5/5	NC	NC
# hits in P. r.	4/5	NC	NC	5/5	NC	NC
Consv. in P. s. only	0.5/5	NC	NC	0/5	NC	NC
Consv. in P.r. only	0/5	NC	NC	**2/5**	NC	NC
Consv. in 3 species	0/5	NC	NC	0/5	NC	NC
Also present in	NA	CL	NA	NA	NA	NA

Table 8: Positional bias and evolutionary conservation of over-represented motifs within germinating cyst gene promoters

			Motifs			
	CGCCGAAG	CGTGTTGC	CTCACTTC	TACATGTA	GCAGC[AT][GA]A	GTGTCACA
Over-rep data						
Total hits	9	7	12	55	21	10
p-value	0.0223	0.0637	0.0046	0.0001	0.0048	0.0132
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	0 (0.576)	1 (0.550)	**7 (0.0134)**	**9 (0.004)**	**7 (0.046)**	3 (0.140)
201-400 bps	**4 (0.076)**	1 (0.550)	0 (0.576)	**12 (0.001)**	4 (0.232)	2 (0.268)
401-600 bps	2 (0.268)	0 (0.576)	1 (0.550)	**12 (0.001)**	2 (0.681)	2 (0.268)
601-800 bps	0 (0.576)	2 (0.268)	3 (0.140)	**11 (0.002)**	5 (0.135)	3 (0.140)
801-1K bps	3 (0.140)	3 (0.140)	1 (0.550)	**11 (0.002)**	3 (0.398)	0 (0.576)
Evol. Consv. Data						
# hits in P. s.	4/5	NC	4/5	5/6	5/6	NC
# hits in P. r.	5/5	NC	4/5	5/6	5/6	NC
Consv. in P. s. only	0.5/5	NC	1/5	1.5/6	1.5/6	NC
Consv. in P.r. only	1.5/5	NC	0/5	1/6	**3.5/6**	NC
Consv. in 3 species	0/5	NC	0/5	0/6	0/6	NC
Also present in	NA	NA	NA	HY, SP, CL, ZOO	NA	NA

Table 8: Positional bias and evolutionary conservation of over-represented motifs within germinating cyst gene promoters

			Motifs	
	T[AG]CCGG[TC]A	**TAATATTA**	**TTTAAAAA**	**TATTAATA[GA]**
Over-rep data				
Total hits	41	9	14	17
p-value	0.0001	0.0223	0.0016	0.002
Positional bias data	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**	**hits (p-value)**
1-200 bps	**6 (0.078)**	0 (0.576)	1 (0.550)	1 (0.771)
201-400 bps	**12 (0.003)**	2 (0.268)	**4 (0.076)**	**7 (0.021)**
401-600 bps	**8 (0.027)**	2 (0.268)	2 (0.268)	**6 (0.037)**
601-800 bps	**7 (0.046)**	3 (0.140)	**5 (0.042)**	2 (0.398)
801-1K bps	**8 (0.027)**	2 (0.268)	2 (0.268)	1 (0.771)
Evol. Consv. Data				
# hits in P. s.	5/5	NC	3/5	6/6
# hits in P. r.	5/5	NC	3/5	6/6
Consv. in P. s. only	0.5/5	NC	0/5	0/6
Consv. in P.r. only	0/5	NC	0/5	0/6
Consv. in 3 species	0/5	NC	0/5	1/6
Also present in	HY, CL	NA	CL	HY

Functional analyses of three overrepresented motifs:

Once the overrepresented motifs in all stage-specific promoter sets were identified and analyzed for their positional bias and evolutionary conservation, I decided to test the functionality of some of the motifs. Three motifs were picked, 'TACATGTA', which is overrepresented in all developmental stages, 'GCTGCTG' which is overrepresented hyphae and sporangia and 'TATTAATA', which is overrepresented in hyphae and germinating cysts. There were several reasons for selecting these motifs. 'TACATGTA' was chosen as it was the only motif overrepresented in all stages. 'GCTGCTG' was chosen as this motif, even though present in both hyphae and sporangia- induced genes, showed evolutionary conservation only in the sporangia set. Similarly, 'TATTAATA' was present in both germinating cyst and hyphae sets and showed positional bias in both but some conservation only in germinating cyst genes. This motif is also overrepresented in the promoters of the RXLR effector genes (Morgan and Kamoun, 2007) that are very important for successful pathogenicity.

To prove the functionality of these overrepresented motifs, an oligo-chimera assay was done. Forty-three bp single-stranded oligonucleotides were designed for each motif (Table 2). These were then annealed with their reverse complements to make them double stranded. These were then inserted in front of the *NIFS* minimal promoter within the *NIFS*+Clone38.2 vector (as described in Chapter I) and the plasmid DNA was used to transform *P. infestans*.

For the ‘GCTGCTG’ motif I stained sporangia and sporulating mycelia from 59 transformants of *P. infestans*, with GUS staining solution as described in the ‘Materials and methods’ section. GUS staining could be detected in 11 out of the 59 transformants and was limited to sporulating mycelia and sporangia tissues. Five of these (Appendices Table F) were studied in details. Evidence of GUS expression could be detected mostly in mature spores (Figs 2A, 2D) and also in maturing spores (Figs 2C, 2D). Nonsporulting mycelia (48 h old in liquid culture) that was stained as a control did not show any staining. GUS staining could be seen in sporulating mycelia. This showed us that this motif is capable of driving the expression of the reporter gene in sporangia with the help of the minimal promoter which by itself is incapable of driving GUS expression (shown in Chapter I).

For the ‘TATTAATA’ motif that is overrepresented in germinating cysts and hyphae genes, I analyzed germinating cyst and sporulating hyphae tissues from 6 expressing transformants out of a total of 37 transformants. Tissues were stained at different time points from 30 min to 12 hr after encystment. No staining could be seen in tissues 30 mins or 1 hr after encystment. It was observed that for five transformants (Appendices Table F) staining was visible 2 hr after encystment (Fig 3) in germinating cysts. The staining disappeared 9 hrs after inducing encystment. This suggested that this motif is functional when the cyst is geminating. The reason behind losing the signal after 9 hr, even though the germ tubes were present, is probably because these cysts have stopped germinating

due to loss of energy, as the assay was done in water. The results might be different in a plant, where the organism can get its nutrients from the plant, even after 9 hrs of encystment with the help of appressoria.

For the 'TACATGTA' motif tissues from sporulating mycelia, sporangia and germinating cysts of 52 transformants were stained at different time points from 30 min to 12 hr. (Fig 4) No staining could be detected in sporangia or geminating cysts. Very light staining was visible in 5 day old hyphae of three of the transformants (Appendices Table F). This might suggest that this motif is functional only in presence of some other motif that is tissue-specific, as it binds a transcription factor that needs the help of other factors for gene expression.

Fig 2 legend:

'GCTGCTG' motif can drive GUS expression in sporangia:

Shown is the result of histochemical staining of sporangia in *P. infestans* transformed with a double stranded oligonucleotide carrying the 'GCTGCTG' motif in front of the *NiFS* minimal promoter. GUS staining can be seen in mature sporangia (A and D) and in maturing sporangia (B and C).

Fig 2

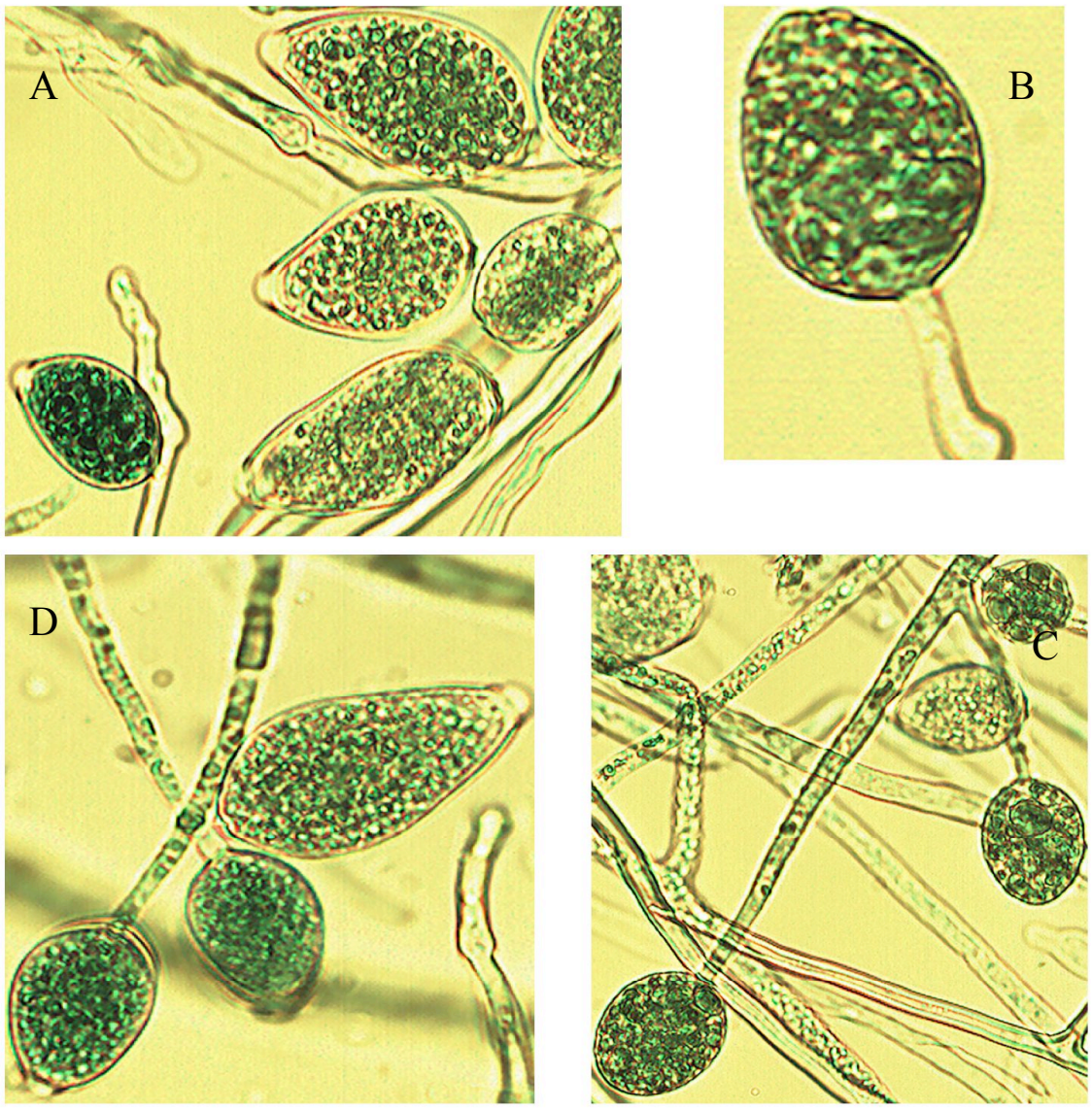


Fig 3 legend:

'TATTAATA' motif can drive GUS expression in germinating cysts:

Shown is the result of histochemical staining of germinating cysts in *P. infestans* transformed with a double-stranded oligonucleotide carrying the 'TATTAATA' motif in front of the *NiFS* minimal promoter. GUS staining can be seen in germinating cysts. The labels (c), (gt) and (a) signifies cyst, germtube and appressoria, respectively.

Fig 3

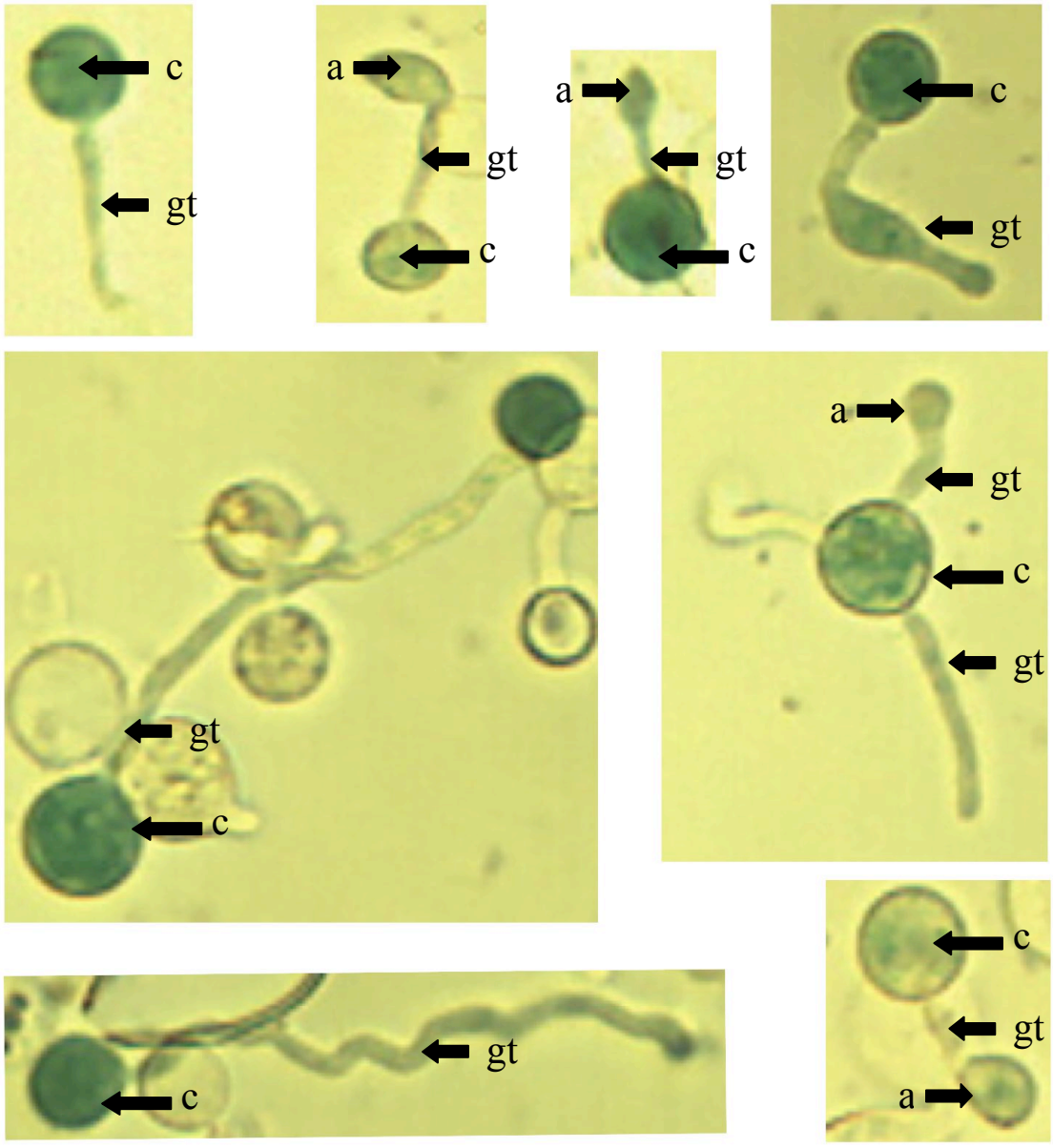
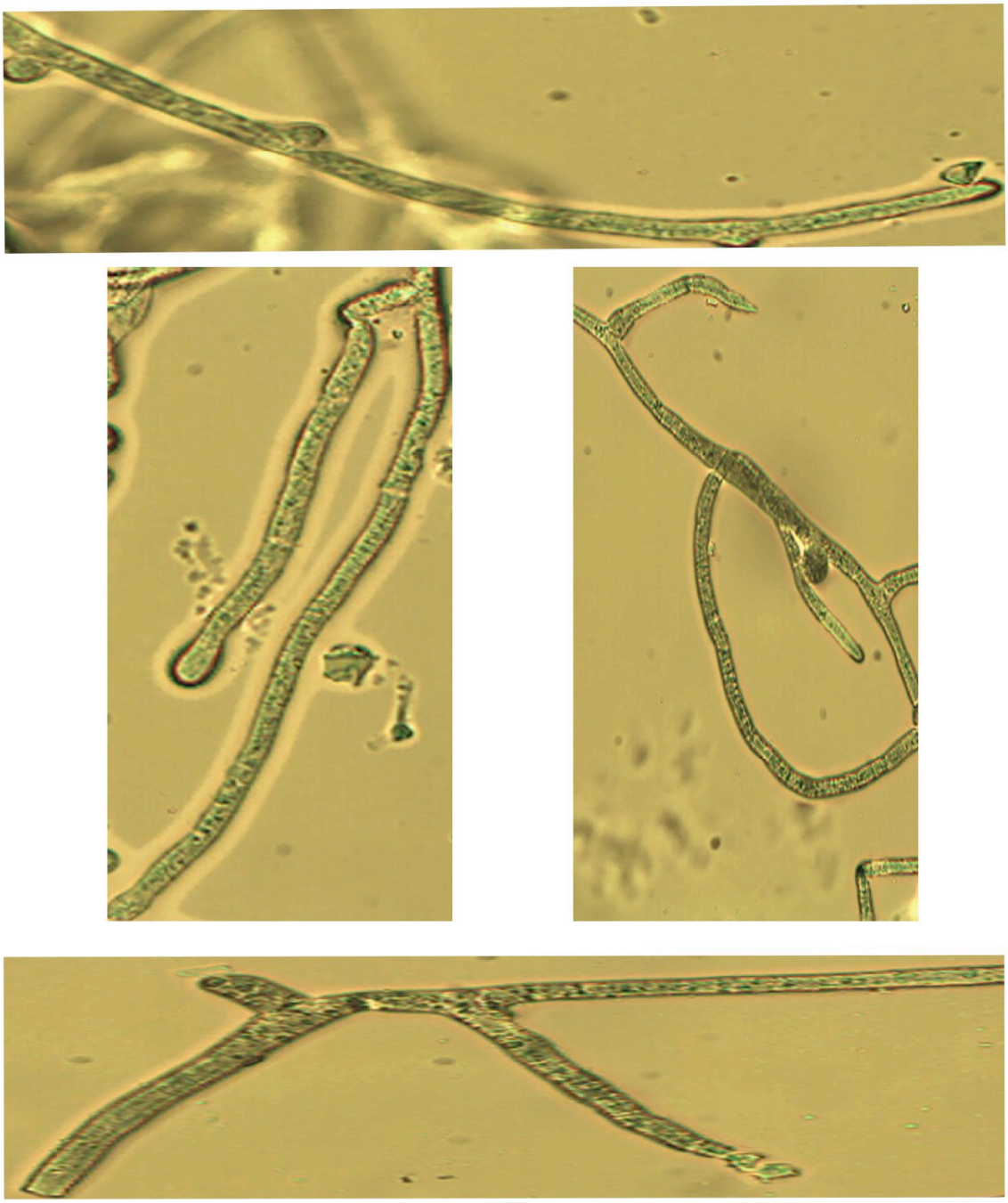


Fig 4 legend:

'TACATGTA' motif can drive GUS expression at very low levels in hyphae:

Shown is the result of histochemical staining of germinating cysts in *P. infestans* transformed with a double stranded oligonucleotide carrying the 'TACATGTA' motif in front of the *NiFS* minimal promoter. Very light GUS staining can be seen in hyphae.

Fig 4



Electrophoretic mobility shift assay the three selected motifs:

To test the binding affinity of the 'TACATGTA' motif for nuclear proteins, sets of double-stranded oligonucleotides carrying this motif were made (mentioned in Table 2). These were then purified, radiolabeled, and then incubated with nuclear extracts from sporulating mycelia, non-sporulating mycelia, sporangia, cleaving sporangia and germinating cyst tissues. The reactions were run in a 4.5% polyacrylamide gel, dried and put under a phosphorimager screen. No band of any significant level (Fig 5) could be detected.

A single band was observed when EMSA was performed with a double stranded oligonucleotide carrying the 'GCTGCTG' motif, that is overrepresented and evolutionarily conserved sporangia. To test the specificity of this band a competition analysis was done, using nuclear extracts from sporangia (Fig 6). The results showed that the band was a due to specific binding affinity of the 'GCTGCT' motif for nuclear proteins. This was evident as the band faded and finally disappeared with increasing concentrations (5x, 25x and 125x of the labeled probe) of unlabeled specific probe. The cold specific probe was able to out-compete the labeled probe but there was no effect on the signals with increasing concentrations of the non-specific and mutated probes. This suggested that the signal was most likely a result of specific binding between the 'GCTGCT' motif and some sporangial nuclear proteins.

