GENETIC DIVERSITY AND EVOLUTION OF DISEASE RESPONSE GENES IN *SORGHUM BICOLOR* L. MOENCH AND OTHER CEREALS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Alejandro Zamora-Meléndez

May 2010

GENETIC DIVERSITY AND EVOLUTION OF DISEASE RESPONSE GENES IN *SORGHUM BICOLOR* L. MOENCH AND OTHER CEREALS

Alejandro Zamora-Meléndez, Ph. D.

Cornell University 2010

Several studies have shown that disease resistance genes diverge under recurrent positive selection as a result of a molecular arms-race between hosts and pathogens. However, these studies have been conducted mainly in animals and few plant genes have been shown to evolve adaptively. The study of plant molecular adaptation for disease resistance is fundamental to our understanding of plant-microbe interactions and to the development of novel plant breeding strategies. Here, we combined information from the expression pattern of *Sorghum bicolor* genes and their divergence to rice in order to identify candidate disease response genes (DRGs). We used evolutionary analyses of orthologous DRG sets from grass species to identify positively selected genes and the targeted residues. Six genes showed a pattern of substitution consistent with positive selection: a thaumatin, a peroxidase and a barley *mlo* homolog, all known antifungal proteins; and a MADS box gene, an eIF5 gene and a gene of unknown function: SESPY. All adaptive sites mapped to the surface of the crystal structures of peroxidase and thaumatin and several are close to the active sites. This information provides a basis for functional validation studies, the identification of accessions having variation at important residues and the rational design of DRGs. Rapid divergence through positive selection should correlate to reduced intraspecific polymorphism. Here we compare the macroevolution and intraspecific polymorphism of positively selected disease response genes and show that the patterns of

polymorphism found are consistent with both selective sweeps and balancing selection. The sorghum *mlo* homolog and SESPY, have old, divergent alleles, while a peroxidase and a gene with a RNA binding domain have significantly reduced diversity suggesting a recent selective sweep. Finally, we show that sorghum DRGs are significantly closer to the telomere and have more exons than a control set of evenly expressed genes. The evidence from gene location; structure; macro-evolution and polymorphism of these DRGs point to the great selective pressure produced by pathogens which has driven the evolution of cereal genome content, order and function.

BIOGRAPHICAL SKETCH

Alejandro Zamora-Meléndez was born in San José, Costa Rica, on November 17, 1970, to Raúl Zamora, a reconstructive surgeon and Marta Meléndez, a high school English teacher. As a child, Alejandro enjoyed camping on the beach, visiting National Parks with his family and having many animal pets, including two female pacas, tropical tailless rodent, the size of a big rabbit but much more intelligent and interesting. He also had fun pretending to do research with his brother, Federico, by setting traps for their dog Zuqui, a nice but sometimes quick-tempered fox-terrier, who, of course, bit him more than once, probably because Alejandro was the youngest of all three. They also set traps for birds, and even banded them with tie-wires, although the recapture rate was low as many of the birds didn't appear to survive for long. Alejandro had good grades in both elementary school and high school, but he spent his favorite hours outside of the classroom playing basketball or soccer with his friends, or riding his bike all around his hometown of Heredia. In high school his favorite subjects were math and sciences, and it was then he knew he wanted to study something that would combine his interest in animals and science in general.

After graduating from high school, he attended the University of Costa Rica, at San José, and studied biology, which from that point on became his life. He had some remarkable classes, but one of the most influential was Organic Evolution, taught by William Eberhard, a class that opened Alejandro's eyes to a different view of the world and relieved him from the myopia that is pervasive in other areas of knowledge. In 1990, Alejandro married his high school sweetheart, Peggy, and whom he had known since kindergarten. Together they had two beautiful and wonderful daughters, Leticia and Lucia. After he got married, Alejandro worked for six years as a natural history tour guide, a great opportunity he was able to take because of the English his

mother had taught him early on and his education in biology. He used to carry Dan Janzen's Costa Rican Natural History and Gary Stile's Birds of Costa Rica (Cornell University Press) everywhere and every tour he learned more and more about the different organisms he saw all over Costa Rica. During that time, Alejandro also took great courses in Ornithology, with Gilbert Barrantes and Mammalogy with José Manuel Mora and used this information to his advantage when working as a tour guide and on several projects in these areas. He also took Herpethology with Federico Bolaños, who became his first Master's project advisor and his boss in the Declining Amphibian Populations project. Alejandro felt like a frog in a pond with many ideas and projects going on, including the development of a Strawberry Dart Poisson frog farm.