I was able to detect binding activity when EMSA assays for the 'TATTAATA' motif was done with nuclear extracts from non-sporulating mycelia, sporulating mycelia, sporangia, and germinating cysts. Two bands were visible when the probe was incubated with nuclear extracts from germinating cyst. The lower band (b) disappeared in competition (Fig 7) with specific cold probe but not with that of non-specific or mutated probes. This suggested that the signal was most likely a result of specific binding between the 'TATTAATA' motif and some germinating cyst nuclear proteins. The upper band (a) was probably due to non-specific binding activity as the band could still be seen with increasing concentration of specific competitor.

Fig 5 legend:

Electrophoretic mobility shift assay with 'TACATGTA' motif:

Results from electrophoretic mobility shift assay with double stranded oligo-nucleotide carrying the 'TACATGTA' motif. NM, MY, SP, CL and GC stands for nuclear extracts from hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cyst respectively, with which the oligonucleotide was incubated.

Fig 5.

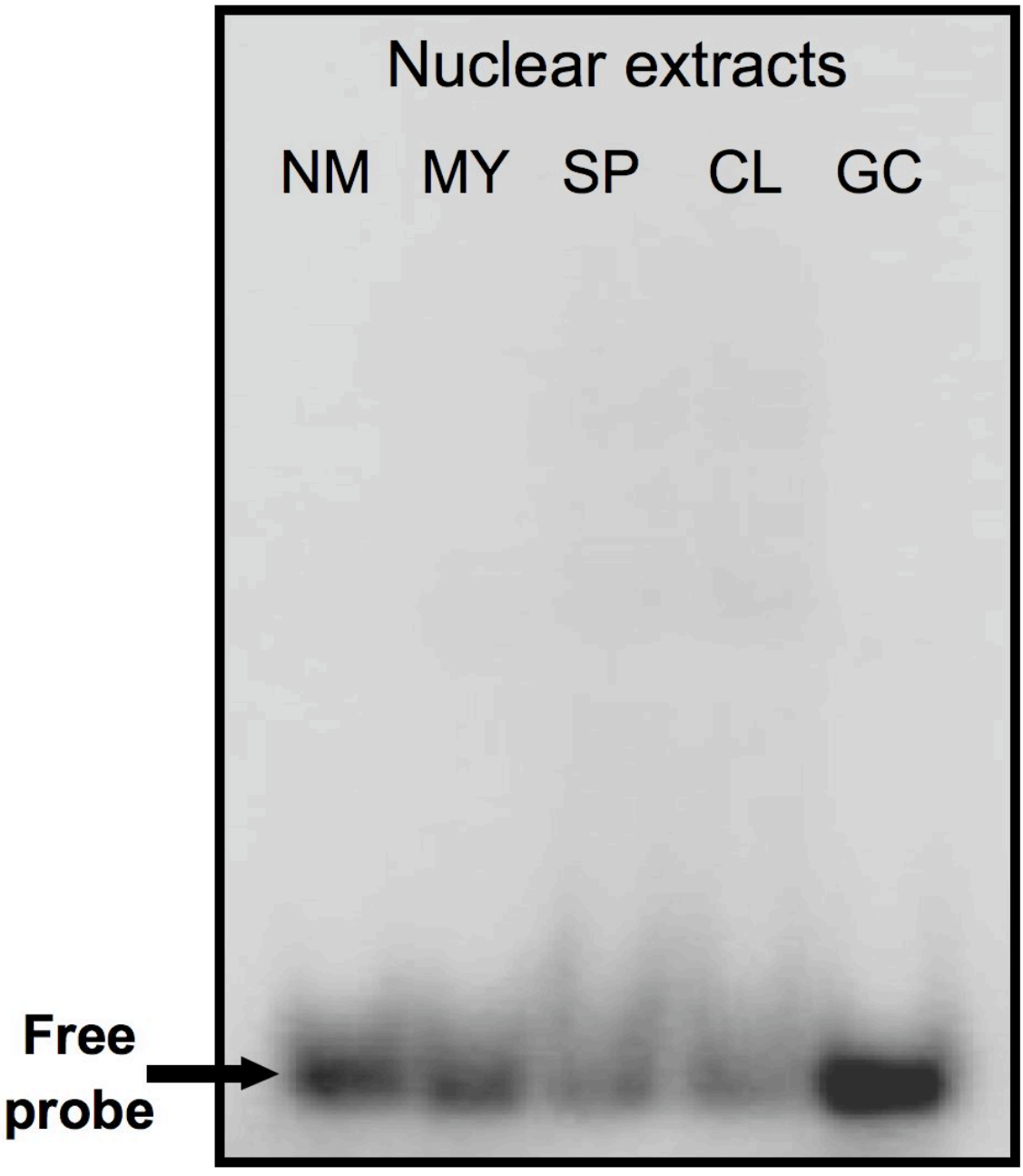


Fig 6 legend:

Electrophoretic mobility shift competition assay with 'GCTGCTG' motif:

Results from electrophoretic mobility shift competition assay with double stranded oligo-nucleotide carrying the 'GCTGCTG' motif. This was incubated with nuclear extracts from sporangia and specific, non-specific and mutated probes. 5x, 25x and 125x signifies the amount of cold competitors added with respect to the hot probe. The specific band is denoted by 'a'.

Fig 6

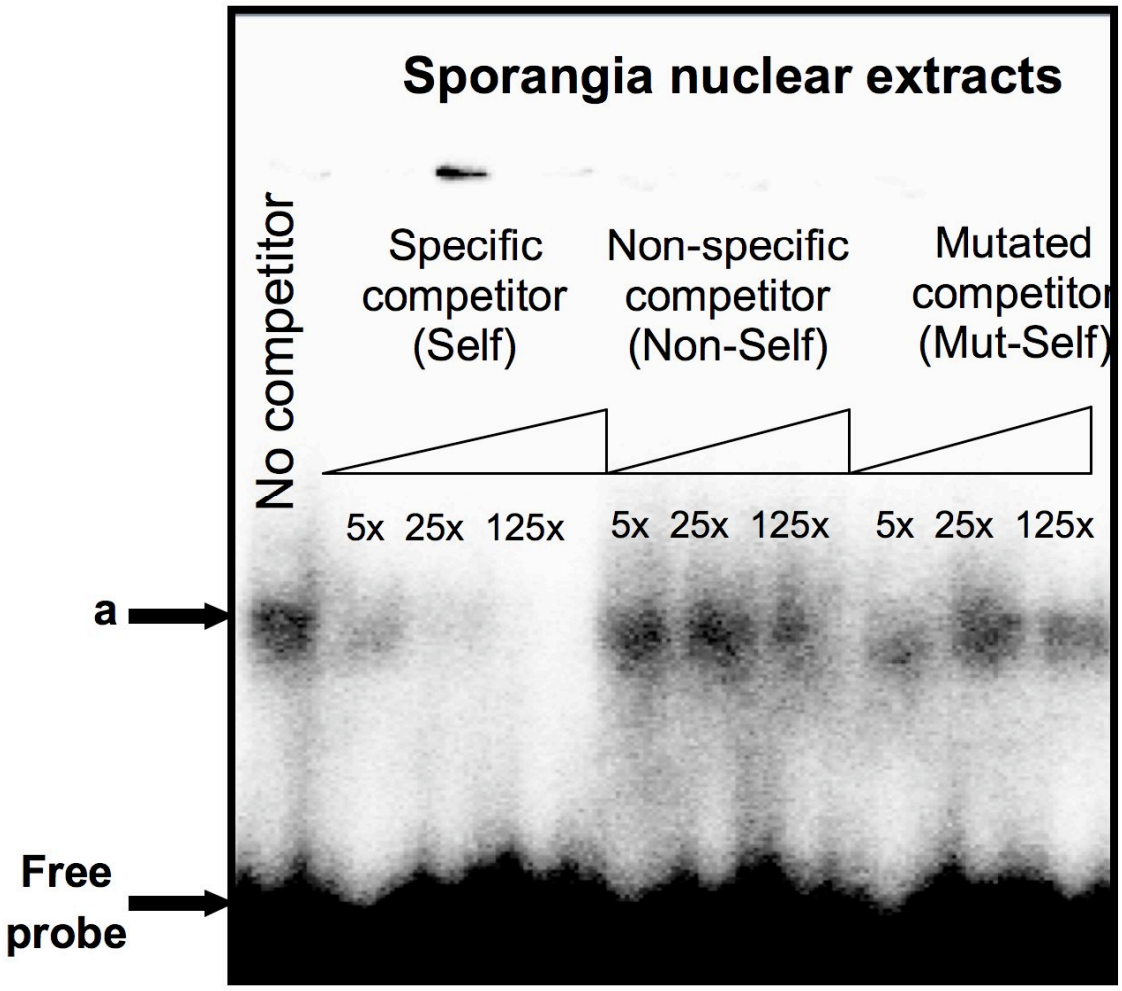
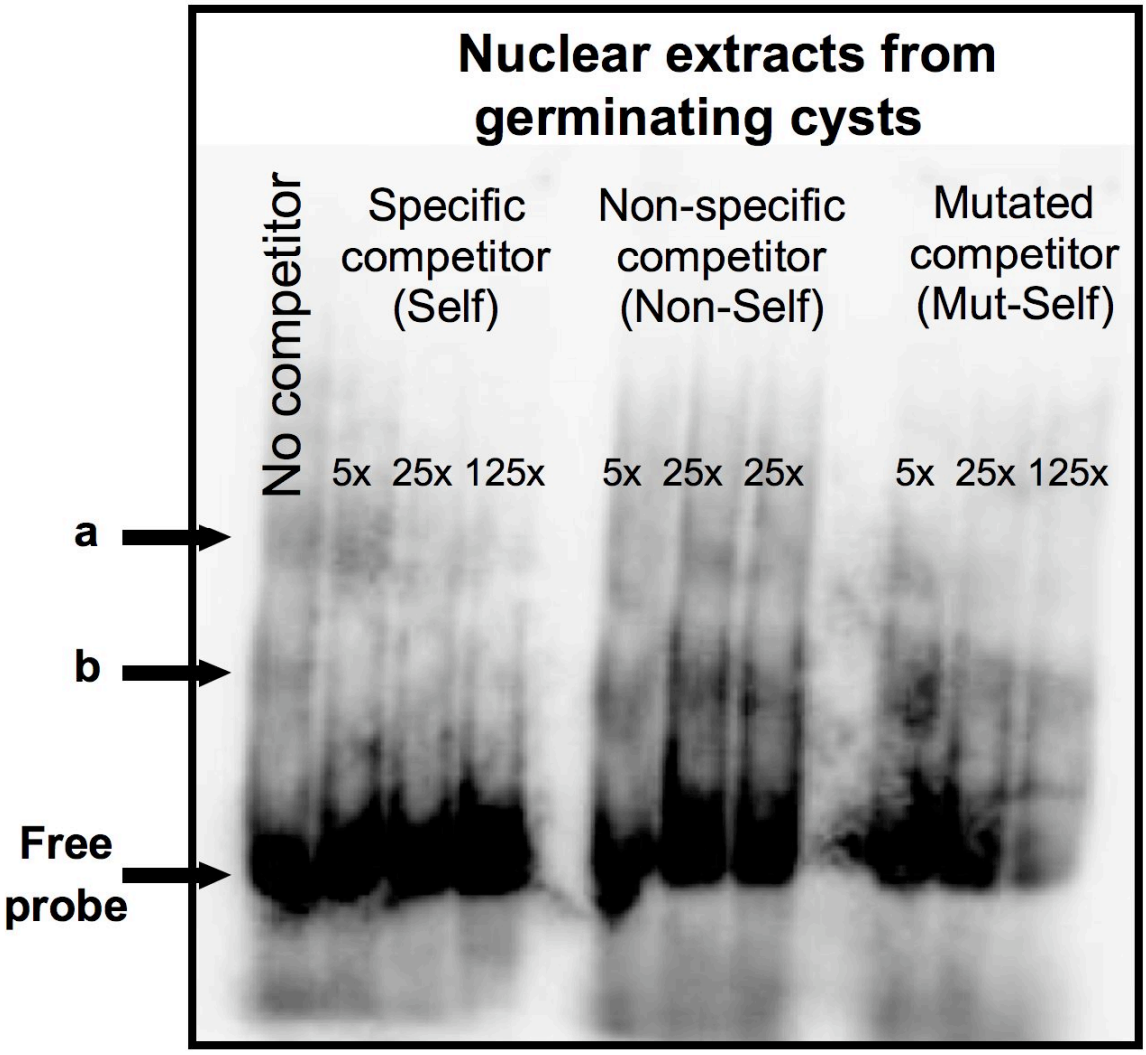

Sporangia nuclear extracts
No competitor
Specific competitor (Self)
Non-specific competitor (Non-Self)
Mutated competitor (Mut-Self)
5x 25x 125x
5x 25x 125x
5x 25x 125x
a
Free probe

Fig 7 legend:

Electrophoretic mobility shift competition assay with 'TATTAATA' motif:

Results from electrophoretic mobility shift competition assay with double stranded oligo-nucleotide carrying the 'TATTAATA' motif. This was incubated with nuclear extracts from germinating cyst and specific, non-specific and mutated probes. 5x, 25x and 125x signify the amount of competitors added with respect to the hot probe. The non-specific band is denoted by 'a' and the specific band is denoted by 'b'.

Fig 7


Nuclear extracts from germinating cysts
No competitor
Specific competitor (Self)
Non-specific competitor (Non-Self)
Mutated competitor (Mut-Self)
5x 25x 125x
5x 25x 25x
5x 25x 125x
a
b
Free probe

A reverse analysis by checking the presence of stage-specific motifs within the promoters of transcription factor genes:

To check if the stage-specific overrepresented motifs that were detected during the course of this analysis, could be found in the promoters of genes up-regulated during the specific stages, a reverse analyses was done. I searched for the known motifs within the promoters of genes. For this analysis I took the promoters of the 18 genes that belong to the bZIP family in *P. infestans*. The expression profiles of 17 of these genes are known (unpublished data Gamboa-Melendez and Judelson). These can broadly be divided into two groups, the nine canonical bZIPs and nine bZIP-like transcription factors. The bZIP like transcription factors show stage-specific expression. Out of the nine bZIP-like transcription factors four (PITG_09198, PITG_09199, PITG_09200 and PITG_09201) show similar expression patterns, these are induced in cysts and germinating cysts. Another gene (PITG_09190) is induced in hyphae along with being induced in cysts and germinating cysts. The 'TATTAATA' motif, which is overrepresented and shows some conservation in germinating cyst was detected in all five promoters, 299 to 763 bases upstream of ATG. In three of the promoters it is present along with the 'TACATGTA' motif (overrepresented in all stages) and the 'TACAGTA' a motif (overrepresented in germinating cyst promoters), approximately 300 bases and 500 bases downstream respectively, of the 'TATTAATA' motif. In the other two genes viz. PITG_09198 and

PITG_09190 the ‘TATTAATA’ motif is present along with ‘GTTGAAG’ (overrepresented in hyphae) and ‘TACCGGTA’ (overrepresented in hyphae, cleavage and germinating cysts) motifs, which are approximately 50 bases downstream of the ‘TATTAATA’ motif. Another bZIP-like transcription factor (PITG_11668) that is induced only in cysts, carried the ‘TACAGTA’ (overrepresented in germinating cysts) motif 215 bases upstream of ATG. Two of the other bZIP like transcription factor genes, PITG_11664 and PITG_13521, show induction during sporangia and both carried sporangia-specific motifs in their promoters, PITG_11664 had ‘CTTC[TC]C’ 151 bases upstream of ATG, whereas, PITG_13521 had ‘C[AG]ACAAC’ motif 256 bases upstream of ATG. PITG_11671 that is induced in germinating cysts had a cleaving sporangia specific motif, ‘AAGC[AG]A’ in its promoter 147 bases upstream of ATG.

Almost all of the nine canonical bZIP encoding genes show high expression in sporangia cleaving sporangia, cysts and germinating cysts. Motifs overrepresented in sporangia and/or cleavage were detected in the promoters of each of these genes. Genes like PITG_09279, PITG_13196, PITG_18417, which showed higher expression in cysts and germinating cysts carried the ‘TACAGTA’, the ‘TACATGTA’ and the ‘TACCGGTA’ motifs in their promoters respectively. The presence of all 41 conserved motifs in the 19 bZIP gene promoters was checked and in most cases the motifs detected were in agreement with their expression profiles i.e the motifs detected were either from the stage where the

gene showed high expression or from a preceding stage. This analysis showed that the motifs detected are indeed highly likely to be real TFBSs.

CONCLUSION:

In this chapter I have done a systematic prediction of putative TFBSs in five key stages of the asexual cycle in one of the most devastating phytopathogens, *P. infestans*. The method described in Chapter I was adopted and each of the five asexual stages was analyzed. I was able to make robust predictions for more than 40 stage-specific motifs that were not only overrepresented, but were also positionally biased and showed evolutionary conservation. It was not surprising that most of these putative TFBSs were within the promoters of genes induced in sporangia and cleaving sporangia stages, as those are the stages when the organism prepares for zoospore release and a lot of genes are induced. Also, most of the motifs that were overrepresented in more than one set were in sets of promoters from consecutive stages.

Another interesting observation was that the number of conserved motifs was not proportional to that of the overrepresented motifs or the number of genes. Four conserved motifs within a set of 47 genes and 13 overrepresented motifs in swimming zoospore stage, were detected, whereas the hyphae set that consisted of 100 genes and 27 overrepresented motifs had only six conserved motifs. This shows that genes within a particular stage like hyphae might be controlled by fewer transcription factors than those in sporangia or cleavage. There are not many genes that are upregulated solely during the swimming

zoospore stage, in the microarray data that was used. This was the reason behind the smaller promoter set for this stage. It was also observed that the motif overrepresented in all the sets did not show much conservation. This motif, 'TACATGTA' when tested for functional activity or binding affinity did not show much staining or any binding activity. I hypothesized that this binds a general transcription factor working in tandem with other stage-specific motifs to drive gene expression. The hypothesis was supported by the presence of this motif in approximately 25% of *P. infestans* genes.

The reverse analysis with the genes encoding for b-ZIP transcripton factors showed that the expression pattern of the genes were consistent with the stage-specificity of the motifs detected in their promoters. The genes with stage-specific motifs showed high expression either in the same or in the following stage. This result was again in congruence with the findings related to overrepresented motifs within multiple stages. In the said analysis too it was found that the motifs overrepresented in more than one stage were actually present in sequential stages.

To conclude, the motifs coming out of this study should help in identification of stage-specific transcription factors and thereby help in having a better idea about the regulatory networks involved in the asexual development of *P. infestans.*

REFERENCES:

1. Schumann GL, D'Arcy CJ (2000) Late blight of potato and tomato. Plant Health Instr. DOI: 10.1094/PHI-I-2000-0724-01
2. Aylor DE, Fry WE, Mayton H, Andrade-Piedra J (2001) Quantifying the rate of release and escape of *Phytophthora infestans* sporangia from a potato canopy. Phytopath. 91: 1189–1196
3. Judelson HS (1997) The genetics and biology of *Phytophthora infestans*: modern approaches to a historical challenge. Fung. Gen. Biol. 22: 65-76
4. Ribeiro OK (1983) Physiology of asexual sporulation and spore germination in Phytophthora. In Phytophthora, its biology, taxonomy, ecology, and pathology (ed. Erwin DC, Bartnicki-Garcia S, Tsao PH) APS Press, St Paul, USA: 55–70
5. Ah-Fong A, Xiang Q, Judelson HS (2007) Motifs regulating sporulation specific expression and transcription start site preference in the promoter of *Phytophthora infestans* Cdc14 gene. Euk. Cell 6: 2222-2230
6. Latijnhouwers M, Govers F (2003) A *Phytophthora infestans* G-Protein β subunit is involved in sporangium formation. Euk. Cell 2: 971-977
7. Latijnhouwers M, Ligterink W, Vleeshouwers VG, van West P, Grovers F (2004) A G-alpha subunit controls zoospore motility and virulence in the potato late blight pathogen *Phytophthora infestans*. Molec. Microbiol. 51: 925-936

8. Blanco FA, Judelson HS (2005) A bZIP transcription factor from Phytophthora interacts with a protein kinase and is required for zoospore motility and plant infection. Molec. Microbiol. 56: 638-648
9. Griffin DH, Breuker C (1969) RNA synthesis during the differentiation of sporangia in the water mold *Achlya*. J. Bact. 98: 689-696
10. Clark MC, Melanson DL, Page OT (1978) Purine metabolism and differential inhibition of spore germination in *Phytophthora infestans*. Can. J. Microbiol. 24: 1032-1038
11. Penington CJ, Iser JR, Grant BR, Gayler KR (1989) Role of RNA and protein synthesis in stimulated germination of zoospores of the pathogenic fungus *Phytophthora palmivora*. Exp. Mycol. 13: 158-168
12. Hardham AR (2001) Cell biology of fungal infection of plants. (In: Howard RJ, Gow NAR (eds)) The Mycota Springer Verl. Heid. pp. 91-123
13. Tyler BM (2002) Molecular basis of recognition between Phytophthora pathogens and their hosts. Ann. Rev. Phytopathol. 40: 137-167
14. Deacon JW, Donaldson SP (1993) Molecular recognition in the homing responses of zoosporic fungi, with special reference to Pythium and Phytophthora. Mycol. Res. 97: 1153-1171
15. Judelson HS, Ah-Fong AMV, Aux G, Avrova AO, Bruce C et al. (2008) Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. Mol. Plant-Microbe Inter. 21: 433–447

16. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Pro. Sec. Int'l. Conf. Intell. Sys. Mol. Biol. pp. 28-36
17. Sinha S, Tompa M (2000) A statistical method for finding transcription factor binding sites. Proc. Int. Conf. Intel. Syst. Mol. Biol. 8: 344-54
18. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput. 6: 127-138
19. Thompson JD, Higgins GD, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680
20. Morgenstern B (2004) DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiServ. Nucleic Acids Res. 32: 33-36
21. Judelson HS, Coffey MD, Arredondo FR, Tyler BM (1993) Transformation of the oomycete pathogen *Phytophthora megasperma* f. sp. *glycinea* occurs by DNA integration into single or multiple chromosomes. Curr. Gen. 23: 211-218
22. Judelson HS, Tani S (2007) Transgene-induced silencing of the zoosporogenesis specific PiNIFC gene cluster of *Phytophthora infestans* involves chromatin alterations. Euk. Cell 6: 1200-1209

23. Tompa M, Li N, Bailey TL, Chruch GM, De Moor B, Eskin E et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23: 137-144

24. Hu J, Li B, Kihara D (2005) Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. 33: 4899-4913

25. Haas PS, Roy NB, Gibbons RJ, Deville MA, Fisher C et al. (2009) The role of X-inactivation in the gender bias of patients with acquired alpha-thalassaemia and myelodysplastic syndrome (ATMDS). Br. J. Haematol. 144: 538-45

26. Bellora N, Farré D, Albà MM (2007) Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Geno. doi: 10.1186/1471-2164-8-459

27. Tharakaraman K, Bodenreider O, Landsman D, Spouge JL, Mariño-Ramírez L (2008) The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. Nucleic Acids Res. 36: 2777–2786

28. Elemento O, Tavazoie S (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. 6: R18

29. Mikula M, Gaj P, Dzwonek K, Rubel T, Karczmarski J et al. (2010) Comprehensive Analysis of the Palindromic Motif TCTCGCGAGA: A Regulatory Element of the HNRNPK Promoter. DNA Res. 17: 245-260

30. Morgan W, Kamoun S (2007) RXLR effectors of plant pathogenic oomycetes. Curr. Opin. Micro. 10: 332–338

APPENDICES:

List A: Stage-specific overrepresented motifs detected by two or more motif finding programs:

Hyphal Motifs:

1. TACATGTA
2. TACCGGTA
3. TAC[AT]GTAC
4. T[AG]CTGTAC
5. TGCCCGGA
6. C[AT]GCAGC
7. AAGCAGCA
8. AAAAAAAT
9. GCTGCAGT
10. TGCTG[TG]C
11. CTTCAAC
12. CAATCAG
13. CGCTGGT
14. CGACGCC
15. TATTTTT
16. AAATAAA
17. ACGGACG
18. AAGTGGT
19. T[AT]TTAATA

20. [AT]GAAGCTG

21. CA[AG]TGC

22. ACGCCGG

23. AGAAAAA

24. ACCGGAA

25. [CG]ATTTTG

26. GTTGAAG

27. GTACTAC

28. CTGGAAA

29. CCTCCAGC

TAC[AT]GTAC and T[AG]CTGTAC merged to T [AG]C[AT]GTAC.
C[AT]GCAGC and AAGCAGCA merged to [AC][AT]GCAGCA.

Sporangia Motifs :

1. [CG]AAGAAG

2. CA[AG]CAAC

3. GCTGC[AT]G

4. CTTC[AG]AC

5. AGC[AG]CAAG

6. GAAGCGAC

7. GAAGCTG

8. CCGTTG

9. CTTCT[TC]C

10. ATGGCTAC

11. AGAGACGC

12. GTGGGGTG

13. T[GC]GAGTTT

14. TGGCTGG
15. CCTGCC

16. TACATGTA

17. GTGC[AT]GCA

18. TTTATTT

19. CTTTTT

20. [CT]GCTCGAG

21. GAAGAGA

22. GAAAAG

23. TTGAAGT

24. GTCGTTT

25. GATCGAG

26. CTGCAAG

GAAGAGA and GAAAAG merged to GAA[AG]AGA
GAAGCTG and GAAGCGAC merged to GAAGC[GT][AG]C
TTTATTT and CTTTTT merged to TT[CT][AT]TTTT
TGGCTGG and CCTGCC merged to TG[CG]CTG[CG]C

Cleavage Motifs:

1. AGAGAGAG
2. TACATGTA
3. TCGTC[GT]TC
4. CGTCGTC
5. TTTAAAAA
6. TACCGGTA
7. CTTCGAG
8. GATGCTG
9. CAACA[GA]CA
10. CTTCAAC
11. AAGCA[AG]A
12. GAGCT[CG]C
13. CGCCACC
14. [AT]GGAGGAG
15. AAAAATAT
16. AAAATGAA
17. TAAATAA
18. [AC]AAGTGGC
19. GAAGT[AC]GA
20. AGTTG[AC]A
21. GCTC[AC]AA

22. ATT[CT]TTA

23. CGTCCTCG

24. AGCATC

25. [AT]CCACCA

26. GAAGCCA

27. GCAGCCG

28. CATCCAT

29. ATC[AC]ACG

30. ACTCG[AG]AG

31. AACTTGC

32. GCAC[CG]AC

33. CTTTTG

34. GCGCTC

35. CTCCTTC

AAAAATAT and AAAATGAA merged to AAAAAT[GA][AT]
[AC]AAGTGGC and GAAGT[AC]GA merged to [ACG]AAGT[ACG]G
ACTCG[AG]AG and AACTTGC merged to ACT[CT]G[AGC]AG
CATCCAT and ATC[AG]ACG merged CATC[ACG]A[CT]G
GAAGCCA and GCAGCCG merged to G[AC]AGCC[AG]
TCGTC[GT]TC and CGTCGTC merged to TCGTC[GT]TC

Swimming Zoospore Motifs:

1. TACATGTA

2. GAAGAAG

3. A[CG]GAAGA[AC]G

4. CCTTCTTC

5. [AC]C[GT]TCTTC

6. CC[CT]TCAGC

7. CCGCAGC

8. GCTGCGGG

9. AGAGCCTG

10. ACCGCGAG

11. AGCTGAAG

12. AGAAACGA

13. CTGTAGCC

14. CG[GC]TGGAG

15. GTCACTGA

16. CCAGCACG

CCTTCTTC and [AC]C[GT]TCTTC merged to [AC]C[CGT]TCTTC
GAAGAAG and A[CG]GAAGA[AC]G merged to A[CG]GAAGA[ACG]G
CC[CT]TCAGC and CCGCAGC merged to CC[CT][TG]CAGC

Germinating Cyst Motifs:

1. ATACTGTA

2. CTACTGTA

3. GTAC[ACT]GTA

4. TACATGTA

5. TACCGGTA

6. GCCGGCA

7. TATTAATA[GA]

8. GCAGCAC

9. CAGCTAA

10. AAAAGAAG

11. CTCACTTC

12. CGCCGAAG

13. AACGGGGT

14. TTTAAAAA

15. GTGTCACA

16. AAAGCTTT

17. CGTGTTGC

18. ACTCGAGC

19. TAATATTA

20. AAAAATAT

ATACTGTA, CTACTGTA and GTAC[ACT]GTA merged to [ACG]TAC[ACT]GTA.
GCAGCAC and CAGCTAA merged to GCAGC[AT][CA]A.
TACCGGTA and GCCGGCA merged to T[AG]CCGG[TC]A

Table A: Shows the genewise conservation for hyphae motifs:

			[AC][AT]GCAGCA					
PITGs	1124	2598	4131	9793	20102			Totals
hit in sojae	1	1	0	1	1			4
hit in ram	1	1	0	1	1			4
consv in soj	0	0	0	0	0			0
consv in ram	0	0.5	0	0	0			0.5
consv in both	0.5	0	0	0.5	0.5			1.5
			T[AT]TTAATA					
PITGs	2972	15390	8193	8577	15239	9824	14003	
hit in sojae	1	1	1	1	1	1	1	7
hit in ram	1	1	1	1	1	1	1	7
consv in soj	0	0	0	0	0	0	0	0
cons in ram	0	0	0	0	0	0	0	0
consv in both	0	0	0	0	0	0	0	0
			[CG]ATTTTG					
PITGs	1124	1752	5616	6454	11883			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0.5	0	0	1			1.5
cons in ram	1	0	0	0.5	0			1.5
consv in both	0	0	0	0	0			0
			AAAAAAAT					
PITGs	1124	5067	14493	14950	15807			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0	0	0.5			0.5
cons in ram	0	0	0	0	0			0
consv in both	0	0	0	0	0			0
			AAATAAA					
PITGs	7160	11294	13448	14721	12551			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0	0	0			0
cons in ram	0	0	0	0	0			0
consv in both	0	0	0	1	0			1
			AAGTGGT					
PITGs	4131	4717	13448	16648	21293			
hit in sojae	0	1	1	1	1			4
hit in ram	0	1	1	1	1			4
consv in soj	0	0.5	0.5	0	0.5			1.5
cons in ram	0	0.5	0	0	0.5			1
consv in both	0	0	0	0	0			0
			ACCGGAA					
PITGs	1329	1398	6454					
hit in sojae	1	1	1					3
hit in ram	0	1	1					2
consv in soj	0	0	0					0
cons in ram	0	0	1					1
consv in both	0	1	0					1

Table A Contd.

			ACGCCGG			
PITGs	1329	2598	4131	9793	15731	Totals
hit in sojae	1	1	0	1	1	4
hit in ram	0	1	0	1	1	3
consv in soj	1	0	0	0	0	1
cons in ram	0	0	0	0	0	0
consv in both	0	1	0	0	0	1
			ACGGACG			
PITGs	6478	9354	12551	18312		
hit in sojae	0	1	1	1		3
hit in ram	0	1	1	1		3
consv in soj	0	0	0	0		0
cons in ram	0	0	0	0		0
consv in both	0	0	0	0		0
			AGAAAAA			
PITGs	1752	2335	6478	14055	15398	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0.5	0	0	0.5
cons in ram	0	0	0	0	0	0
consv in both	0	1	0	0	0	1
			[AT]GAAGCTG			
PITGs	10250	11883	14357	17507	18312	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0.5	0.5	0	0	1
cons in ram	0.5	0	0	0.5	0	1
consv in both	0	0	0	0	0	0
			CAATCAG			
PITGs	1752	10643	11450	12808	17161	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0.5	1	1	1	0	3.5
cons in ram	0	0	0.5	0.5	0	1
consv in both	0	0	0	0	0	0
			CCTCCAGC			
PITGs	127	2047	7340	954	21293	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	0.5	0.5
cons in ram	0	1	0	0	0	1
consv in both	0	0	0	0	0	0
			CGACGCC			
PITGs	1329	4010	6357	13762	14055	
hit in sojae	1	1	1	1	1	5
hit in ram	0	1	1	1	1	4
consv in soj	0	0.5	0	0	1	1.5
cons in ram	0	0.5	0.5	0	0	1
consv in both	0	0	0	0	0	0

Table A Contd.

	CGCTGGT					
PITGs	5067	10674				Totals
hit in sojae	1	1				2
hit in ram	1	1				2
consv in soj	0	0				0
consv in ram	0	0				0
consv in both	1	1				2
	CTGGAAA					
PITGs	754	2598	5117	7521	17507	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0.5	0.5	1
cons in ram	0.5	0	0	0	0	0.5
consv in both	0	1	0	0	0	1
	CTTCAAC					
PITGs	1397	1398	7068			
hit in sojae	1	1	1			3
hit in ram	1	1	1			3
consv in soj	0	0	0			0
cons in ram	0	0	0			0
consv in both	1	1	1			3
	GCA[AG]TGC					
PITGs	833	2277	5616	6570	14436	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0.5	0	0	1	0	1.5
cons in ram	0	0	0.5	0	0	0.5
consv in both	0	1	0	0	0.5	1.5
	GCTGC[TA]GT					
PITGs	5117	7340	10674	11573	13660	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0.5	0.5	0	0	0	1
cons in ram	0	0	0	0.5	0	0.5
consv in both	0	0	0	0	1	1
	GTACTAC					
PITGs	1218	2277	5007	12808	14695	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	0	0
cons in ram	0.5	0	0	0	0.5	1
consv in both	0	0	0	0	0	0
	GTTGAAG					
PITGs	287	693	3639			
hit in sojae	1	1	1			3
hit in ram	1	1	1			3
consv in soj	0	0	0			0
cons in ram	0	0	0			0
consv in both	1	0	1			2

Table A Contd.

	T[AG]C[AT]GTAC								
PITGs	754	7068	7160	14950	15731				Totals
hit in sojae	1	1	1	1	1				5
hit in ram	1	1	1	1	1				5
consv in soj	1	0	0	0.5	0				1.5
cons in ram	0.5	0	0	0	0				0.5
consv in both	0	0	0	0	0				0
	TACATGTA								
PITGs	833	1329	8704	9354	12562	14357	16926	17956	
hit in sojae	1	1	1	1	1	1	1	1	8
hit in ram	1	1	1	1	1	1	1	1	8
consv in soj	0	0	0	0	0	0	0	0	0
cons in ram	0	0	0	0	0	1	0	0.5	1.5
consv in both	0	0	0	0	0	0	0	0	0
	TACCGGTA								
PITGs	15607	17161	18083	18299	21410				
hit in sojae	1	1	1	1	1				5
hit in ram	1	1	1	1	1				5
consv in soj	0	0	0	0	0				0
cons in ram	0	0	0	1	0				1
consv in both	0	0	0	0	0				0
	TATTTTT								
PITGs	1218	4570	14055	14124	14375				
hit in sojae	1	1	1	1	1				5
hit in ram	1	1	1	1	1				5
consv in soj	0	0	1	0.5	0.5				2
cons in ram	0.5	0	0	0	0.5				1
consv in both	0	0	0	0	0				0
	TGCCCGGA								
PITGs	1329								
hit in sojae	1								1
hit in ram	0	14	hits	in	Pi				0
consv in soj	0								0
cons in ram	0								0
consv in both	0								0
	TGCTG[TG]C								
PITGs	2335	2598	5117	8599	10643				
hit in sojae	1	1	1	1	1				5
hit in ram	1	1	1	1	1				5
consv in soj	0	0	0	0	0				0
cons in ram	1	0	0	0.5	0				1.5
consv in both	0	0	0	0	0				0

Table B: Shows the genewise conservation for sporangia motifs:

			[CG]AAGAAG				
PITGs	2249	8419	9847	12827	20749		Totals
hit in sojae	1	1	1	1	1		5
hit in ram	1	1	1	1	1		5
consv in soj	0	0	0	0	0		0
consv in ram	0	0	0	0	0		0
consv in both	0	0	1	1	1		3
			AGAGACGC				
PITGs	2784	4075	7288	16466	17315		
hit in sojae	1	1	1	1	1		5
hit in ram	1	0	1	1	1		4
consv in soj	0	0	0	0	0		0
cons in ram	0	0	1	0	1		2
consv in both	1	0	0	0	0		1
			AGC[AG]CAAG				
PITGs	2460	3760	12568	14635	20221		
hit in sojae	0	0	1	1	1		3
hit in ram	0	0	1	1	1		3
consv in soj	0	0	0	0	0		0
cons in ram	0	0	0	0	0		0
consv in both	0	0	0	0	0		0
			ATGGCTAC				
PITGs	10162	12827	13461	13895			
hit in sojae	1	1	1	1			4
hit in ram	1	1	1	1			4
consv in soj	1	0	0	0			1
cons in ram	0	0	0	0			0
consv in both	0	0	1	0			1
			[CT]GCTCGAG				
PITGs	5447	10162	12577	17002	17975	12568	
hit in sojae	1	1	1	1	1	1	6
hit in ram	1	1	1	1	1	1	6
consv in soj	0	0	0	1	1	1	3
cons in ram	1	0.5	0	0	0	0	1.5
consv in both	0	0	0	0	0	0	0
			CTGCAAG				
PITGs	2784	3634	7288	9086	12827	16262	
hit in sojae	1	1	1	1	1	1	6
hit in ram	1	1	1	1	1	1	6
consv in soj	0	0	0	0	1	0	1
cons in ram	0	0	0	0	0	0	0
consv in both	0	0	0	0	0	0	0
			CA[AG]CAAC				
PITGs	5835	13011	15414	15858	17002	17975	
hit in sojae	1	1	1	1	1	1	6
hit in ram	1	1	1	1	1	1	6
consv in soj	0	0	0	0	1	0	1
cons in ram	0	1	0	0	0	0	1
consv in both	0	0	0	0	0	0	0

Table B Contd.

			CCGTTG					
PITGs	123	17002	16002	4121	9833	12567	14222	Totals
hit in sojae	1	1	1	1	1	1	1	7
hit in ram	1	1	1	0	1	1	1	6
consv in soj	0.5	0	1	1	1	0	1	4.5
cons in ram	1	0	0	0	0.5	1	0	2.5
consv in both	0	1	0	0	0	0	0	1
			CTTC[AG]AC					
PITGs	46	14222	1416	3886	4335	13769	9960	
hit in sojae	1	1	1	1	1	1	1	7
hit in ram	1	1	1	1	1	1	1	7
consv in soj	0	1	0	0	0	0	0	1
cons in ram	1	0	1	0	1	0	0	3
consv in both	0	0	0	1	0	1	1	3
			CTTCT[TC]C					
PITGs	2952	13895	16664					
hit in sojae	1	1	1					3
hit in ram	1	1	1					3
consv in soj	0	0	1					1
cons in ram	1	0	0					1
consv in both	0	1	0					1
			GAA[AG]AGA					
PITGs	2784	12827	14588	5477	15414			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0	0	0			0
cons in ram	0	0	0	0	0			0
consv in both	1	0	1	0	1			3
			GAAGC[GT][AG]C					
PITGs	1416	13461	14365	15867	15875			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0.5	0	0	0			0.5
cons in ram	0	0.5	0	0	0			0.5
consv in both	1	0	0	0	1			2
			GATCGAG					
PITGs	9847	9960	12182	15402	20800			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0.5	0	0			0.5
cons in ram	0	0	0	0	0			0
consv in both	1	1	0	0	1			3
			GTGC[AT]GCA					
PITGs	672	2952	10184					
hit in sojae	1	1	1					3
hit in ram	1	1	1					3
consv in soj	0	0	0					0
consv in ram	0	0	0					0
consv in both	0	1	1					2

Table B Contd.