However, his life took a devastating turn when Peggy died in August 16, 1996, leaving Alejandro with two young daughters, 5 and almost 2, a broken heart, a business he knew nothing of to run and an unfinished Master's. His father told Alejandro to hold on to work as a life saver and to keep moving to stay afloat. It took him many years to really move on, probably because of his father's untimely death to Chronic Myeloid Leukemia, only a few years later.

Meanwhile, he became a businessman by virtue of destiny, managing the Da Vinci Art Gallery and Frame shop, which later also became an art-school and an art-supplies store, making Alejandro realize that to survive in the small business world you have to use time and space in multiple ways. He worked in this business for five years, at the end of which he learned to love the smooth texture of a high quality frame, the value of a tight inventory and the importance of team work. He sometimes misses the times some people called him Sr. Da Vinci, or his pseudonym of Armando Marcos.

Alejandro started a different Master's project in 1997, under the direction of Ana Mercedes Espinoza, who nurtured and supported Alejandro, helping him regain confidence in life and science, and who, with her example, instilled in Alejandro a deep interest in plant genetics, disease resistance and a desire to use Costa Rica's genetic resources in an intelligent and sustainable way. In 2001, Alejandro finished his M.Sc., got married to a beautiful young lady, Carolina Delgado, and was awarded a special Fulbright scholarship in biotechnology to conduct doctoral studies, all in the month of June. The scholarship, financed by the Costa Rica-USA Foundation for Cooperation (CRUSA), gave Alejandro the opportunity to do a Ph.D. in one of the best universities in the world, something that he had desired for 10 years and which was an essential part of his plan in life. At this moment, he left his business, his extended family and friends, and Costa Rica, with its warm oceans, rain forests, frogs, birds, wild rice and other things he loved, and came to Ithaca, New York, to study Plant Breeding and Genetics at Cornell University, his first and only choice.

At Cornell, he was influenced and inspired by many brilliant professors like Charles Aquadro, Ed Buckler, Carlos Bustamante, Andy Clark, Jeff Doyle, Jason Mezey, Susan McCouch, Ross McIntyre, Rebecca Nelson, Lynda Nicholson, Rasmus Nielsen, Mark Sorrels, and Steve Tanksley. This perhaps unusual combination of influences was due to the design of Steve Kresovich, Alejandro's Ph.D. advisor, who together with Charles Aquadro had devised a plan to generate Ph.D.'s with a hybrid background in plant breeding and comparative and evolutionary genomics. Alejandro's previous interest in evolution and his experience in plant genetic resources made him a well suited candidate for this new project. Alejandro was also inspired by others who were no longer at Cornell, like Barbara McClintock, whose deep love for plant genetics and her unparalleled sensing of the hidden workings of plant genomes still echoes around campus; and by others who were not personally at Cornell but who

wrote the books that were on Alejandro's desktop for many years, like Dan Graur, Wen-Hsiung Li, Michael Lynch, Bruce Walsch, Richard N. Strange, D.S Falconer, Trudy Mackay, Phil Hedrick and Stephen J. Gould.

From fall 2006 to spring 2008, Alejandro was a Teaching Assistant in the Introductory Biology Course at Cornell, an experience he described as amazing, if extremely time-demanding. This job put Alejandro in contact with lots of great biology professors, such as Kuei-Chiu Chen, Dick Ecklund, Cole Gilbert, Carl Hopkins, Charles Walcott, and young challenging students which magnified Alejandro's biofilia and without a doubt made him a better teacher and experimental researcher.

Alejandro is living proof that anyone can do amazing things in life if he or she is given a decent set of genes, and above all, good support and opportunities. He is very grateful for all that.

Dr. Zamora is now a postdoctoral researcher at the Boyce Thompson Institute for Plant Research, in Peter Moffet's lab, where he is learning all there is to know about post-transcriptional gene silencing in plants and its role in extreme resistance to viral pathogens.

Now he only wishes he could sleep a little more.

To my dear wife Carolina, and my daughters

Leticia, Lucia and Tatiana,

in great appreciation

for all their love,

patience

and

support all these years.

ACKNOWLEDGMENTS

No one can do any work by himself and I have been blessed to have many people offer their time and ideas, as well as their emotional and economic support. This work would not have been possible without them. Some people's contribution is so big that I couldn't describe it appropriately here but you know who you are. I want to thank you all.