	GTCGTTT							
PITGs	121	1416	5835	17975	14222	14281		Totals
hit in sojae	1	1	1	1	1	1		6
hit in ram	1	1	1	1	1	1		6
consv in soj	0	0	0	0	1	0		1
consv in ram	0	0	0.5	0	0	0		0.5
consv in both	0	1	0	1	0	0		2
	GCTGC[AT]G							
PITGs	937	4454	12567	10162	9833			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0	0	0			0
consv in ram	0	0	0	0	0			0
consv in both	1	0	0	0	1			2
	GTGGGGTG							
PITGs	121	13001	15858	16338				
hit in sojae	1	1	1	1				4
hit in ram	1	1	1	1				4
consv in soj	0	0	0	0				0
consv in ram	0	0	0	0				0
consv in both	0	1	0	1				2
	TACATGTA							
PITGs	1101	3192	4949	5192	13196	6087		
hit in sojae	1	0	1	1	1	1		5
hit in ram	1	0	1	0	1	1		4
consv in soj	0	0	0	1	0	0		1
consv in ram	0	0	0	0	0	0		0
consv in both	0	0	0	0	0	0		0
	T[GC]GAGTTT							
PITGs	2342	2460	3760	6741	7288	10352	10746	
hit in sojae	1	0	0	1	1	1	1	5
hit in ram	1	0	0	1	1	1	1	5
consv in soj	0	0	0	0	0	1	0	1
consv in ram	0	0	0	0	0	0	0	0
consv in both	0	0	0	0	0	0	0	0
	TG[CG]CTG[CG]C							
PITGs	12567	13769	14087	17975	20527			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	1	0	0	0			1
consv in ram	0	0.5	0	0	0			0.5
consv in both	0	0	0	0	0			0
	TT[CT][AT]TTTT							
PITGs	2342	4075	13183	15875	16089	20042		
hit in sojae	1	0	1	1	1	1		5
hit in ram	1	0	1	1	1	1		5
consv in soj	0	0	0	0	0	0		0
consv in ram	0	0	0	0	0	0		0
consv in both	0	0	0	0	0	0		0

Table B Contd.

	TTGAAGT							
PITGs	937	1416	5525	12567	12827	15875	16262	Totals
hit in sojae	1	1	1	1	1	1	1	7
hit in ram	1	1	1	1	1	1	1	7
consv in soj	0	1	0	1	1	0	0.5	3.5
consv in ram	0.5	0	0	0	0	0	0.5	1
consv in both	0	0	0.5	0	0	1	0	1.5

Table C: Shows the genewise conservation for cleavage motifs:

			[AT]CCACCA			
PITGs	5714	10630	12293	13755	12524	Totals
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	1	1	0	0	0	2
consv in ram	0	0	0	1	0	1
consv in both	0	0	1	0	1	2
			AAAAAT[GA][AT]			
PITGs	891	3346	6236	10523	11238	
hit in sojae	1	1	1	0	1	4
hit in ram	1	1	1	0	1	4
consv in soj	0	0	0	0	0	0
cons in ram	0	0.5	0	0	0	0.5
consv in both	0	0	0	0	1	1
			AAGCA[AG]A			
PITGs	591	10507	20886	20590		
hit in sojae	1	1	1	1		4
hit in ram	1	1	1	1		4
consv in soj	0	0	0	0		0
cons in ram	0	0	0	0		0
consv in both	1	1	0	1		3
			[ACG]AAGT[ACG]G			
PITGs	6965	12293				
hit in sojae	1	1				2
hit in ram	1	1				2
consv in soj	0	0				0
cons in ram	0	0				0
consv in both	1	1				2
			ACT[CT]G[AGC]AG			
PITGs	4477	16727	10847	18680	16356	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	0	0
cons in ram	0	0	1	0	0	1
consv in both	0	0	0	0	1	1
			AGAGAGAG			
PITGs	3034	8258	17344			
hit in sojae	1	0	1			2
hit in ram	1	1	1			3
consv in soj	0	0	0			0
cons in ram	0	1	0			1
consv in both	1	0	0			1
			AGCATC			
PITGs	2227	2028	6965	8404	16727	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	0	1	4
consv in soj	0	0	0	0	0.5	0.5
cons in ram	0	0	1	0	0	1
consv in both	1	0	0	0	0	1

Table C Contd.

			AGTTG[AC]A			
PITGs	2227	5149	13149	1266	12352	Totals
hit in sojae	1	1	1	1	0	4
hit in ram	1	1	1	1	0	4
consv in soj	0	1	1	0	0	2
cons in ram	0	0	0	0	0	0
consv in both	1	0	0	0	0	1
			[AT]GGAGGAG			
PITGs	12507	21207	4477	10507	12293	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	0	0
cons in ram	1	1	0	0	0	2
consv in both	0	0	0	0	0	0
			ATT[CT]TTA			
PITGs	891	12507	13149	21207		
hit in sojae	1	1	1	1		4
hit in ram	1	1	1	1		4
consv in soj	1	0	0	0		1
cons in ram	0	0	1	0		1
consv in both	0	1	0	1		2
			CAACA[GA]CA			
PITGs	5149	8404	11238	11239	21452	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	0	4
consv in soj	1	1	0.5	1	1	4.5
cons in ram	0	0	1	0	0	1
consv in both	0	0	0	0	0	0
			CATC[AC]A[CT]G			
PITGs	2227	6835	7444	9979	20710	
hit in sojae	1	1	0	1	1	4
hit in ram	1	1	0	1	1	4
consv in soj	0	0	0	0	0	0
cons in ram	0	0	0	0	0	0
consv in both	1	0	0	0	0	1
			CGCCACC			
PITGs	10847	12507	20590	21207		
hit in sojae	1	1	1	1		4
hit in ram	1	1	1	1		4
consv in soj	0	1	1	0		2
cons in ram	1	0	0	0		1
consv in both	0	0	0	0		0
			TTTAAAAA			
PITGs	5111	5714	12293	11238	20886	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	0	0
cons in ram	0	0	0	0	0	0
consv in both	0	0	1	0	0	1

Table C Contd.

		CGTCCTCG				
PITGs	16231	17591	3525	18680		Totals
hit in sojae	1	1	1	1		4
hit in ram	1	1	1	1		4
consv in soj	0	0	0	0		0
consv in ram	0	0	0	0		0
consv in both	1	1	0	1		3
		CTCCTTC				
PITGs	5296	18680	9899	13601	16967	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	1	1
cons in ram	0	1	0	0	0	1
consv in both	1	0	0	0	0	1
		CTTCAAC				
PITGs	591	5670	9979	18174	20590	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	1	0	0	0	1
cons in ram	0	0	1	0	0	1
consv in both	1	0	0	0	1	2
		CTTCGAG				
PITGs	5714	8404	18386	21207	21452	
hit in sojae	1	1	1	1	1	5
hit in ram	1	0	1	1	1	4
consv in soj	0.5	0	0	0.5	0	1
cons in ram	0	0	0	0	0	0
consv in both	0	0	0	0	0	0
		CTTTTG				
PITGs	2030	3467	3590	5111	13601	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0.5	0.5	0	0	1
cons in ram	0	0	0	0	0.5	0.5
consv in both	0	0	0	1	1	2
		G[AC]AGCC[AG]				
PITGs	2008	5149	11238	16356	11400	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	1	0.5	0	0	0	1.5
cons in ram	0.5	0	0	0	0	0.5
consv in both	0	0	0	0	0	0
		GAGCT[CG]C				
PITGs	3590	5205	12507	21207	13419	
hit in sojae	1	1	1	1	1	5
hit in ram	1	1	1	1	1	5
consv in soj	0	0	0	0	0	0
cons in ram	0	0	0.5	0.5	0	1
consv in both	1	0	0	0	0	1

Table C Contd.

	GATGCTG							
PITGs	5714	11504	12507	18386	18393			Totals
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	0			4
consv in soj	1	0	0	0	1			2
cons in ram	0	0	0	0	0			0
consv in both	0	0	0	0	0			0
	GCAC[CG]AC							
PITGs	2110	3590	5205	19483	20590			
hit in sojae	1	1	1	0	1			4
hit in ram	1	1	1	0	1			4
consv in soj	0	0	0	0	0.5			0.5
cons in ram	0	0	0	0	0			0
consv in both	0	0	0	0	0			0
	GCGCTC							
PITGs	2008	6049	7355	9899	12507			
hit in sojae	1	0	1	1	1			4
hit in ram	1	0	1	1	1			4
consv in soj	0.5	0	0	0.5	0.5			1.5
cons in ram	0	0	0	0	0.5			0.5
consv in both	0	0	1	0	0			1
	GCTC[AC]AA							
PITGs	5149	5714	15282	18386	5296			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0	0	0			0
cons in ram	0	0	0	0	1			1
consv in both	0.5	0	0	0	0			0.5
	TACATGTA							
PITGs	891	3346	5111	7444	8707	10571	13491	
hit in sojae	1	1	1	0	0	0	1	4
hit in ram	1	1	1	0	0	0	1	4
consv in soj	0	0	0	0	0	0	0	0
cons in ram	0	0	0	0	0	0	0	0
consv in both	0	0	0	0	0	0	0	0
	TCGTC[GT]TC							
PITGs	591	3034	4281	20590				
hit in sojae	1	1	1	1				4
hit in ram	1	1	1	1				4
consv in soj	1	0	0	1				2
cons in ram	0	0	0	0				0
consv in both	0	0	0	0				0
	TAAATAA							
PITGs	2029	6835	16967	18428	20710			
hit in sojae	1	1	1	1	1			5
hit in ram	1	1	1	1	1			5
consv in soj	0	0	0	0	0			0
cons in ram	0	0	0	0	0			0
consv in both	0	0	1	0	0			1

Table C Contd.

	TACCGGTA				
PITGs	17591	17675	13036	19483	Totals
hit in sojae	1	1	0	0	2
hit in ram	1	1	0	0	2
consv in soj	0	0	0	0	0
consv in ram	0	0	0	0	0
consv in both	0	0	0	0	0

Table D: Shows the genewise conservation for swimming zoospore motifs

			[AC]C[CGT]TCTTC				
PITGs	971	3267	13041	13283	16978		Totals
hit in sojae	0	1	1	1	0		3
hit in ram	0	1	1	1	1		4
consv in soj	0	0.5	0	0.5	0		1
consv in ram	0	0.5	0	0	1		1.5
consv in both	0	0	0	0	0		0
			A[CG]GAAGA[ACG]G				
PITGs	9124	11231	14267	15838	17040	18473	
hit in sojae	1	1	0	1	1	1	5
hit in ram	1	1	1	1	1	1	6
consv in soj	0	0	0	0	0	0	0
cons in ram	0.5	0	0	0	0	0	0.5
consv in both	0	1	0	0	0	0.5	1.5
			AGAAACGA				
PITGs	1526	4486	9954				
hit in sojae	1	1	1				3
hit in ram	1	1	1				3
consv in soj	0	0	0				0
cons in ram	0	0	0				0
consv in both	0	1	1				2
			AGAGCCTG				
PITGs	11231	11594	13283	16978	17040		
hit in sojae	1	1	1	0	1		4
hit in ram	1	1	1	1	1		5
consv in soj	0	0	0	0	0		0
cons in ram	0	0	0	1	1		2
consv in both	0	0	0	0	0		0
			CC[CT][TG]CAGC				
PITGs	4486	6369	10035	12832	13283		
hit in sojae	1	1	1	1	1		5
hit in ram	1	1	1	1	1		5
consv in soj	1	0	0	1	0		2
cons in ram	0	0	0	0	0		0
consv in both	0	0	1	0	1		2
			GTCACTGA				
PITGs	7866	8306					
hit in sojae	1	1					2
hit in ram	1	1					2
consv in soj	0	0					0
cons in ram	0	0					0
consv in both	1	1					2
			TACATGTA				
PITGs	1360	6370	14172	17040	18999		
hit in sojae	1	1	1	1	1		5
hit in ram	1	1	1	1	1		5
consv in soj	0	0	0	0	0		0
cons in ram	0	0	0	0.5	0.5		1
consv in both	0	0	0	0	0		0

Table E: Shows the genewise conservation for germinating cyst motifs:

			[ACG]TAC[ACT]GTA				
PITGs	339	6099	7387	11891	13196		Totals
hit in sojae	1	1	1	1	1		5
hit in ram	1	1	0	1	1		4
consv in soj	0	0	0.5	0	0		0.5
consv in ram	0	0	0	0	0		0
consv in both	0	0	0	0	0		0
			AAAGCTTT				
PITGs	2332	5881	1101	7143	4325		
hit in sojae	1	1	1	1	1		5
hit in ram	1	1	1	1	1		5
consv in soj	0	0	0	0	0		0
cons in ram	1	1	0	0	0		2
consv in both	0	0	0	0	0		0
			CGCCGAAG				
PITGs	698	13374	13636	14720	17063		
hit in sojae	1	1	1	1	0		4
hit in ram	1	1	1	1	1		5
consv in soj	0	0.5	0	0	0		0.5
cons in ram	0	0	1	0.5	0		1.5
consv in both	0	0	0	0	0		0
			CTCACTTC				
PITGs	1101	5603	9173	10037	14360		
hit in sojae	1	1	1	1	0		4
hit in ram	1	1	1	1	0		4
consv in soj	0	1	0	0	0		1
cons in ram	0	0	0	0	0		0
consv in both	0	0	0	0	0		0
			TACATGTA				
PITGs	6099	6476	9173	13636	15847	17252	
hit in sojae	0	1	1	1	1	1	5
hit in ram	0	1	1	1	1	1	5
consv in soj	0	0	0.5	0.5	0	0.5	1.5
cons in ram	0	0	0	0.5	0	0.5	1
consv in both	0	0	0	0	0	0	0
			GCAGC[AT][GA]A				
PITGs	2972	6099	8879	11891	11944	14238	
hit in sojae	1	1	0	1	1	1	5
hit in ram	1	1	0	1	1	1	5
consv in soj	0	0	0	1	0.5	0	1.5
cons in ram	0.5	0.5	0	0.5	1	1	3.5
consv in both	0	0	0	0	0	0	0
			T[AG]CCGG[TC]A				
PITGs	5603	6476	7383	13196	13680		
hit in sojae	1	1	1	1	1		5
hit in ram	1	1	1	1	1		5
consv in soj	0	0	0.5	0	0		0.5
cons in ram	0	0	0	0	0		0
consv in both	0	0	0	0	0		0

Table E Contd.

	TTTAAAAA					
PITGs	7452	8879	14003	14360	6467	Totals
hit in sojae	0	0	1	1	1	3
hit in ram	0	0	1	1	1	3
consv in soj	0	0	0	0	0	0
cons in ram	0	0	0	0	0	0
consv in both	0	0	0	0	0	0

	TATTAATA[GA]						
PITGs	2972	8193	8577	9824	14003	15239	
hit in sojae	1	1	1	1	1	1	6
hit in ram	1	1	1	1	1	1	6
consv in soj	0	0	0	0	0	0	0
cons in ram	0	0	0	0	0	0	0
consv in both	0	1	0	0	0	0	1

Table F: Shows the list of transformants used for the functional analyses of the three motifs viz. GCTGCTG, TATTAATA and TACATGTA

Sporangia specific motif: GCTGCTG (SP1)					
Construct name	**Transformant #**				
SP1	1.19	1.20	1.21	3.10	3.20
Germinating cyst motif: TATTAATA (GC2)					
Construct name	**Transformant #**				
GC2	2	10	13	17	22
All stages motif: TACATGTA (GC1)					
Construct name		**Transformant #**			
GC1	4	16	25		

Chapter III

Core promoter elements in Irish potato famine pathogen *Phytophthora infestans:* their consensus, effect in gene expression and distribution within the Heterokontophyta

ABSTRACT:

The core promoter, which is the ultimate target for most factors controlling transcriptional activity, usually draws less attention than proximal elements in analyses of promoters. A computational analysis of the core promoter regions of *Phytophthora infestans,* the most devastating potato pathogen, is presented in this chapter. *P. infestans* belongs to the Oomycete class of the phylum Heterokontophyta. Sets of core promoter regions, 50 bases on either side of the putative transcription start sites based on EST data were assembled. These sets were searched *in silico* for overrepresented motifs and positional bias of the motifs. A genome-wide analysis was also done by searching 200 bases upstream of the translation start sites of all *P. infestans* genes including those lacking EST data. This resulted in a better *Phytophthora*-specific consensus for elements like Initiator (Inr) and Flanking Promoter Region (FPR) that were previously identified in studies involving a limited number of oomycete genes. A novel seven base element, Downstream Promoter Element for *Pythiales* (DPEpyth) was detected. Genes with none of the three core promoter elements or with just Inr had equal probability of being constitutively or differentially expressed. Whereas, those with either FPR or DPEpyth in addition to Inr, were in

most cases differentially expressed between developmental stages. The distribution of Inr, FPR and DPEpyth within eight other Heterokonts (five oomycetes including two *Phytophthora species*, two diatoms and a brown alga), were checked. While FPR was found in most oomycetes, DPEpyth was detected primarily in the *Pythiales (Pythium and Phytophthora)*. Core promoter elements identified in other organisms, like TATA-box, MTE, DPE, etc., were not detected at any significant level in *P. infestans*.

INTRODUCTION:

Precise control of gene expression at the transcriptional level is required for proper growth and development of any organism. In eukaryotes, the binding sites for the transcription machinery usually are genomic DNA sequence elements that act as signals for regulation and are named enhancers, proximal promoters, or core promoter elements depending on their location (Ohler and Wassarman, 2010). Core promoter elements, unlike enhancers and proximal elements which are found at varying positions with respect to the transcription start site (TSS), are present within ~50 bases on either side of the TSSs in most genes transcribed by RNA polymerase II. The core promoter can be defined biochemically as the minimal stretch of DNA that is sufficient to accurately direct basal levels of transcription initiation by Pol II *in vitro* on naked DNA templates containing a single well-defined transcription start site (TSS; Müller 2007, Butler and Kadonaga 2002, Struhl 1987, Weis and Reinberg 1992, Smale 1997, 2001, Smale et al. 1998, Burke et al. 1998, McLeod et al. 2004, Burke and Kadonaga, 1997). Most

known core promoter elements are found in focused or single-peak core promoters, where initiation occurs within a short region (Juven-Gershon et al., 2008). On the other hand, dispersed core promoters like XCPEI can direct initiation only with the help of other sequence-specific activators (Juven-Gershon et al., 2008). The latter play a key role in orchestrating accurate transcription initiation (McLeod et al., 2004, Burke and Kadonaga, 1997), by directing the assembly of general transcription factors, Mediator (McLeod et al., 2004), and several other factors that make up the basal transcription machinery (Hochheimer and Tjian, 2003, Woychik and Hampsey 2002, and Hampsey, 1998). There is considerable structural and functional diversity in the core promoters that have been studied (Smale and Kadonaga, 2003) and this makes the detection of core promoter elements extremely difficult (Levine and Tjian, 2003). Still, there are some well-characterized core promoter elements, with TATA-box and initiator (Inr) being the two most widely studied. No known sequence motif is universal (Juven-Gershon et al., 2008) or taxon-specific. Therefore, an insight into core promoter elements in taxa that are not well-studied can lead us to a better understanding of the evolution of the basal transcription machinery.

Very little is known about the core promoter elements responsible for transcription of genes in the kingdom *Heterokontophyta,* which includes brown algae, chrysophytes, diatoms and some protozoa in addition to oomycetes. Oomycetes include both plant and animal pathogens in addition to saprophytes. *Phytophthora infestans,* an oomycete, is one of the most devastating

phytopathogens, responsible for the late blight disease in potato. The *P. infestans* genome was sequenced recently (Haas et al., 2009), which resulted in a plethora of data related to this organism. In this study the data that is already available was used to conduct a genome-wide study of the core promoter elements in *P. infestans,* to obtain a better understanding of the transcription mechanism in this economically important pathogen.

I looked for overrepresented motifs in the region where one would detect the core promoter elements, i.e. ~50 bases on either side of the putative TSS, and then checked if those elements show some positional bias within the said region. Once the core promoter elements were detected, the expression patterns of genes with different combinations of the three core promoter elements, were checked, to get an idea of the effects that these have on gene expression. Differences in expression of genes with different combinations of core promoter elements has been reported previously (Burke et al., 1998). I have also looked into the genomes of five other oomycetes (*Phytophthora sojae*, *Phytophthora ramorum*, *Pythium ultimatum*, *Hyaloperonospora arabidopsidis* and *Saprolegnia parasitica*), two diatoms (*Thalassiosira psuedomona* and *Phaeodactylum tricornutum*) and one brown alga (*Ectocarpus siliculosus*) to find out the extent of conservation of the three elements.

MATERIALS AND METHODS

Sequences:

P. infestans and *S. parasitica* sequences were obtained from the Broad Institute (http://www.broadinstitute.org/annotation/genome). *P. sojae*, *P. ramorum*, *T. pseudonana* and *P. tricornutum* sequences were from the Joint Genome Institute (http://genome.jgi-psf.org). *H. arabidopsidis* sequences were from the Virginia Bioinformatics Institute (http://vmd.vbi.vt.edu), *E. siliculosus* sequences were from the University of Gent genome portal (https://bioinformatics.psb.ugent.be/gdb/ectocarpus) and *P. ultimatum* sequences were obtained from the Michigan State University database (http://pythium.plantbiology.msu.edu/).

Softwares:

MEME (Baily and Elkan, 1994) was used for the detection of overrepresented motifs. ClustalW (Thompson et al., 1994) was used for multiple sequence alignments to check for evolutionary conservation. PERL scripts that were developed in-house were used for sequence extraction and positional bias analyses.

Assembly of gene sets and sequence extraction:

Translation start positions of all the genes and ESTs in the database were extracted and mapped with PERL scripts. The EST start sites were considered to be putative Transcription Start Sites (TSSs). Based on the EST evidences three different gene sets were assembled: a) High Confidence Set: genes for which

there were two or more EST evidences for the putative TSSs; b) Expanded Set: genes for which there was one EST evidence for the putative TSSs; c) Total Set: all genes in the database, which is actually a superset of the first two sets that also includes the genes for which there were no EST evidences.

For the High Confidence and the Expanded sets, sequences 50 bases on either side of the putative TSSs were extracted from the *P. infestans* database with a PERL script that was developed in-house. In case of the Total Set, 200 bases upstream of the translation start codon (ATG) were extracted for all genes.

Apart from the three afore-mentioned sets of sequences from *P. infestans*, seven other sequence sets, which included 200 bases upstream of ATGs for all *P. sojae, P. ramorum. P. ultimatum, H. arabidopsis, S. parasitica, T. pseudonana*, *P. tricornutum and E. siliculosus* genes respectively, were assembled. Some of these were extracted with the same script used for the *P. infestans* Total Set. For others the coordinates for the translation start sites were extracted based on their respective .gff files and then the sequences were extracted from the databases with the help of other PERL scripts developed in-house.

Detection of overrepresented motifs:

Stand-alone MEME version 4.3.0, with a minimum width (minw) of 5 and a maximum width (maxw) of 8 was used to look for overrepresented motifs. The default gap opening cost (wg) and gap extension cost (ws) for multiple alignments of 11 and 1 respectively were used. The distribution of motifs (mod)

used was “anr” with a default E-value cut-off (evt) of 1e-05 and 5 as the maximum number of EM iterations (maxiter) to run. The minimum sites for each motif (minsites) used were 5 with the rest of the parameters being default.

The expected value for each motif was calculated by dividing the total number of bases being looked at, with the probability of finding the motif randomly. This expected value was then used to calculate the observed to expected ratio.

Detection of positional bias:

Positional bias of all overrepresented motifs for the High Confidence and Expanded sequence sets were checked by detecting their frequencies within each of the ten 10 bp windows that the 100 bp sequences were divided into. The positions where the motifs end were used for calculating their frequencies in each window. The average and the expected frequencies were taken into consideration while checking for the positional bias of these motifs.

For the *P. infestans* Total Set and that of all the eight other heterokonts, the 200 bases upstream of the ATGs were divided into four fifty base windows for the purpose of checking the positional bias for each of the motifs overrepresented in the *P. infestans* High Confidence and Expanded sets.

Analyzing the effects of Inr, FPR and DPEpyth on gene expression :

Effects of core promoter elements on the expression of genes were checked with the help two different analyses. First, maximum expression of the genes in five key asexual stages of the *P. infestans* life cycle, viz. hyphae,

sporangia, cleavage, swimming zoospores and germinating cysts, was looked at. For the second analysis, the maximum fold-change in expression between any two stages was calculated with the help of per-gene normalized expression data. Previously published microarray data (Judelson et al., 2008) was used for both analyses.

Checking for distribution of *P. infestans* core promoter elements in other heterokonts:

To analyze the distribution of the *P. infestans* core promoter elements in the eight other genomes, I checked a) if the *P. infestans* core elements were overrepresented, and b) if these showed similar positional bias in the other genomes. A PERL script developed in-house was used to determine the frequency of these motifs in each of the eight other genomes. The equality of proportions was checked for the observed and expected frequencies for each motif and a p-value cut-off of 0.05 (95%) was used to determine the significance.

Looking for core promoter elements found in other eukaryotes:

The Total Set was used to look for the presence and distribution of the most common core promoter elements that have been detected in other organisms previously. A PERL script was used to detect the frequency and then the overrepresentation was calculated by assessing the expected frequency of the motifs in the total number of bases searched.

RESULTS:

Assembly of gene sets and sequence extraction:

The primary goal of this study is the detection of core promoter elements in *P. infestans*. Core promoter elements are found ~50 bases on either side of the TSS. Therefore, having an idea of where the TSSs for the genes are is essential. Since TSSs are not annotated in the *P. infestans* database, I looked for EST evidence to predict TSSs for the genes. I searched for ESTs that had their 5' termini within 29 to 150 bases upstream of the ATG (the region of interest). It has been shown in yeast and rice that the minimum distance between the translation start site (ATG) and the TSS required for the RNA pol II to function is approximately 30 bases (Zhang and Dietrich, 2005, Zhu et al., 1995). It is worth mentioning that ESTs often do not reach the actual 5' end of the gene. Therefore, the EST start positions may be close to, but not precisely define, the real TSSs. This is one rationale for using a 100 bp window for motif searching in this study. After matching the starting positions of all genes (predicted genes in the Broad Institute database when the analysis was conducted) with that of all ESTs, it was found that there were only 3129 ESTs, ~4% of the total (74,135), that had their 5' terminus within the region of interest.

Based on the EST data two different gene sets were assembled as mentioned in the methods section. The High Confidence Set had 121 genes for which there were strong EST evidences for the TSS. In this set two or more ESTs started within two bases of each other within the 100-bp window; the most-

upstream EST was used to define the TSS. The second set, called the Expanded Set, is a collection of 571 genes that had a single EST evidence for the TSSs, i.e. no two EST starts were at the same point or within two bases of each other, within the 100-bp window. As with the first set, the position that was most upstream of the ATG was considered to be the TSS. A third group, the Total Set, was assembled that consisted of all (18178) genes in the database.

For the High Confidence and Expanded Sets, 100 bases (50 bases on either side of the putative TSS) were extracted for motif searches. The 200 bases upstream of the ATG were searched from the Total Set.

Overrepresented motifs within the core promoter region and their positional bias:

There are multiple tools available for detecting overrepresented motifs. Here only MEME was used, as a relatively small region was searched, and the expectation maximization technique that MEME uses was likely to create less redundancy than Gibbs sampling or enumerative search methods, with fewer false positives, compared to other methods.

Most of the overrepresented motifs in the High-confidence and Expanded sets were very similar to the initiator (Inr) and flanking promoter region (FPR) motifs that were first detected in *P. infestans* genes *ipiB* and *ipiO* by Pieterse and his colleagues (Pieterse et al.1994). This was improved upon by McLeod et al. in 2004, who used 15 oomycete genes to define the oomycete Inr consensus as YCATTYY (McLeod et al., 2004). Our results from MEME suggested that there

could be a C instead of a T at the fourth position. Therefore, the frequency and distribution of variations of the Inr were checked, by changing the C and the two T residues around the third position purine (A) (Fig 1, Table 1) to degenerate pyrimidines (C to T/Y, T to C/Y). The effect of degeneracy at the third position was not checked as McLeod et al. (2004) have shown that this was the transcription start site for most oomycete genes with Inr; the phenomenon of transcription starting at the A within the Inr, has been observed in other organisms (Smale and Kadonaga, 2003, Parry et al., 2010). Fig 1 shows that most Inr-like motifs were within ten and twenty bases upstream of the predicted TSSs in the High-confidence and Expanded sets, respectively. The fact that the peaks for the Inr were somewhat broad may reflect that most of the ESTs are not full-length.

By changing the Inr consensus from YCATTYY to YCAYTYY, the number of occurrences in the High-confidence sequence set increased from 35 to 53 in the 20-nt region just upstream of the putative TSS. Therefore, more Inr sequences were captured near EST verified TSSs. It should be mentioned that the expected to observed ratio as a result of this change was comparable to that of the of the McLeod definition. A similar increase resulted within the Expanded Set (Fig 1). But, changing the C and T at positions 2 and 5 to any pyrimidine (Y) did not increase the number of occurrences to such an extent (Table 1) in any of the two sets, when compared to the hits outside the 20 base window.

Table 1:

Different Inr definitions and their frequencies within the core promoter region:

The table shows the frequency of the different Inr definitions within each 10 nt window of the genes that belongs to the High-confidence set in *P. infestans*. Starting with the Inr as defined in *Drosophila,* degeneracy was introduced into each of the pyrimidine sites. 3rd column from the left shows the definition by Peiterse et al. (TCAYTTY; 1994), the 4th shows the definition from McLeod et al. (YCAYTYY; 2004) and the 5th column from the left shows our definition (YCAYTYY; bold). TSS is considered as +1. The 6th and the 7th column shows adding more degeneracy does not increase the frequencies much within the twenty base window (-20 to +1).

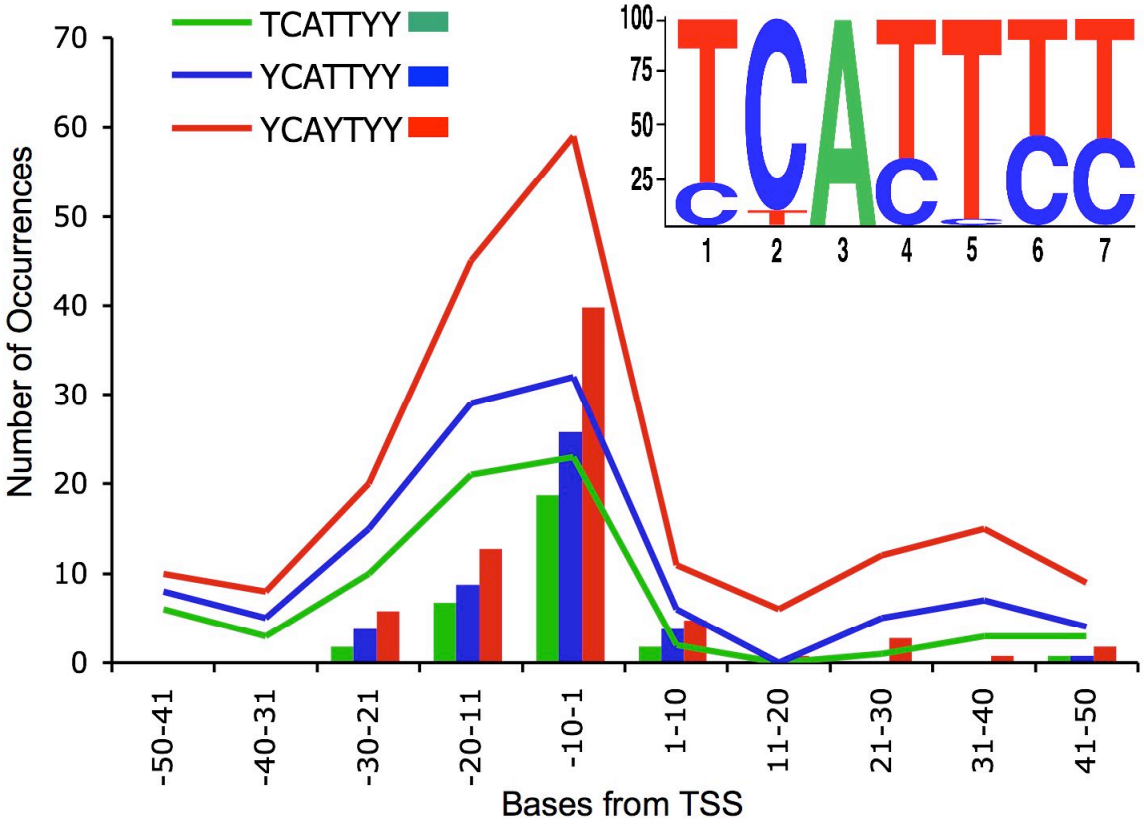
Bases from TSS	TCATTYY	TCAYTTY	YCATTYY	**YCAYTYY**	YYAYTYY	YYAYYYY
-50-41	0	0	0	0	2	4
-40-31	0	0	0	0	3	3
-30-21	2	0	4	6	6	9
-20-11	7	6	9	13	14	15
-10-1	19	17	26	40	43	44
1-10	2	3	4	5	5	7
11-20	0	0	0	1	1	2
21-30	0	0	0	3	3	8
31-40	0	0	0	1	5	9
41-50	1	0	1	2	3	7

Fig 1 legend:

Distribution of three possible Inr definitions in *P. infestans* gene core promoter regions:

Shown is the distribution of our suggested definition (YCAYTYY) of Inr in *P. infestans*. It is compared with the two other definitions, that of Pieterse et al. (1994; TCATTYY) and McLeod et al. (2004; YCATTYY), within the High Confidence (bars) and the Expanded (lines) sets. The logo (upper right hand corner) is derived from the High Confidence Set results.

Fig 1


TCATTYY
YCATTYY
YCAYTYY
Number of Occurrences
0
10
20
30
40
50
60
70
-50-41
-40-31
-30-21
-20-11
-10-1
1-10
11-20
21-30
31-40
41-50
Bases from TSS
100
75
50
25
1
2
3
4
5
6
7

The same is true for the Total Set even though the increase in frequency is within a region broader than 20 bases in case of the total set. This is, expected as for this set I was looking at the regions upstream of the ATGs for all the genes in the database instead of the regions around the putative TSSs. I also checked for the frequency of each of the bases in all the degenerate positions, for the Inrs within the High Confidence Set, to come up with a better definition for oomycete specific Inr (Fig 1) which happens to fit into the consensus for the more generalized eukaryotic Inr i.e. $YYA_{+1}NWYY$.

The FPR (Flanking Promoter Region) is a core promoter element that was detected, in 16 oomycete promoters by McLeod et al. (2004), and the suggested consensus is CAWTTTNYY. Our results from the High Confidence Set indicate that the pyrimidine at the eighth position is ~90% of the time a cytosine. A cytosine at that position also decreases the false positives outside the 20 base window (-10 to +10) for which the FPR shows a bias. Also, the seventh position N is a guanine approximately 60% of the time and a cytosine was detected ~20% of the time in place of the second position Adenine (Fig 2, Table 2). As in case of Inr, to find out if more FPRs could be captured near the EST verified TSSs I checked for a truncated version by taking off two terminal bases. The increase in occurrences was significantly more than expected by random chance for the truncated motif MWTTTNC (obtained by removing the first and the last bases from the McLeod definition, introducing a degeneracy at the first position and taking degeneracy off the last position) (Fig 2), suggesting that the two bases on

either side of this motif may not critical. A similar increase in occurrences was observed in case of the Expanded Set (Fig 2) and the Total Set (results not shown) of genes. FPR was found to be highly overrepresented (at least five and a half times higher than any other 20 base window) in the region -10 to +10, which is just downstream of each of the two windows where the Inr is overrepresented. This is consistent with the results reported previously by McLeod et al. (2004)

Table 2:

Different FPR definitions and their frequencies within the core promoter region:

The table shows the frequency of the different FPR definitions within each 10 nt window of the genes that belong to the High-confidence set in *P. infestans*. The second column from left shows FPR (CAWTTTNYY), as defined by McLeod et al. in 2004. The third column shows that the 2nd position to the right of the three thiamines, is almost always a cytosine. The fourth column shows that the base to the right of the three thiamines is a guanine ~60% of the times. The fifth column shows that the second position of the McLeod defined FPR can also be a M. The next two columns to the right supports the presence of a guanine and a cytosine next to the three thiamines ~60% and ~90% of the times. The last but one column from the left shows the truncated version and our definition MWTTTNC (bold) of the FPR. The last column shows that a guanine is present next to the three thiamines ~60% of the times.

Table 2: Frequencies of different FPR definition

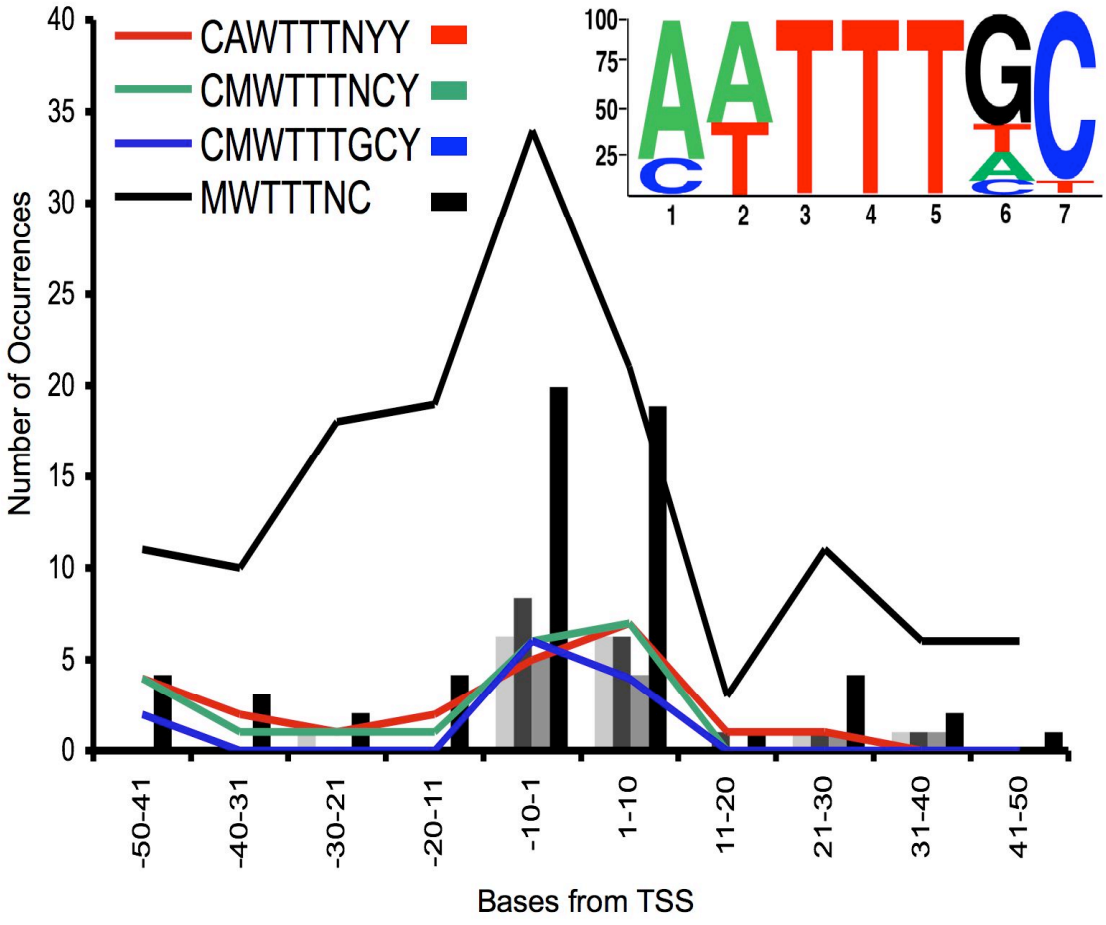
Bases from TSS	CAWTTTNYY	CAWTTTNCY	CAWTTTGCY	CMWTTTNYY	CMWTTTNCY	CMWTTTGCY	**MWTTTNC**	MWTTTGC
-50-41	0	0	0	0	0	0	4	1
-40-31	0	0	0	1	0	0	3	2
-30-21	1	0	1	1	0	0	2	1
-20-11	0	0	0	0	0	0	4	0
-10-1	6	6	3	8	8	5	19	7
1-10	6	5	4	7	6	4	18	11
11-20	0	0	0	1	1	0	1	0
21-30	1	1	0	1	1	1	4	1
31-40	1	0	0	1	1	1	2	0
41-50	0	0	0	0	0	0	1	1

Fig 2 legend:

Distribution of four possible FPR definitions in the *P. infestans* gene core promoter regions:

Shown is the distribution for the definition for FPR put forward by McLeod et al. (2004) in *P. infestans* (CAWTTTNYY) and its comparison with (CMWTTTNCY) within the High Confidence (bars) and the Expanded (lines) sets. Distribution of (CMWTTTGCY) shows that the seventh position N is a Guanine in about 60% cases. Also shown is the distribution of the truncated version and our definition of FPR (MWTTTNC). The logo (upper right hand corner) is derived from the results of the High Confidence Set.