My daughters, Leticia, Lucia and Tatiana, who were born during my B.Sc, M.Sc and Ph.D, respectively, are my greatest inspiration and my motivation to learn more biology, because I see the miracle of life in them. I love you all and always will.

My wife, Carolina, has supported me unconditionally during all these years. I would have never been able to do any of this without her help, love and encouragement, and would have been, without a doubt, much colder and bored all those long winter nights. I never felt Ithaca's cold because you were there by my side. I love you, Gue.

My mother, Marta and don Jeni for all their support. Also, my brother Federico, his beautiful family, and all the Zamora aunts and cousins for their emotional support and encouragement through the years.

Also, a very special thank you to my father, for pointing me in the right direction, and my grandmother, Abue, for giving me a piece of paper with the ad for the Fulbright scholarship.

Guillermo and Lizette, for the many good times we shared here in Ithaca and in

Albany, and for their constant support, friendship and shrewd advice.

Floriana Blanton was somehow always able to get me out of any difficulty and fixing the problem even before I knew I was in trouble. She, David and all the Blantons welcomed me into their home from the very first time I came to Ithaca and have been my adoptive family ever since. They are Tatiana's Godparents, so we're more like a real family now. I will always cherish all the times we spent together particularly watching or playing tennis. Go Roger!

Todd Mattison has been a great friend and a great example. He is without a doubt the single most unbiased and more foreigner-friendly person I've ever met, and I admire that a lot. He is a great father and a conscientious, hard working guy who is also incredibly witty and who made me realize that I was a wimp (by actually telling me so), not only because I couldn't run 5K's every day but because I didn't get up way before the sun did to go to work. Todd I promise I'll try to do these things, someday.

Joe, we had many good tennis matches, I'm sure you are the best non-chinese member of the Chinese tennis club, and I will certainly never forget the ski chair incident or the time I lost my hearing after a visit to the mosh pit at the Haunt.

Herb, we always had a good conversation about hunting turkey or deer. I never did any of these things, but I wanted to. I shared with you the excitement of only 14 more weeks until fishing season and the Colts winning the game last Sunday.

Herman Faith and others at the Costa Rica USA foundation and the Cultural Affairs Office of the Embassy of United States of America, who gave me the opportunity to

come to Cornell as an exchange scholar, through a special Fulbright scholarship in biotechnology.

The National Science Fundation for the grant 0115903 to Steve Kresovich that supported this investigation.

Ana Mercedes Espinoza, Ana Sittenfeld, Jorge Lobo and others at the University of Costa Rica and the CIBCM, who encouraged me to come to do a Ph.D. and who supported me in this quest for knowledge.

Chepe and Franklin, my good friends, who kept me in the loop by email and phone-out, and who, along with my dear cousins Eduardo, Rafa and Emilia, threw great parties the few times I visited Costa Rica during these years.

Paul Dick Ecklund, a great teacher of teachers, was my mentor in the Intro Bio Class. I was always amazed at the depth of his knowledge and his dedication to teaching biology. I enjoyed getting out of character for a while and participating in the Photosynthesis play.

All my students in Intro Bio, who made me a better teacher and a better biologist. I learned plenty because of them and from many of them. Hyeongsu Park was my best student by far and one who made me realize what responsibility, hard work and discipline are, the Minjok way. In addition to being great at math and programming he is also a really nice guy and I'm honored to count him as my friend.

All the people at the Institute for Genomic Diversity, who taught me and helped me all the time, particularly Martha Hamblin and Sharon Mitchell, for all their good advice and for sharing their vast knowledge on molecular biology and population genetics; Hong Sun, who was my bench mate for a couple of years, Alex Casa, who was the Brazilian sun that warmed up the hearts at IGD with her hugs, Charlotte Acharya and Wenyan Zhu who were the sweet engines that did all the hard work; and my fellow graduate students Randy Wisser, Seth Murray, Patrick Brown and Maria Salas, all amazing people, who will be doing great things in science and breeding. We shared a lot of good times and enriching discussions.

Steve Kresovich, my advisor, is very sharp and seems to have the amazing ability to do many jobs at the same time. I only wish I knew, after all this years, what the nature of his amazing power to get things done is. I think getting up at 2am and working non-stop 'til late at night, without lunch, may have something to do with it. However, for him, a single phone call seems to be enough to get anything approved, so there's more to that story. Thank you for allowing me to come to your lab and interact with so many wonderful people, thank you for your economic support, and particularly thank you for your patience and understanding.

I'll always have a great appreciation for the time I spent with Charles "Chip" Aquadro and Rebecca Nelson, for the knowledge on population genetics and plant pathology they transmitted to me and for the great input they had on this work.

Steve Tanksley, Mark Sorrells and Ed Buckler for interesting discussions about plant breeding, comparative genomics and for their great example as teachers and new age breeders.

Susan McCouch is not only THE master of rice breeding and genomics, but it is also the most eloquent, trust-worthy and coherent person I have ever met. Her work in pulling together the areas of computational science and plant biology has been instrumental in developing great tools like Gramene. But her most important work is helping breeders all around the world understand and release the potential of wild rice species to make better varieties. There is no one more committed to helping the hungry with the right technology. My deepest respect is for you Susan.

Also, Jeff Doyle, Ross McIntyre, Rasmus Nielsen, Carlos Bustamante, Lynda Nicholson, Qi Sun and many other amazing professors, technicians, staff and students, who really make Cornell the great institution I have been honored to be a part of, and that will be a part of me forever.

I want to give special thanks to Peter Moffett, who gave me the opportunity to work in his lab, and who, without any self-interest reviewed and gave me keen comments on all chapters and allowed me to take some time off to finish this dissertation. Finally, all the people in the Moffett and Klessig labs at the Boyce Thompson Institute, who welcomed me with open arms and shared their great knowledge on plant defense mechanisms.

I thank you all for contributing to making my time in Ithaca and Cornell exciting, challenging many times, but always incredibly fulfilling. Doing a Ph.D. at Cornell was beyond my wildest dreams and without a doubt one of the best times in my life.

And we've only just begun.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BAC	Bacterial artificial chromosome
BLAST	Basic Local Alignment Search Tool
BSU	Biotic Stress Upregulated
cDNA	complementary DNA
CDS	Coding DNA Sequence
CIM	Composite interval mapping
DRG	Disease Response Gene
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTGS	High Through-put Genome Sequence
NBS	Nucleotide Binding Site
LD	Linkage disequilibrium
LLR	Leucine Rich Repeat
MAS	Marker assisted selection
NIL	Near isogenic line
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PS	Positively Selected
QTL	Quantitative Trait Loci
RBD	RNA Binding Domain
R-gene	Canonical Resistance gene with NBS-LRR
SNP	Single Nucleotide Polymorphism
SSR	Simple sequence repeat
TE	Transposable Element
TIR	Toll-Interleukin Receptor

LIST OF SYMBOLS

α	alpha
Å	Angstrom
μ	mutation rate
Ne	effective population size
ρ	population recombinaition parameter
ω	omega
π	nucleotide diversity (Nei 1987)
Θ	population mutation parameter, 4Neμ ,Theta

PREFACE

The study of plant-pathogen interactions has been for long a favorite of evolutionary biologists and plant breeders. Understanding the mechanisms that make the difference between life and death of a plant when in the presence of microorganisms, has far reaching consequences for our own survival, with indirect effects through plant epidemics and famine, but also more directly, due to the effect pathogen derived toxins can have in our own health and also through the information we can reveal about pathogen action and how we can use it to cure and prevent human disease. Although the differences between metazoan and angiosperm disease resistance mechanisms are clear and numerous, abundant evidence has accumulated recently showing that there are many similarities, particularly with respect to the innate responses, and some genes first identified in plants have been shown to exist in humans as well. As a great side effect of the vital importance of research in molecular plant-pathogen interactions and co-evolutionary history, we get to discover what adaptation means at the molecular, genomic and population levels and satisfy our curiosity and need for explaining natural phenomena.

CHAPTER ONE

INTRODUCTION

> "*We know about the components of genomes...We know nothing, however, about how the cell senses danger and initiates responses to it that are often truly remarkable*." (McClintock, B. 1984. Significance of responses of the genome to challenge. *Science* **226**: 792−801).

Plants are at the base of food chains in every ecosystem and, as such, must endure the attack of an almost innumerable diversity of organisms, including pathogens such as viroids, viruses, phytoplasmas, bacteria, fungi, oomycetes (see Ingram (1999) for a review), protozoans and even parasitic angiosperms. They also have to fend off pests such as nematodes, insects, mites and even large mammalian herbivores such as ourselves, to name only a few. Although there is a large list of organisms considered pathogens, only about 1-5% of the estimated diversity of their respective taxonomic groups has been formally described. Furthermore, plants have intricate relationships with endophytic bacteria and fungi that under certain conditions behave as benefactors or apparent commensals, but may become pathogenic given certain conditions.