Fig 2


CAWTTTNYY
CMWTTTNCY
CMWTTTGCY
MWTTTNC
Number of Occurrences
0
5
10
15
20
25
30
35
40
-50-41
-40-31
-30-21
-20-11
-10-1
1-10
11-20
21-30
31-40
41-50
Bases from TSS
100
75
50
25
1
2
3
4
5
6
7

The region of overrepresentation for FPR, like Inr, broadens as one goes from high confidence to low confidence to the total gene sets due to the decrease in confidence about the TSS from one set to another with the increase in the number of genes. Like with Inr, a better oomycete specific consensus (MWTTTNC) for FPR was also developed by looking at the frequency of the bases at each degenerate position within the High confidence set.

The search for overrepresented motifs also identified a new seven base putative core promoter element, SAASMMS, that was named DPEpyth. This is present in one-third of the genes in the region from 11 to 40 bases downstream of the putative TSSs in both High Confidence and Expanded sets (Fig 3), and is overrepresented within the first 50 bases in case of the total set (Fig 6C). In the promoters where this element is detected Inr is present almost half (~43%) of the time; in few cases (~8%) where no Inr could be found, but FPR was present. FPR and Inr both are present in ~13.5% of the cases and in ~35% DPEpyth is present by itself (Fig 5A). In genes where this element is present either with Inr or FPR or both, it is present downstream of both Inr (26 nt on an average) and FPR (17 bases on an average) and therefore downstream of the TSSs of those genes (Fig 4A). The position and conservation of these three elements on a gene Adenosylhomocysteinase (PITG_10198) and its orthologs in *P. sojae* and *P. ramorum* shows that the DPEpyth is not only conserved but, P.sojae has two copies of this element right next to each other, the alignment also supports our

definitions of Inr and shows that the FPR in *P. infestans* would not be detected with the McLeod definition (Fig 4B).

Meme detected another motif, GARGMR, that was overrepresented in both the High-confidence and the expanded sets. This was not regarded as a core promoter element after closer examination indicated that it was usually found to be very close to the ATG, both upstream and downstream of it. This might be a motif that plays a role in translation rather than transcription.

Fig 3 legend:

Distribution of DPEpyth in the *P. infestans* gene core promoter regions:

Shown is the distribution of the novel *P. infestans core* promoter element DPEpyth within the High Confidence (bars) and the Expanded (lines) sets. The logo is derived from the High Confidence Set results**.**

Fig 3

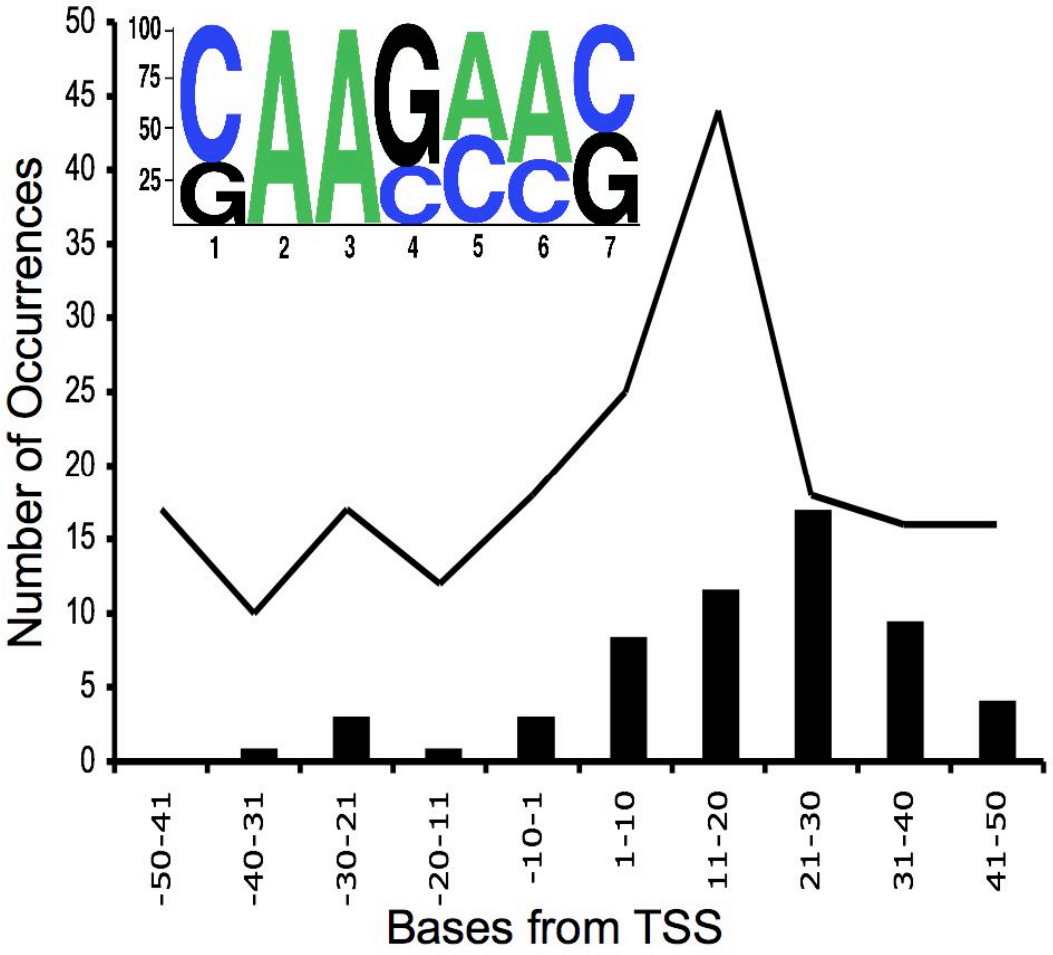

Number of Occurrences
Bases from TSS
-50-41
-40-31
-30-21
-20-11
-10-1
1-10
11-20
21-30
31-40
41-50
0
5
10
15
20
25
30
35
40
45
50
100
75
50
25
1
2
3
4
5
6
7

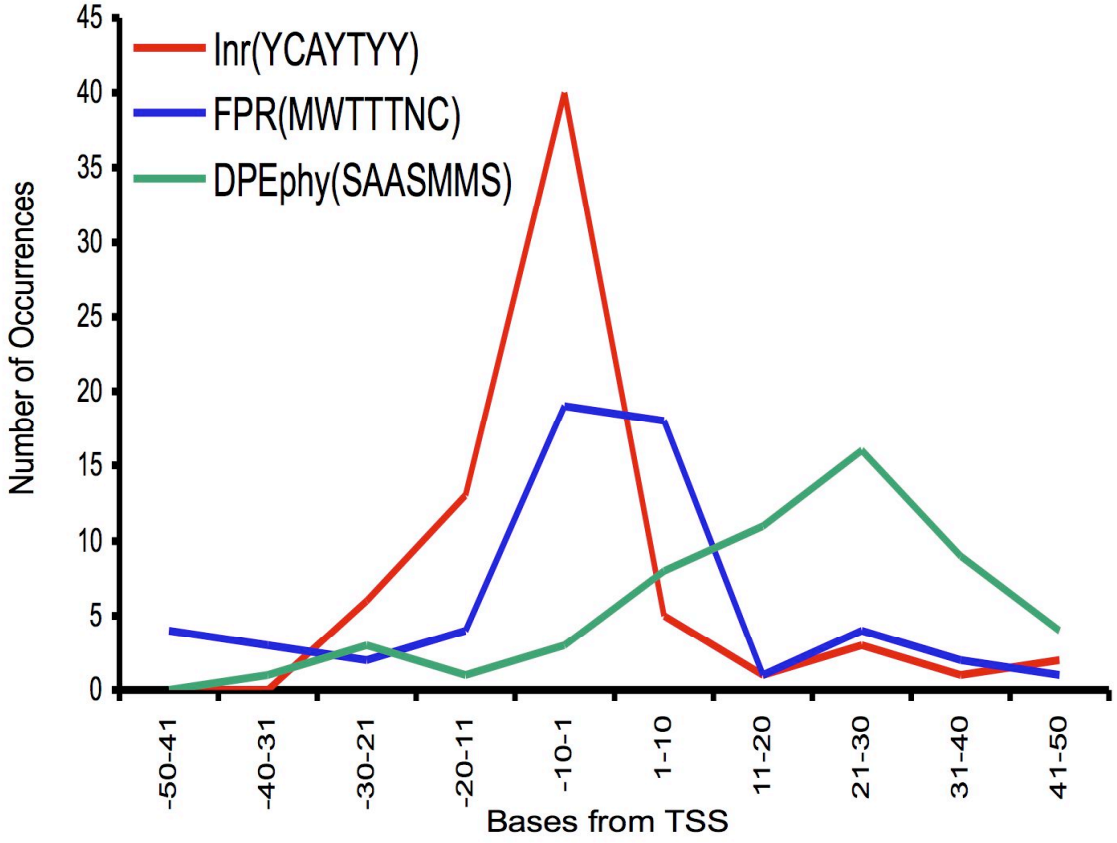
Fig 4 legend:

FPR is present downstream of Inr and DPEpyth is present downstream of both Inr and FPR.

A) Shown is the positional bias of the three *P. infestans* core promoter elements within the core promoter regions of the High Confidence Set. B) Shown is the conservation of the three core promoter elements in the *P. infestans* gene encoding Adenosylhomocysteinase (PITG_10198) and in its orthologs in *P. sojae* and *P. ramorum.* It also shows the relative distance between each of these elements and that from the translation start site (ATG).

Fig 4

A



B

```
P.inf   ACAGGCCACTTTGGAATTTGCGTCCACAAGCCCAGCCACC
P.soj   TCGGCTCACTTTGGAATTCGCGTCTAGAAGCCCAAGCACC
P.ram   ACTGCTCACTTTGGAATTCGCTCTTCAGTAGCCCAAGCAC
Cons          YCAYTYY   MWTTTNC          SAASMMS
                Inr       FPR             DPEpyth

P.inf   TTTCACCCCCCCCTACTT---TCAACAATG
P.soj   GCCTTCACCCCCTCTCTTCTTCCAACAATG
P.ram   CTTTTCACCCCCC--AGTCAACTAACCATG
```

Correlation between core promoter elements and gene expression:

To check the effects of the three core promoter elements (CPE), two analyses were performed using High Confidence Set genes. These genes were divided into seven different sets based on the presence of different core promoter elements, for the analyses. There were 46 genes out of 121, for which there was no core promoter element detected in the regions where each of the three elements are overrepresented; these made the 'No CPE' (Fig. 5b) set. The Inr, FPR and the DPEpyth sets had 19, 8 and 13 genes respectively (Fig. 5a). The set with Inr and FPR had 11 genes, that with Inr and DPEpyth had 16 and the one with FPR and DPEpyth had only three genes (Fig. 5a). Five genes had all three elements (Fig. 5a).

Expression profiles of the genes in the seven sets described above were extracted from previously reported microarray data (Judelson et al. 2008). The maximum expression of the genes in any of the five asexual developmental stages (hyphae, sporangia, cleaving sporangia, swimming zoospore and germinating cysts) was checked to see if the presence of any element affected expression levels. Expression data was not available for two genes each in the 'Inr' and 'Inr+DPEpyth' sets. For all other sets, data were unavailable for one gene each, except for that of 'FPR + DPEpyth', where data for all genes were available. There were no data for eight genes that belong to the 'No CPE' set.

It was observed that the median of the maximum expression for genes with just Inr (615.4) was more than 2-fold lower than that of the 'No CPE' set

(1355.8). But, the median for the gene sets with FPR, either on its own (2775.5) or with Inr (2381.2) or DPEpyth (2921.5), was much higher than that of the 'No CPE' set (Fig 5b). The median for the 'DPEpyth' (1229.6), ,and 'No CPE' sets were similar. That of 'ALL 3 CPE set'(1609.3) and 'Inr + DPEpyth' (967.5) were slightly higher and lower than the 'No CPE set' respectively. This showed that FPR probably has the greatest effect on the expression of the genes. It should be mentioned that even though there were only 3 genes in the 'FPR + DPEpyth' set, the expression pattern of genes with these two elements, were checked within the expanded set and the results were comparable. It should also be mentioned that the range for the maximum expressions was pretty broad, and the patterns that were observed might change with the increase in the number of genes, even though the patterns held true for the groups (with fewer genes) within the Expanded Set, that were checked.

Whether the presence of a certain element was associated with constitutive or developmentally-regulated expression was also examined. To analyze this, a per-gene normalized data was used to look at the maximum fold-change for each gene within the five different developmental stages in all the sets. The median for each set was checked after that. About two-thirds of the genes with only Inr or FPR showed a maximum fold-change of more than 10 fold (Fig 5b); whereas only one third of those with DPEpyth showed such a difference. But, it was highly interesting to find that many more genes showed a maximum fold change of greater than 10 when there was FPR with Inr (~90%),

or a DPEpyth with Inr (~80%) (Fig 5b). This suggested that the presence of any one of the two elements, downstream of the Inr, makes the genes much more likely to show more extreme changes than with Inr alone or with none of the three elements. The numbers of genes in the ‘All 3 CPE’ and the ‘FPR + DPEpyth’ sets were too low to draw any firm conclusions.

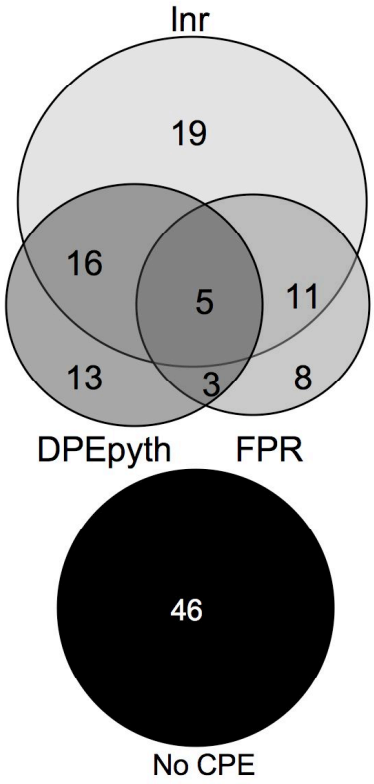
Fig 5 legend:

Core promoter elements within the High Confidence Set and their effect in gene expression patterns:

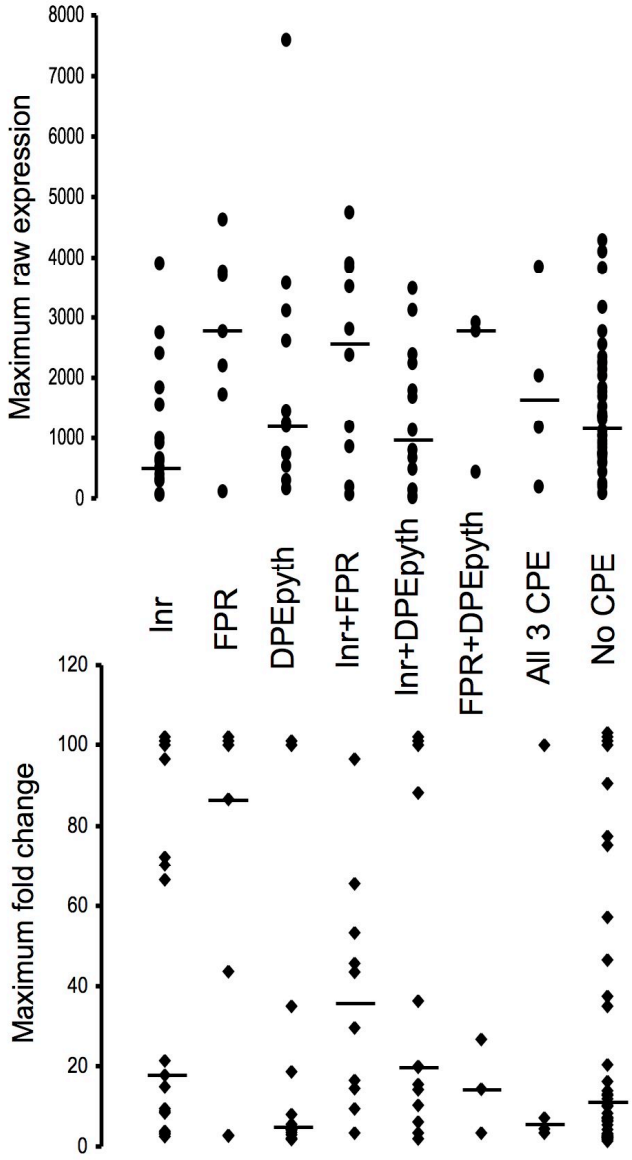
A) The Venn diagram shows the number of genes within the High Confidence Set with different combinations of the three core promoter elements. B) The upper panel shows the maximum expression of each gene (circle) within the High Confidence Set divided into different groups depending on the core promoter elements present in those genes. The horizontal bars represent the median for each group. The lower panel shows the maximum fold change in expression for a gene (diamond) within the five different stages of the asexual cycle. The bars represent the median for each group.

Fig 5

A


Inr
19
16
5
11
13
3
8
DPEpyth
FPR
46
No CPE

B


Maximum raw expression
8000
7000
6000
5000
4000
3000
2000
1000
0
Maximum fold change
120
100
80
60
40
20
0
Inr
FPR
DPEpyth
Inr+FPR
Inr+DPEpyth
FPR+DPEpyth
All 3 CPE
No CPE

Distribution of Inr, FPR and DPEpyth in heterokonts:

I decided to look for the three *P. infestans* elements in other heterokonts and compare their frequency and distribution with that of *P. infestans.* The average distance between the ATG and the TSS (the 5' UTR) in *P. infestans* is considered to be 41 bases (Win et al., 2006). Therefore, I decided to look for these elements within 200 bases upstream of ATG, which should include the core promoter region i.e. 50 bases on either side of the TSSs. Hence, 200 bases upstream of ATG of all predicted genes for each of the eight heterokont species were extracted and searched for Inr, PFR, and DPEpyth. The GC contents of 1-kb upstream sequences of each of the said genomes were calculated and taken into account while calculating the expected frequency of the motifs.

It was observed that the overrepresentation of Inr in all Phytophthora species viz. *P. infestans, P. sojae and P . ramorum* (Fig 6a) and all other heterokonts except for the diatom *P. tricornutum* and the brown alga *E. siliculosus,* using a p-value cut-off of 95%. The absence of Inr from the latter species may be due to the inaccuracy of their gene models, or the stringency of the p-value cut-off for significance in an equality of proportions test. In fact, the observed frequencies for Inr in *P. tricornutum* within the first 100 bases are much higher than the background (p = 0.075 for 51-100 bases and 0.12 for 1-50 bases). Also the fact that, the observed values for Inr in *T. pseudonana*, the other diatom checked is not significantly higher (p = 0.05 for 51-100 bases and 0.06 for

1-50 bases) than that of *P. tricornutum* makes us believe that Inr is actually overrepresented within 51-100 bases in *P. tricornutum* too (Table 3).

Table 3:

Distribution of P. infestans core promoter elements in other Heterokont species:

Occurrence of the three *P. infestans* core promoter elements within the first 100 bases upstream of ATG in other species of the phylum Heterokontophyta is shown above. The observed values per hundred genes is shown. The p-value is derived from the z-score obtained by checking the equality of proportions between the observed and the expected values. The p-values shown in bold were found to be significant at 95% confidence interval.