Considering the diversity of extant potentially pathogenic micro-organisms, their large effective population sizes and therefore, their rapid response to selection, it is intriguing, if not amazing, that multicellular organisms, with much smaller effective population size do not succumb completely. Plants in particular, being the primary producers are constantly attacked by a large array of organisms trying to get the energy they have gathered. Although it has proven difficult to determine the amount

of energy a plant looses to pathogens, and therefore its reduction in fitness, a global average of 30% loss has been estimated for agricultural plants due to pathogens (Agrios 2005). However, in some particular cases, yield losses, *i.e.* selective pressure, can be close to a 100%, as have happened during severe epiphytotics that have been documented such as that on potato by the oomycete *Phythophtora infestans,* causing the Irish Famine of 1846-47; or some localities during the Southern Corn leaf blight epidemic in the US in 1970, caused by the fungal pathogen *Cochlibolus heterostrophus;* or the Bengal Famine of 1943, caused by the rice brown spot epidemic, also caused by a species of *Cochliobolus* (Agrios 2005). While the incidence of a particular disease may vary from year to year, it is almost invariably present and therefore, it is safe to argue that each plant species is under a strong and relatively constant selective pressure by a diverse array of pathogens.

Plants have a large repertoire of genes involved in resistance to pathogens (DRGs)

Because of their sessility and lack of an adaptive immune system, the genome of plants must carry numerous disease resistance genes (DRGs) in order to deal with many different potential pathogens during their lifetime. Up to 20 percent of a plant's genome may be involved in disease resistance, which corresponds to ca. 5800 genes in *Arabidopsis thaliana*, and close to 8000 genes in the genome of rice (Goff et al. 2002; Yu et al. 2002a), ca. 600 of which are canonical R genes, with Nucleotide Binding Site and Leucine Rich Repeat domains (NBS-LRR). Clearly, it wouldn't be energetically feasible for a plant to express all of these genes constitutively and therefore most of them are expressed as needed. This transcriptional reprogramming occurs rapidly, directed by the integration of information from multiple signal transduction pathways,

and has lead to the identification of many DRGs by the analysis of changes in their expression profile (Pratt et al. 2005).

Strategies used to identify disease resistance genes

Several different strategies have been used to find disease resistance genes including map based cloning; gene tagging with insertional mutagens; candidate gene analysis; bioassays to test for in vitro antimicrobial activity of purified proteins; search for signal transduction regulators and their overexpression or modification; and finally whole genome analysis using genome subtraction and expression profiling. More recent strategies involve the comparison of syntenic regions of fully and partially sequenced genomes where there is evidence of QTL involved in the trait of interest. Probes and degenerate primers have been used successfully to clone and sequence resistance gene analogs (RGAs) from several species based on conserved domains in the genes known (McIntyre et al. 2004). While these strategies are very useful, they are limited to those types of genes already known and that have not diverged significantly.

More recently reverse genetic strategies have been devised to detect the effect of selection in particular genes or regions of the genome. These analyses can be conducted without a priori hypothesis of the function of these genes, but it may also be possible to try to enrich the sample of candidate genes studied to include genes having some evidence of being used in disease resistance. One such source of evidence is the expression profile of a particular gene. Traditionally, northern blots have been used to study the expression profile of a single gene under particular circumstances. But more recently, high-throughput methods of analysis such as microarrays and EST-based expression profiles (a.k.a. electronic northerns) have been used to associate particular genes with functions. Clearly, the annotation of genes of interest should be confirmed

by multiple methods, but high through put schemes allow for the analyses of tens of thousands of genes simultaneously.

Statistical analysis and methods for detection of positive selection

Different amino acid sites in a protein have different structural and functional roles. Domains of the protein that are in the extracellular space and interact with different molecules have different constraints than those that form the protein core or help the protein to be anchored in the cellular membrane. Therefore the changes in the protein sequence at these domains may have, more or less, an impact on the shape or function of the protein as well as different selective pressures.

Pathogens and their hosts are involved in a never ending coevolutionary race in which pathogens keep in constant change to avoid recognition while their hosts need to devise specific recognition and response mechanisms to protect themselves from the infection that can result in reduced fitness or death. Therefore, particular genes from both kinds of organisms are under strong selective pressure and must evolve rapidly and constantly.

The main hypothesis in this study is that at least some genes involved in disease resistance should evolve faster than average due to the selective pressure imposed by multiple pathogens. Consequently, the identification of disease response candidate genes through the analysis of expression profiles; prior annotation, and divergence values, that additionally show signals of selection, particularly positive Darwinian selection, is here seen as a complementary strategy for the identification of genes that may be currently important and/or have been important in the evolutionary past in disease resistance. These candidate genes can then be compared to the information generated from other strategies such as QTL mapping, whole genome association studies and candidate gene cloning, and should also be functionally validated using

directed mutagenesis, knock-outs, overexpression, complementation and other biochemical methods.

Many tests of selective neutrality are based on the analysis of polymorphism, divergence and linkage disequilibrium within and around the gene in question, and require knowledge of these parameters for the average genome or for neutral regions. However, even within neutral regions, non-equilibrium demographic effects can produce patters of polymorphism identical to any of those produced by natural selection. This confusion of demographic and selective effects increases false positives and hinders the discovery of genes that have been really under selection in current times.

Some methods for identifying the fingerprint of natural selection are based on detecting regions of low variability, as a result of purifying selection. Methods such as Hudson-Kreitman-Aguadé test (Hudson et al, 1987) and Tajima's D (Tajima, 1989), have been used widely as tests of an equilibrium neutral model, *i.e.* assuming constant population size and selective pressure. However these are strictly tests of neutrality, have limited power and rely on some assumptions that are often unrealistic (Nielsen 2005). The HKA test, for instance, assumes that the effective population size remains constant throughout the entire evolutionary process, a condition which is often not true (Nei and Kumar 2000). It also requires a large sample of sequences to calculate the ratios. Tajima's D (1989) assumes a constant population size and an equal probability of mutation for all nucleotides. Other methods for testing neutrality that could be used are the McDonald-Kreitman (1991) test, and maximum likelihood estimations (Goldman and Yang, 1994). Goldman and Yang's method assumes an equal ratio of non-synonymous to synonymous substitutions along the protein sequence (Nielsen and Yang, 1998; Yokoshuki and Gojobori, 1999).

However, selection can also produce a diversifying effect, where different lineages or alleles in a locus can have more non-synonymous than synonymous substitutions at certain positions of the polypeptide. This is a stringent and unequivocal sign of positive Darwinian selection in molecular evolution (Nielsen and Yang, 1998; Swanson *et al.,* 2001). Thus the ratio of non synonymous to synonymous substitutions rates (dN/dS) is key to identifying the type and extent of selection in a gene, as well as the amino acid positions that have been subject to selective pressure.

Some methods (Nei and Gojobori, 1986; Goldman and Yang, 1994) have been used to test the neutral model of evolution using the d_N/d_S ratio, but they have considered the ratio to be uniform throughout the protein, an unrealistic assumption since different amino acids have different roles and are subject to different constraints. Some positions may be highly conserved, while in others, non-synonymous substitutions that confer a superior fitness to the individual will be selected, hence increasing the d_N/d_S ratio in the comparison of different lineages. Examples of the latter situation have been observed in the abalone sperm lysins (Lee et al., 1995), the HIV-1 genes (Yang et al, 2000), and the female reproductive proteins in mammals (Swanson *et al.,* 2001).

Nielsen and Yang (1998) and Yang *et al.* (2000) developed methods for identifying amino acid positions that are under the effect of selection. These methods are based on modeling the evolution of the nucleotide sequence as a continuous time Markov chain that has a state space in the set of all the possible codons. One of the advantages of these likelihood models over previous treatments is that they consider the differences in selective pressures among the sites and use a model where there are different categories of sites in the sequence, each with a different ω value (where $\omega = d_N/d_S$). This is important because in most proteins where selection was demonstrated

there was one or a few critical positions and therefore a general value of d_N/d_S can be estimated.

Hence, comparative methods that use the d_N/d_S rate ratio are useful for detecting the effect of numerous events of positive selection in the evolution of a gene in different lineages, where selective sweeps have occurred repeatedly leading to the fixation of replacement amino acid substitutions in the same gene and the same site(s) as a result of the interaction between host and pathogen. On the other hand, methods based on the polymorphism found in current populations can help pinpoint the effect of selection on a particular mutation that is increasing in frequency in the population or that has become recently fixed due to strong positive selection. By combining these two strategies, and in this order to avoid false positives generated by drift, we can identify disease resistance genes that show adaptive evolution across millions of years of evolution (the sum of the length of the branches in the species phylogeny) and then test to see if there are signs of recent selective events.