Table 3: Distribution of core promoter elements in Heterokonts

Species	Bases from ATG	Inr per 100 genes	p-value	FPR per 100 genes	p-value	DPEphy per 100 genes	p-value
P. infestans	1-50	15.59	**0.001**	11.23	0.069	17.24	**0.039**
	51-100	15.30	**0.001**	12.15	**0.049**	12.44	0.191
P. sojae	1-50	16.83	**0.003**	9.01	0.068	21.68	**0.024**
	51-100	16.40	**0.004**	9.50	**0.050**	14.82	0.192
P. ramorum	1-50	17.86	**0.002**	10.31	**0.042**	21.78	**0.022**
	51-100	18.89	**0.001**	10.62	**0.036**	13.85	0.242
Py. ultimum	1-50	11.25	**0.050**	9.44	0.108	24.19	**0.007**
	51-100	16.74	**0.005**	12.44	**0.032**	16.47	0.102
H. arabidopsidis	1-50	11.21	0.063	7.52	0.272	13.02	0.225
	51-100	13.06	**0.030**	8.70	0.186	12.70	0.244
S. parasitica	1-50	21.13	**0.000**	20.41	**0.002**	16.55	0.128
	51-100	5.96	0.292	6.48	0.148	11.03	0.467
T. pseudomona	1-50	11.36	0.063	4.88	0.331	29.65	**0.001**
	51-100	11.97	**0.049**	7.03	0.364	20.48	**0.018**
Ph. tricornutum	1-50	9.62	0.120	10.40	0.118	13.11	0.822
	51-100	10.86	0.075	11.44	1.410	13.41	0.881
E. siliculosus	1-50	3.72	0.396	3.50	0.456	19.24	0.063
	51-100	5.41	0.383	5.25	0.315	18.63	0.052

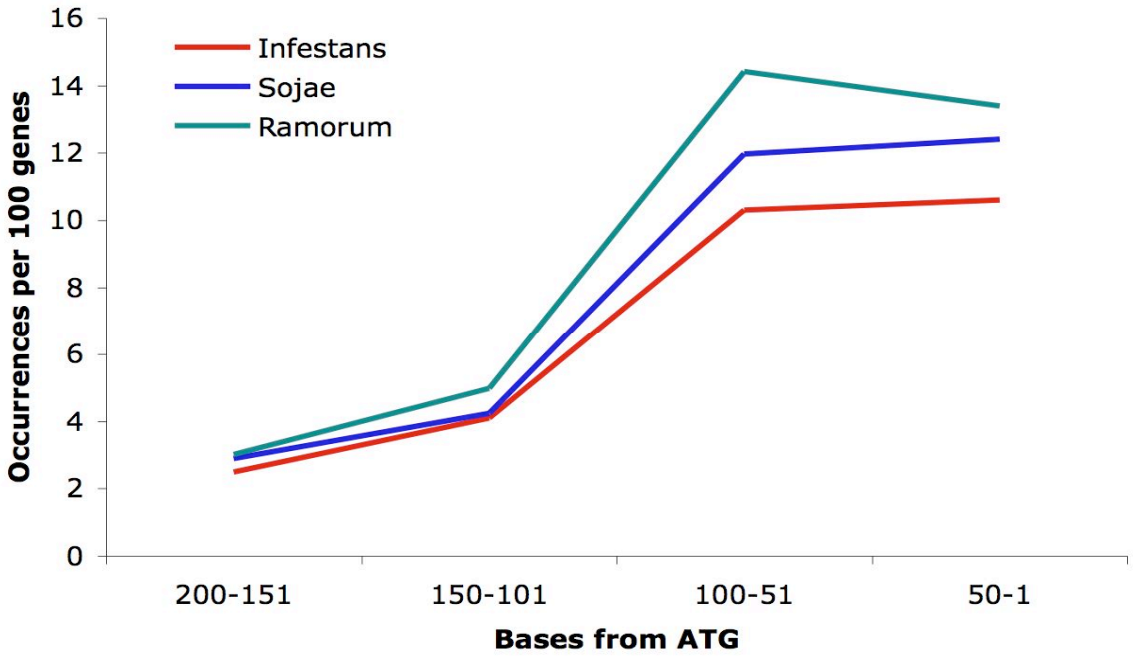
Fig 6 legend:

Distribution of *P. infestans* core promoter elements in the two other Phytophthora species:

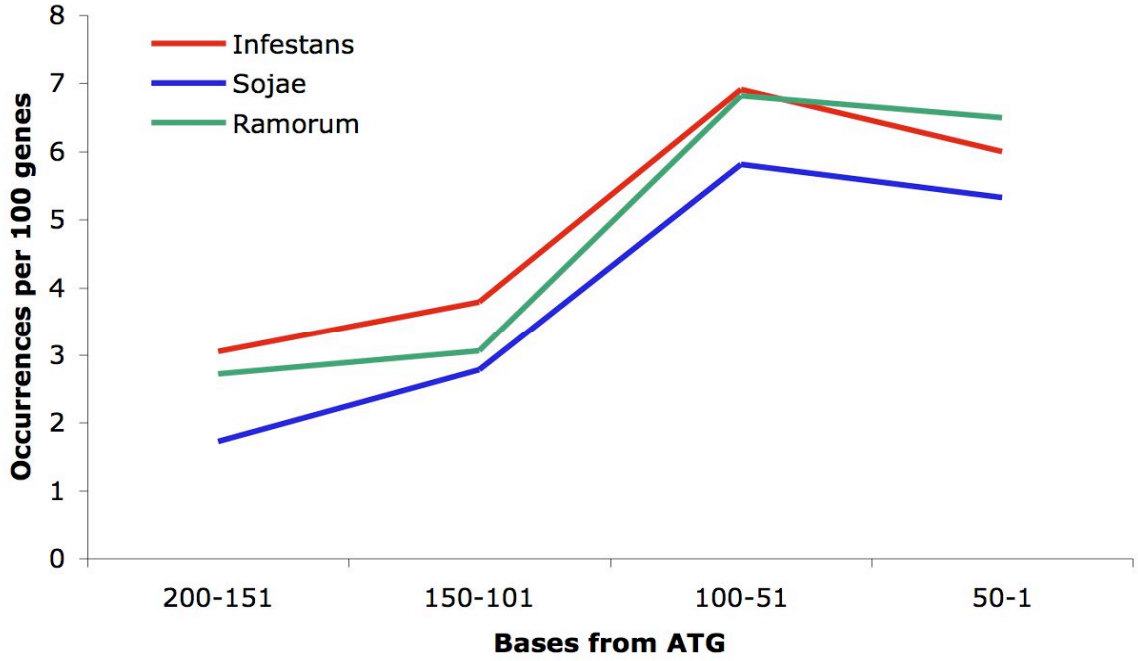
A) Shown is the distribution of Inr within the first 200 bases of all genes (Total Set) and the bias for the first 100 bases in all three Phytophthora species : *P. infestans, P. sojae* and *P. ramorum.* B) Shows the distribution of FPR within the first 200 bases of all genes (Total Set) and the bias for the first 100 bases in all three *Phytophthora* species. C) Shows the distribution of DPEpyth within the first 200 bases of all genes (Total Set) and the bias for the first 50 bases in all three *Phytophthora* species.

Fig 6

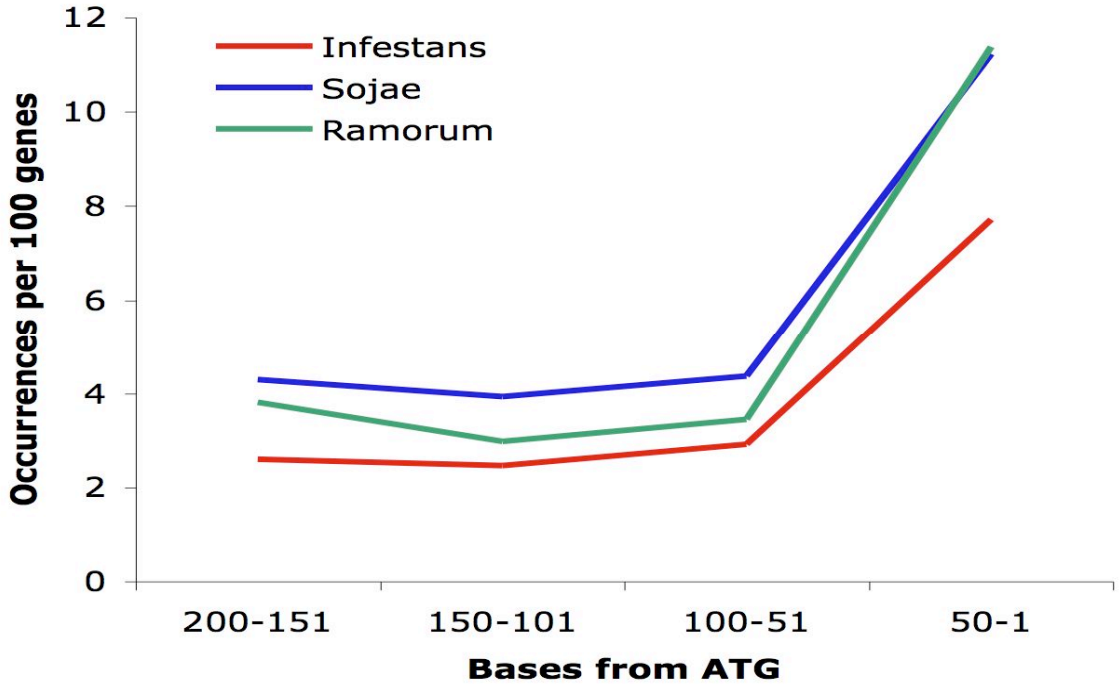
A


16
14
12
10
8
6
4
2
0
Occurrences per 100 genes
Infestans
Sojae
Ramorum
200-151
150-101
100-51
50-1
Bases from ATG

B


8
7
6
5
4
3
2
1
0
Occurrences per 100 genes
Infestans
Sojae
Ramorum
200-151
150-101
100-51
50-1
Bases from ATG

C


Occurrences per 100 genes
12
10
8
6
4
2
0
Infestans
Sojae
Ramorum
200-151
150-101
100-51
50-1
Bases from ATG

Overrepresentation of FPR was detected not only in all three Phytophthora species (Fig 6b), but also within *Py. ultimum* and *S. parasitica.* However, no overrepresentation for this motif could be detected in *H. arabidopsidis*. DPEpyth is found to be overrepresented within the first 50 base pairs only in the species that belong to the order Pythiales i.e. the Phytophthora species (Fig 6c) and *Py. ultimatum.* The only other heterokont where an overrepresentation for DPEpyth could be detected was one of the two diatoms, *T. pseudonana.*

Other known core promoter elements in *P. infestans*: The searches for over-represented motifs described above did not reveal any sequences resembling core promoter elements described in other eukaryotes besides the Inr. These include the widely distributed TATA-box and others such as the BREu, BREd, DRE, and Y-patch that have been reported in a more narrow range of species (Smale and Kadonaga, 2003 ; Civan and Svec, 2009).

To test more directly whether such elements might exist as functional motifs in *P. infestans,* I tested whether they are over-represented. No TATA-like sequence appeared over-represented in the High Confidence, Expanded and Total sets (Table 4). When the sequence was detected (possibly by random chance), no positional bias was observed within the 200 base pairs upstream of all *P. infestans* genes (Total set). No peaks were observed in the region where the TATA-box is usually detected, i.e. approximately 26 to 30 bases upstream of

the TSS in all organisms studied, other than *S. cerevisiae*, (Ohler and Wassarman, 2010; Jin et al., 2006; Ohler, 2006), when I searched the High Confidence and the Expanded sets. In *S. cerevisiae* it appears in a wider region 40 to 120 bases upstream of the TSS, and this region, for most genes was included within the total set where no peaks were detected. Different variations of TATA-box detected in previous studies were looked for in all three different sets of genes for *P.infestans* but no overrepresentation in any particular region could be detected. Other elements that are found upstream of the TSS in some organisms (BREu, BREd and DRE), were also not detected more often than they are expected by random chance (Table 4). A similar result was obtained for elements normally found downstream of TSS like MTE and DPE (Table 4). Variations and degenerate definitions of these core promoter elements, along with that of plant core promoter elements such as Y-patch and Arabidopsis motif 5 and motif 7, were checked but overrepresentation for none of these could be detected. For example, the plant core promoter element Y-patch was checked using a considerably degenerate definition (CYTCYYYCCYC) (Civan and Svec, 2009), but not a single occurrence could be detected in the High Confidence set. In the Total Set only 1% of the genes had this element (Table 4) which is much less than one would expect by random chance. Similarly, the most degenerate versions of two Arabidopsis core promoter elements, motif 5 and motif 7 (Molina and Grotewold, 2005) were found to be present in only 0.06% and 0.1% of the genes respectively when the Total Set was searched.

Table 4:

Distribution of Inr, FPR, DPEpyth and other well-known eukaryotic core promoter elements in *P. infestans*:

The table above shows the percentage of P. infestans genes that carry the well-known core promoter elements. It also shows if these elements have a bias towards any particular position, within the first 200 bases upstream of the translation start sites, of the genes where these are present.

Table 4: Distribution of well known eukaryotic core promoter elements in *P. infestans*

Motif name	Motif definition	Expected location relative to TSS	Transcription Factor	Occurrence in % of genes in P. infestans	Positional Bias
Inr	YCAYTYY	-2	TAF1, TAF2	~42%	Yes
FPR	MWTTTNC	8	unknown	~22%	Yes
TATA-box	TATAWAWR	-31	TBP	~2.5%	No
DPE	RGWCGTG	30	TAF6, TAF9	~4.5%	No
DPEphy	SAASMMS	+11 to +40	unknown	~31%	Yes
MTE	CSARCSSAACGS	+20 to +30	TFIID	~.05%	No
BRE	SSRCGCC	-39	TFIIB	~8.5%	No
DRE	WATCGATW	Upstream (variable)	DREF-TRF2 complex	~1.5%	No
Y-patch	CYTCYYCCYC	Variable	unknown	~1%	No
Motif 5 Arabidopsis	AAACCCHARN	Variable	Myb related telomeric DBP	~0.6%	No
Motif 7 Arabidopsis	ARGCCAW	Variable	unknown	~1%	No

DISCUSSION:

Early core promoter studies established the idea of the TATA-box being universal, but the views regarding core-promoter organization in eukaryotes have changed in recent times. The TATA-box is no longer considered to be a general polymerase II core-promoter feature (Gross and Oelgeschlager, 2006). Other core promoter elements like Inr, DPE, MTE, etc. have also been identified over the years, but have also proved to not be universal. Nevertheless, some of these motifs may function through similar mechanisms. TATA, TATA-Inr, Inr and Inr-DPE in all likelihood initially evolved as functionally equivalent recognition sites for TFIID subunits and their evolutionary precursors (Smale and Kadonaga, 2003). It has been suggested that these may continue to function as interchangeable TFIID recognition sites in some promoters (Smale and Kadonaga, 2003). Most of the *P. infestans* promoters seem to be in the Inr class as very few TATA or DPE motifs could be detected.

To the best of our knowledge this is the first comprehensive *in silico* genome-wide analysis of core promoter elements in any oomycete or heterokont. The two main challenges regarding computational promoter analyses are: TSS prediction which is required for accurate localization of the core promoter, and the discovery of motif within that search space. To address the first issue, I have considered EST data related to each gene to locate the TSS. The somewhat broad (over a 20 base window) distribution for the different core promoter elements, instead of specific positions with respect to the TSS, is probably due to

the lack of knowledge for the exact TSSs for the genes. To counter the second challenge I have followed some of the guidelines suggested by a prior study (Juven-Gershon et al., 2006). This includes checking whether the putative core promoter elements are in several different gene promoters, and if the motifs are at a specific position (positional bias) with reference to the putative TSS. For elements like Inr, which is a known binding site for subunits of TFIID (Smale and Kadonaga, 2003), and FPR which is specific to oomycete promoters, these were detected in increased number of genes by checking for certain variations which were obtained from MEME results. As suggested previously by Mcleod et al. (2004) it was found that the FPR is situated exactly eight bases downstream of the Inr, but only in about one-third of the genes that the Inr. Conservation of distances between core promoter elements has been observed in the case of Inr and DPE in Drosophila (Burke and Kadonaga, 1997). Inr is found in two times more genes than FPR in *P. infestans,* which implies that FPR is not necessary for Inr function although there could be a combinatorial effect. This is evident from the analysis of expression data where most the genes with Inr and FPR have higher maximum expression than the ones with only Inr. This is similar to the effect that TATA-box has in Drosophila genes with Inr, where the expression is higher than that of the TATA-less genes (Burke et al., 1998). There are some *P. infestans* genes where FPR is found without an upstream Inr, which suggests that FPR might be able to initiate transcription independently by binding to an element within the RNA polymerase II complex. FPR has a base composition

very similar to that of Inr which might help it bind. The expression data are consistent with this hypothesis, since FPR was correlated with higher levels of expression. The definitions for Inr and FPR include a lot of degeneracy, which might suggest that the elements within the basal transcription machinery that bind to these elements are flexible in their choice of binding sites. This hypothesis is supported by the findings of McLeod et al. (2004), which shows that a guanine at the fourth position of Inr (YCAYTYY) instead of a cytosine or thiamine works fine. Similarly, a guanine at the eighth position of the McLeod-defined FPR (CAWTTTNYY) is able to initiate transcription, even though a cytosine was detected at that position, almost always.

Overrepresentation of none of the other well-known eukaryotic core promoter elements could be detected. Lack of the TATA-box in *P. infestans* is not particularly surprising, due to the abundance of Inr, which is more commonly found in TATA-less genes. Inr, like the TATA-box, is the recognition site for the multi-subunit TFIID complex (Smale and Kadonaga, 2003), which contains the TATA-binding protein (TBP) and several TBP-associated factors (TAFs; Burke et al., 1998; Smale et al., 1998). Therefore, it is probably somewhat redundant for an organism to have both, even though classes of genes with both an Inr and TATA-box have been identified in *Drosophila* and some mammals. There is considerable evidence that the two subunits of TFIID complex viz. TAF2 and TAF1 interact with the Inr (Verrijzer et al., 1994, 1995; Kaufmann et al., 1998; Chalkley and Verrijzer, 1999) in a sequence-specific manner (Kaufmann and

Smale, 1994; Martinez et al., 1994; Purnell et al., 1994, Burke and Kadonaga, 1996; Oelgeschlager et al., 1996). It has also been observed that purified RNA polymerase II recognizes Inr and mediates transcription in the absence of TAFs (Carcamo et al., 1991; Weis and Reinberg, 1997), suggesting that Inr is required for different steps in the process of transcription when it interacts with TFIID and RNA polymerase II (Butler and Kadonaga, 2002).

DPE (Butler and Kadonaga, 2002; Juven-Gershon et al., 2006) and MTE (Juven-Gershon et al., 2006) are mostly seen in non-oomycete TATA-less promoters. Both DPE and MTE are believed to act co-operatively with Inr (Juven-Gershon et al., 2006) as without the Inr neither exhibit core promoter activity. Therefore, it might be expected that these elements would be overrepresented in *P. infestans* core promoters, which is not the case. This is probably due to the fact that most of these consensus sequences are derived from studies in vertebrates and insects. A previous study (Judelson et al., 1992), which tested promoter sequences from a range of species for their activities in oomycetes, has shown that non-oomycete promoter elements do not work in oomycetes. It has been suggested that in oomycetes the DNA binding specificity of key elements of the transcription machinery are either different from their orthologs in higher fungi and eukaryotes or these proteins may be not be present (Judelson et al., 1992).

A new putative core promoter element, DPEpyth, which is a seven base element found downstream of the TSS, was detected. Although mostly detected in promoters where there was either Inr or FPR or both, there were quite a few

exceptions suggesting that DPEpyth might be able to function independently of the other motifs. The results from the expression analyses suggest that DPEpyth on its own may not have a lot of impact on the level of expression of a gene. DPEpyth is found to be overrepresented within the first 50 bases upstream of the ATG among all the Phytophthora species and in *Py. ultimatum,* which is very close to Phytophthora in terms of phylogenetic distance suggesting that this may be specific to the order Pythiales. However, its overrepresentation in one of the two diatoms suggests that its distribution might be broader. The reason behind the absence of DPEpyth in the other diatom is probably due to the phylogenetic distance between the two diatoms (McDonald et. al., 2010) studied. Therefore, further study for this element might be interesting and might provide some clue sabout the evolution of the transcription machinery in the different organisms. DPEpyth is always found downstream of the Inr and FPR and is fairly close to the translation start, therefore the possibility that it might have some role in translation can not be excluded. But, the fact that it is found at a region where another well-characterized core promoter element, DPE, is found in Drosophila, tells us that it most likely is involved in transcription. Also, like DPE, DPEpyth in most cases is found either with Inr or with FPR (that has very high sequence similarity with Inr) which strengthens this belief.

To conclude, this study shows us that some core promoter elements like FPR might be specific to a very small class like oomycetes. Also, there might be a species-specific or group-specific consensus for other well characterized

elements like Inr and that is probably the reason why the consensus for Inr in Drosophila differs from that of Inr in mammals. It also shows that an entirely different sequence (like DPEpyth) can be detected at a region where another known core promoter element (DPE) is found in other organisms. Therefore, searching for a consensus identified in another organism may not be the optimal approach for detecting core promoter elements in a phylogenetically distant organism. A better approach is looking for different variations and also searching for new elements.