Positively selected disease resistance genes previously identified in plants

From a theoretical perspective, identifying genes under positive selection is important to determine the proportion of them in a functional category and to increase our understanding of the selective pressure and patterns of evolution in the plant's genome. From a more practical point of view, the detection of positively selected genes and particular codons within them can be useful in sorting functional polymorphisms from neutral variation.

Although the number of plant disease response genes identified so far is rather large, there are only three categories of genes that have been shown to evolve under positive selection, these are the two main categories of canonical disease Resistance (R) genes (Mondragón-Palomino *et al*. 2002); chitinases (Bishop, Dean, and Mitchell-Olds

2000), and polygalacturonase inhibitors (Stotz *et al*. 2000). Interestingly, Bakker *et al*. (2008) studied 27 genes previously shown to be involved in the salicylic acid pathway of signal transduction, one of the most important pathways in resistance against biotrophic pathogens, and found that most are highly conserved and show almost no signs of positive selection.

Most of the plant disease resistance genes (R genes) cloned so far have two conserved domains: a nucleotide binding site (NBS) and a carboxy terminal leucine rich repeat (LRR) (Ellis *et al.*, 2000). Two categories of NBS-LRR exist, those that have a Toll/interleukin-1-receptor (TIR) homology region and those that don't. *Toll* is a transmembrane protein of *D. melanogaster* involved in growth regulation and is homologous to the vertebrate interleukin-1 receptor (Lewin, 1998) (see Martin, Bogdanove, and Sessa (2003) for a good review on the structure and function of R genes). At least 1 percent of the *Arabidopsis* genome is composed of NBS-LRR type genes (ca. 200) and around 75% of those genes are TIR-NBS-LLR (Grube *et al.*, 2000; Ellis *et al.*, 2000). R-genes lacking the TIR domain often have another conserved domain, the coiled coil (CC) domain, which is plant specific and has a function analogous to that of TIR (Fluhr, 2001).

The grasses we call cereals

Crop species that today we call cereals were domesticated in several different places in the world some 10-14000 years from wild species of grasses. Wheat was domesticated in the Fertile Crescent around 14000 years ago by the Sumerians; sorghum was domesticated probably in what today is Ethiopia some 9000 years ago; rice was domesticated independently in South east Asia and Northeast Asia from two differentiated subspecies at least 8000 years ago (Sweeney and McCouch 2007) and more recently in Africa. These are of course only a subset of the plant species humans

have transformed genetically through selection, in greater or lesser degree, to use them more efficiently as food, feed, forage and fiber, among other purposes. Civilization as we know it could not exist without these species and the capacity they provide for grain storage, which can be used in winter, low yield seasons and times of war (Diamond 1999).

But why are the cereals so disproportionally important for our agriculture?

One of the reasons is that cereals were domesticated from wild species that lived in dense stands, tolerate many conspecifics around them and actually may use this mechanism to out-compete other plants in the environment. These are also species that thrive in open spaces and are particularly well adapted to colonize recently disturbed land, such as river margins and forest openings. They are derived from what today we call weeds, and some of the wild species behave as weeds in their domesticated sister species' plantations in many parts of the world (e. g. red rice), particularly in their centers of origin. Maize may be an exception to this rule since the closest wild relative, *Zea maize ssp. parviglumis,* is extremely different in morphology. And even though these weeds are today a severe economic problem, there is evidence that introgression from wild species, through the wild-weedy complex may have provided farmers with valuable alleles for domestication, improvement and disease resistance (Elias et al. 2001). As a matter of fact, in the case of wheat, the wild weedy complex may have yielded, over thousands of years, new polyploid species of immense value to humans. In fact the evolution of cereal species (as that of most or all plants, see (Vision, Brown, and Tanksley 2000; Paterson et al. 2006)) is dominated by whole genome duplication events, in many cases through allo-polyploidization. Additionally, several grass weed species have become adapted to

cultivated environments through Vavilovian mimicry, and have later been adopted by farmers as cereal crops (Simmonds and Smartt 1999).

Cereals as a model system for comparative genomics of disease resistance

The cereals have become an excellent model system for comparative genomics due to the availability of the full genome of rice (Goff et al. 2002) and sorghum (Paterson et al. 2009) and large amounts of genomic sequence from maize, sugarcane, wheat and barley. Additionally, physical and genetic maps for each species are available as well as web-based comparative maps (e.g. Gramene and Phytozome) that make the determination of orthology more feasible. Extensive research in disease resistance in the cereals with information of expression profiles based on EST libraries (Jantasuriyarat et al. 2005; Pratt et al. 2005) and several different methodologies provide a unique opportunity to determine whether there are signals of positive selection in several types of disease resistance genes.