REFERENCES:

1. Ohler U, Wassarman DA (2010) Promoting developmental transcription. Development 137: 15–26
2. Müller F, Demény MA, Tora L (2007) New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. J. Biol. Chem. 282: 14685–14689
3. Butler JEF, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. 16: 2583-2592
4. Struhl K (1987) Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast. Cell 49: 295–297
5. Weis L, Reinberg D (1992) Transcription by RNA polymerase II: Initiator-directed formation of transcription-competent complexes. FASEB J. 6: 3300–3309
6. Smale ST (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim. Biophys. Acta. 1351: 73–88
7. Smale ST (2001) Core promoters: Active contributors to combinatorial gene regulation. Genes Dev. 15: 2503–2508
8. Smale ST, Jain A, Kaufmann J, Emami KH, Lo K et al. (1998) The initiator element: A paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harb. Symp. Quant. Biol. 58: 21–31

9. Burke TW, Willy PJ, Kutach AK, Butler JEF, Kadonaga JT (1998) The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. Cold Spring Harb. Symp. Quant. Biol. 63: 75–82
10. McLeod A, Smart CD, Fry WE (2004) Core Promoter Structure in the Oomycete *Phytophthora infestans*. Eukaryotic Cell 3: 91-99
11. Burke TW, Kadonaga JT (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by $TAF_{II}60$ of *Drosophila.* Genes Dev**.** 11: 3020–3031
12. Juven-Gershon T, Hsu JY, Theisen JWM, Kadonaga JT (2008) The RNA polymerase II core promoter – the gateway to transcription. Curr. Op. Cell Biol. 20: 1–7
13. Hochheimer A, Tjian R (2003) Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev. 17: 1309–1320
14. Woychik NA, Hampsey M (2002) The RNA polymerase II machinery: structure illuminates function. Cell 108: 453–463
15. Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol. Mol. Biol. Rev 62: 465–503
16. Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. Annu. Rev. Biochem. 72: 449–479
17. Levine M, Tjian R (2003). Transcription regulation and animal diversity. Nature 424: 147–151

18. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE et al. (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461: 393-398
19. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Sec. Int'l. Conf. Intell. Sys. Mol. Biol. pp. 28-36
20. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680
21. Judelson HS, Ah-Fong AMV, Aux G, Avrova AO, Bruce C et al. (2008) Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. Mol. Plant-Micro. Inter. 21: 433–447
22. Zhang Z, Dietrich FS (2005) Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. Nucleic Acids Res. 33: 2838–2851
23. Zhu Q, Dabi T, Lamb C (1995) TATA box and initiator functions in the accurate transcription of a plant minimal promoter in vitro. The Plant Cell 7: 1681-1689

24. Pieterse CM, van West P, Verbakel HM, Brassé PW, van den Berg-Velthuis GC et al. (1994) Structure and genomic organization of the ipiB and ipiO gene clusters of *Phytophthora infestans*. Gene 138: 67-77
25. Parry TJ, Theisen JWM, Hsu JY, Wang YL, Corcoran DL (2010) The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. Genes Dev. 24: 2013–2018
26. Win J, Kanneganti TD, Torto-Alalibo T, Kamoun S (1996) Computational and comparative analyses of 150 full-length cDNA sequences from the oomycete plant pathogen *Phytophthora infestans*. Fung. Gen. Biol. 43: 20-33
27. Jin VX, Singer GAC, Agosto-Pérez FJ, Liyanarachchi S, Davuluri RV (2006) Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. BMC Bioinf. 7: 114
28. Ohler U (2006) Identification of core promoter modules in *Drosophila* and their application in accurate transcription start prediction. Nucleic Acids Res. 34: 5943–5950
29. Civan P, Svec M (2009) Genome-wide analysis of rice (*Oryza sativa* L. subsp. *japonica*) TATA box and Y patch promoter elements. Genome 52: 294–297
30. Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. BMC Genomics 6: 25
31. Gross P, Oelgeschlager T (2006) Core promoter-selective RNA polymerase II transcription. Biochem. Soc. Symp. 73**:** 225-236

32. Juven-Gershon T, Hsu JY, Kadonaga JT (2006) Perspectives on RNA polymerase II core promoter. Biochem. Soc. Transactions 34: 1051-1054
33. Verrijzer CP, Yokomori K, Chen JL, Tjian R (1994) *Drosophila* $TAF_{II}150$: Similarity to yeast gene TSM-1 and specific binding to core promoter DNA. Science 264: 933–941
34. Verrijzer CP, Chen JL, Yokomori K and Tjian R (1995) Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. Cell 81: 1115–1125
35. Kaufmann J, Ahrens K, Koop R, Smale ST and Müller R (1998) CIF150, a human cofactor for TFIID-dependent initiator function. Mol. Cell Biol. 18: 233–239
36. Chalkley GE, Verrijzer CP (1999) DNA binding site selection by RNA polymerase II TAFs: A TAFII250–TAFII150 complex recognizes the Initiator. EMBO J. 18: 4835–4845
37. Kaufmann J, Smale S (1994) Direct recognition of initiator elements by a component of the transcription factor IID complex. Genes Dev. 8: 821–829
38. Martinez E, Chiang CM, Ge H and Roeder RG (1994) TAFs in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. EMBO J. 13: 3115–3126
39. Purnell BA, Emanuel PA, Gilmour DS (1994) TFIID sequence recognition of the initiator and sequences farther downstream in *Drosophila* class II genes. Genes Dev. 8: 830–842

40. Burke TW, Kadanoga JT (1996) *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 10: 711–724
41. Oelgeschläger T, Chiang CM, Roeder RG (1996) Topology and reorganization of a human TFIID–promoter complex. Nature 382: 735–738
42. Carcamo J, Buckbinder L, Reinberg D (1991) The initiator directs the assembly of a transcription factor IID-dependent transcription complex. Proc. Natl. Acad. Sci. USA 88: 8052–8056
43. Weis L, Reinberg D (1997) Accurate positioning of RNA polymerase II on a natural TATA-less promoter is independent of TATA-binding protein associated factors and initiator-binding proteins. Mol. Cell Biol. 17: 2973–2984
44. Judelson HS, Tyler BM, Michelmore RW (1992) Regulatory sequences for expressing genes in oomycetes. Mol. Gen. Genetics 234: 138-146
45. McDonald SM, Plant JN, Worden AZ (2010) The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: A case study of *Micromonas*. Mol. Biol. Evol. 27: 2268-2283

CONCLUSION

Identification of regulatory elements, especially that of transcription factor binding sites (TFBSs), has been one of the most intriguing problems related to gene regulation that bioinformaticians have tried to solve (Li and Tompa, 2006). Multiple algorithms have been developed over the years for detection of overrepresented sequences within promoter datasets, but each of these have their own share of pros and cons (Tompa et al., 2005). One major problem is that these short degenerate sequences do not carry much information on their own. To tackle the issues related to the problem of TFBS identification, and to increase the reliability of the predictions, scientists have tried different approaches over the years. Some of these approaches such as looking at the positional bias of a short string of DNA sequence known as 'motif' (Bellora et al, 2007; Tharakaraman et al., 2008), considering phylogenetic information (Cliften et al. 2003; Hong et al., 2003; McCue et al., 2002, Dermitzakis et al., 2002; Guo and Moose, 2003; Bowser and Tobe, 2007), or looking at the chromatin structure (Whitington et al., 2009), are very interesting. High-throughput molecular techniques like Chromatin immunoprecipitation (ChIP; Kaufman et al., 2010), hybridization to microarrays (ChIP-chip; Chen et al., 2010) and direct sequencing (ChIP-sequencing, Raha et al, 2010) that look at protein DNA interactions, have also been used for identification of TFBSs. The cost involved with the high-throughput molecular techniques is high, therefore, in recent years the focus has

been on combining bioinformatics and molecular techniques for prediction and validation of TFBSs (Vallania et al. 2008).

In this study I have presented a method (Chapter I) that combines data from regulatory genomics, positional regulomics and comparative genomics, for the robust prediction of candidate TFBSs. The validation of functionality of the predicted candidates is then done with data from relatively fast functional genomics and protein-DNA binding affinity experiments. It was shown that the method is fast and robust in terms of predictions. In terms of validation, it is inexpensive and not very labor intensive, when compared to other molecular techniques. With the cost of sequencing going down and the number of sequenced genomes going up by the day, this method can be applied in case of any organism to great effect. This method was used to identify the proximal and core promoter elements present in *Phytophthora infestans.*

Even if its historical importance is set aside, its economical importance makes *Phytophthora infestans* an organism worth studying. It remains a critical threat to world food security and causes a loss of ~ 6.7 billion dollars (Haas et al., 2009) by infecting potato, the world's largest non-cereal foodcrop. *P*. *infestans* genome was sequenced in 2009 (Haas et al., 2009), this opened the doors for genomics and bioinformatics studies related to this organism, which has the largest and most complex genome among the chromalveolates sequenced to date (Haas et al. 2009). *P. infestans* genome (~240 Mb) is much larger when compared to that of the other related *Phytophthora* species (95 Mb

in *P. sojae* and 65 Mb in *P. ramorum)*, but there is not much difference in the number of protein coding genes (17797 in *P. infestans* , 16988 in *P. sojae* and 14451 *P. ramorum;* Haas et al. 2009), makes *P. infestans* an interesting organism to study from a bioinformatician's point of view as well.

Phytophthora exhibits fungus-like growth that involves the formation of spores on the termini of specialized hyphae called sporangiophore (Judelson and Blanco, 2005). In fact, most oomycetes look superficially like fungi due to the filamentous thread-like mycelia. This coupled with absorption being the common nutritional mode, were the main reasons why they were once classified with true fungi like yeast and *Neurospora*. However, there are several differences between true fungi and oomycetes. One of the main differences is diploidy in oomycetes, unlike in true fungi, which impairs the use of mutagenesis while studying development. Unlike in other diploid organisms, while doing reverse genetics, one has to rely on gene silencing rather than knock-outs since homologous recombination of transgenes is extremely rare in *P. infestans*. This is one of the main reasons behind limited research with this group.

This research, to the best of our knowledge, is the first systematic study of promoter structure in any oomycete. I have studied the core elements within the promoters of *P.infestans* genes and the proximal elements within the promoters of genes upregulated in the five key asexual stages. A previous study (Judelson et al. 1992), which tested promoter sequences from a range of species for their activities in oomycetes, has shown that non-oomycete promoter elements do not

work in oomycetes. The fact that non-oomycete promoter elements do not work in oomycetes makes the study even more relevant.

Our study of the core promoter regions (Chapter III) revealed the presence of a seven base putative core promoter element (DPEpyth), in *Phythophthora* and *Pythium*, which has not been reported thus far. The elements that have been detected within oomycete core promoters to date are Inr (YCATTYY; McLeod et. al, 2004) and FPR (CAWTTTNYY; McLeod et. al, 2004). These elements are very similar to each other if sequence similarity and their position within the core promoter are considered, suggesting that they might be binding either to the same or very similar factors within the transcription machinery. The DPEpyth motif (SAASMMS), on the other hand, is not only different from Inr and FPR at the sequence level, it also is ~25 bp closer to the translation start site when compared to the Inr, suggesting that this might be involed in binding a different protein. The presence of Inr and/or FPR in most of the genes where DPEpyth is found strengthens this belief. The expression data shows that the genes, carrying DPEpyth without Inr or FPR, on an average have higer maximum expression than those carrying Inr but lacking FPR and DPEpyth. This suggests that the DPEpyth might be an important core promoter element for the non-Inr and non-TATA genes. It has also been able to confirm the absence of TATA-box within most oomycete genes, and to put forward better definitions for the two other core promoter elements present in oomycetes, Inr and FPR.

As for the proximal promoter region, after a systematic study of the five key asexual stages in *P. infestans* (Chapter II), robust predictions for more than 41 stage-specific motifs were made, these are not only overrepresented but, are also positionally biased. Though, it is mention worthy that neither overrepresentation nor positional bias guarantees that a motif is a real TFBS. The fact that these elements show evolutionary conservation with either one or both of the two other *Phytophthora* species checked suggest that these elements have a very high probability of being real TFBSs. After doing functional analyses for five of these putative TFBSs viz. 'TACATGTA', 'TATTAATA', 'CGTCCTCG', 'GCTGCTG' and 'CTTCAAC', the biological activity of the last four elements were confirmed. 'TATTAATA' was found to be active in germinating cyst. 'GCTGCTG' and 'CTTCAAC' were active in mature and early sporangia respectively, whereas 'CGTCCTC' was active during cleavage. Both sporangia and cleavage are essential for the release of zoospores, the principal innoculum of the disease caused by *P. infestans,* and cysts are essential for the formation of infection structures. All of these elements have a bias for a region that is less than 600 bases from the translation start site, which is in accordance with the previous studies (Ah-Fong et al., 2007; Tani and Judelson, 2007) that have identified functional motifs. Unlike in human, there has been no evidence of distal regulators in *P. infestans*, in whatever little data that is available related to oomycete promoters. Also, the average intergenic region in *P. infestans* is 603 bp (Haas et al., 2009). Therefore, it is highly likely that these promoter elements

would work in close co-ordination with the core promoter elements, from a short distance to control gene expression.

It was found that ‘TACATGTA’, a motif overrepresented in all stages, but capable of driving very little or no reporter expression, in any, was present in tandem with the ‘TATTAATA’ and the ‘TACAGTA’ motifs in the promoters of the genes that encodes for the bZIP-like TFs and are induced in germinating cysts. The fact that the bZIP-like genes are up-regulated in the germinating cyst stage and the ‘TATTAATA’, ‘TACAGTA’ motifs are overrepresented within the same set suggests that ‘TACATGTA’ might be a binding site for a general transcription factor that needs help from other elements in different stages for regulating gene expression. The reverse analysis also showed that the motifs found in the promoters of the bZIP transcription factor genes were in congruence with their expression pattern. Genes that were upregulated in sporangia, cleaving sporangia and germinating cysts in all cases carried motifs that were found to be overrepresented in these stages.

I believe that this research should lead to the identification of transcription factors for some of the overrepresented, positionally biased and evolutionarily conserved putative TFBSs that have been predicted, with the help of biochemical approaches. The study of pathways that activate these TFs, with genetic, biochemical and cell-biological methods, should eventually lead to a detailed understanding of the mechanisms that trigger the formation of spores, the principal inoculum for the disease, and give some insight on the other

developmental stages. An understanding of the pathways involved in the asexual development *P. infestans* should throw some light on the signaling pathways that regulate development in oomycetes as a whole. In terms of broader impact, this should lead to new and improved strategies for blocking the disease.

Transgenic plants that degrade molecules found to trigger development, or chemicals that block the receptors of those molecules can be used to arrest the disease cycle. It would be a major achievement if the spore cycle can be blocked. Not only *P. infestans* but most oomycetes and fungus-like species, without spores, can neither move to a new habitat or host, nor form infection structures. This is the reason why interfering with the spore cycle has been a proven strategy for controlling disease in other systems (Kim et al., 2000; Matheron et al., 2000; Reuveni, 2003; Errampalli, 2004; Munkvoid and Marois, 1993; Wheeler et. al., 2003). Another strategy can be blocking the transcription factors responsible for driving the RXLR effector genes. The RXLR effectors are secreted and translocated into the plant cell, to suppress both PAMP-triggered and Effector-triggered immunity, by oomycete plant pathogens like *P. infestans* (Birch et al. 2008). I did not look at the RXLR effector genes specifically, but there were many RXLR effector genes within the promoters of the genes upregulated in germinating cyst. The 'TATTAATA' motif, functionality of which has been validated in this study, was found to be overrepresented within the promoters of those genes. With the advances in chemical genomics and the increasing availability of chemical libraries to screen, the information on TFs to

be blocked to stop the spore cycle or the expression of the RXLR effector genes might be of immense importance. The day may not be far when late blight is finally eradicated.

REFERENCES:

1. Li N, Tompa M (2006) Analysis of computational approaches for motif discovery. Alog. Mol. Bio. doi: 10.1186/1748-7188-1-8

2. Tompa M, Li N, Bailey TL, Church GM, De Moor B (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23: 137-44

3. Bellora N, Farré D, Albà MM (2007) Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Geno. doi: 10.1186/1471-2164-8-459

4. Tharakaraman K, Bodenreider O, Landsman D, Spouge JL, Mariño-Ramírez L (2008) The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. Nucleic Acids Res. 36: 2777–2786

5. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71-76

6. Hong RL, Hamaguchi L, Busch MA, Weigel D (2003) Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell 15: 1296-1309

7. McCue LA, Thompson W, Carmack CS, Lawrence CE (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res. 12: 1523-1532

8. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. Mol. Biol. Evol. 19: 1114-1121

9. Guo H, Moose SP (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell 15: 1143-58

10. Bowser PR, Tobe SS (2007) Comparative genomic analysis of allatostatin-encoding (Ast) genes in Drosophila species and prediction of regulatory elements by phylogenetic footprinting. Peptides 28: 83-93

11. Whitington T, Perkins AC, Bailey TL (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Res. 37: 14-25

12. Kaufmann K, Muino JM, Osteras M, Farinelli L, Krajewski P et al. (2010) Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat. Prot. 5: 457-72

13. Chen K, van Nimwegen E, Rajewsky N, Siegal ML (2010) Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. Genome Biol Evol. 2: 697-707

14. Raha D, Hong M, Snyder M (2010) ChIP-Seq: a method for global identification of regulatory elements in the genome. Curr. Prot. Mol. Biol. 2010 Chapter 21:Unit 21.19.1-14

15. Vallania F, Schiavonea D, Dewildea S, Pupoa E, Garbay S (2009) Genome-wide discovery of functional transcription factor binding sites by comparative genomics: The case of Stat3. Proc. Natl. Acad. Sci. 106: 5117-22

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

17. Judelson HS, Blanco FA (2005) The spores of *Phytophthora*: weapons of the plant destroyer. Nat. Microbiol. Rev. 3: 47-58

18. Judelson HS, Tyler BM, Michelmore RW (1992) Regulatory sequences for expressing genes in oomycete fungi. Mol. Gen. Genet. 234: 138–146

19. McLeod A, Smart CD, Fry WE (2004) Core promoter structure in the oomycete *Phytophthora infestans*. Euk. Cell 3: 91-99

20. Ah Fong A, Xiang Q, Judelson HS, (2007) Motifs regulating sporulation specific expression and transcription start site preference in the promoter of *Phytophthora infestans* Cdc14 gene. Euk. Cell 6: 2222-30

21. Judelson HS, Tani S (2007) Transgene-induced silencing of the zoosporogenesisspecific PiNIFC gene cluster of *Phytophthora infestans* involves chromatin alterations. Euk. Cell 6: 1200-1209

22. Kim BS, Lee JY, Hwang BK (2000) In vivo control and in vitro antifungal activity of rhamnolipid B, a glycolipid antibiotic, against *Phytophthora capsici* and *Colletotrichum orbiculare*. Pest Manag. Science 56: 1029-1035

23. Matheron ME, Porchas M (2000) Impact of azoxystrobin, dimethomorph, fluazinam, fosetyl-Al, and metalaxyl on growth, sporulation, and zoospore cyst germination of three *Phytophthora* spp. Plant Dis. 84: 454-458

24. Reuveni M (2003) Activity of the new fungicide benthiavalicarb against Plasmopara viticola and its efficacy in controlling downy mildew in grapevines. Eur. J. Plant Pathol. 109: 243-251

25. Errampalli D (2004) Effect of fludioxonil on germination and growth of *Penicillium expansum* and decay in apple cvs. Empire and Gala. Crop Protect. 23: 811-817

26. Munkvold GP, Marois JJ (1993) The effects of fungicides on *Eutypa lata* germination, growth, and infection of grapevines. Plant Dis. 77: 50-55

27. Wheeler IE, Hollomon DW, Gustafson G, Mitchell JC, Longhurst C et al. (2003) Quinoxyfen perturbs signal transduction in barley powdery mildew (*Blumeria graminis* f sp *hordei*). Mol. Plant Pathol. 4: 177-186

28. Birch PRJ, Boevink PC, Gilroy EM, Hein I, Pritchard L et al. (2008) Oomycete RXLR effectors: delivery, functional redundancy and durable disease resistance. Curr. Op. Plant Biol. 11: 373-379