Sorghum (*Sorghum bicolor*) the super crop

Sorghum is a very important crop in Africa, particularly in sub-Saharan Africa, where it is second only to pearl millet (*Pennisetum glaucum*) in drought tolerance. There are many varieties, selected for multiple uses over thousands of years since its domestication in East Africa. African varieties of sorghum have been selected for thousands of years for multiple uses, including broom-making, beer brewing, injera bread quality, sweet stem, pop sorghum, and both low and high tannin content (the last two: palatable and anti-bird, respectively) (Engels, Hawkes, and Worede 1991; BOSTID 1996). Additionally, many varieties grow up to 5m tall and are used in construction of houses or storage spaces. Other traits that show variation include the

number of seeds per caryopside and sweetness of stem (Engels, Hawkes, and Worede 1991).

Worldwide, sorghum is the fifth most important crop, with a global production of ca. 70 million tons annually from 50 million hectares (BOSTID 1996) for an average yield of only 1.4 tons per hectare (FAO 2005), a reflection of the little resources this crop has received for its improvement. The production of sorghum in the US is valued at 2 billion dollars per year, and it is likely to increase in the next few years due to the great potential of sorghum for biofuel generation in the form of ethanol derived from grain, the sweet stalk as in sugarcane, and probably from lignin and cellulose in the near future as well (Murray et al. 2008).

Sorghum Genetic Diversity

From East Africa, cultivated sorghum was taken to all other regions in Africa, where different varieties appeared and where it crosses with wild *Sorghum bicolor,* producing wild-weedy swarms, that as with many other cereal species, may have helped to generate new cultivated varieties and in general, increase the diversity of accessions.

There is plenty of genetic diversity for this species, as as many as 17 different subspecies were described initially by botanists, and hybrid grain sorghum in the US typically produces 6.5 tons per hectare. Even more tantalizing, record yields of up to 12 tons per hectare have been reported (BOSTID 1996).

Genetics and genomics of sorghum

Sorghum genomic information has increased dramatically during the last few years while this document was written, allowing for cutting edge research in this species and making it an invaluable tool for comparative genomics in the cereals. Sorghum has

10 chromosomes in its haploid complement (2n = 20) and it is an ancient paleo-polyploid that currently behaves meiotically as a diploid. Multiple mapping populations were devised during the last decade by different groups interested in the advance of sorghum genomics (Whitkus, Doebley, and Lee 1992; Chittenden et al. 1994; Pereira et al. 1994; Xu et al. 1994; Lee 1996; Dufour et al. 1997; Ming et al. 1998; Tao et al. 1998; Boivin et al. 1999; Crasta et al. 1999; Peng et al. 1999; Bhattramakki et al. 2000; Kong, Dong, and Hart 2000; Haussmann et al. 2002; Menz et al. 2002; Bowers et al. 2003). Moreover, the genome of sorghum has been aligned to that of other cereal species, revealing extensive macrocolinearity. Kim (2005) used Fluorescent In Situ Hibridization (FISH) to establish the position of single copy probes derived from BACs and previously anchored to the different mapping populations, and in order to generate a new nomenclature based on chromosome size, that could be used for all the research community.

Sorghum genome sequencing

The US Department of Energy Joint Genome Institute, under its 'Community Sequencing Program' has generated an 8X sequence coverage of the genome of *Sorghum bicolor* L. genotype, BTx623, available for comparative analyses through Phytozome. The 10 million sequences generated have been assembled into 'contigs' (contiguous sequences without gaps) and 'scaffolds' (reconstructed stretches with any gaps spanned by at least two end-sequenced clones, and prealigned with publicly available sequences. They have also been integrated with physical and genetic maps to yield genetically-oriented pseudo-molecules that cover over 90% of *Sorghum* chromosomes and previously predicted expressed genes, as determined by an extensive EST project (Pratt et al. 2005).

Comparison of the 8x *Sorghum bicolor* contiguous sequence assemblies with known genes and protein sequences of several other species was used to generate the gene models in sorghum. The pipeline used resulted in 28,003 complete gene models. In addition, 6493 candidate genes that lack a start and/or stop codon were assigned as partial models, which don't overlap with complete models. Therefore a total of 36,338 transcript models at 34,496 loci are mapped in the current genome sequence. Partial gene models may result from several, not mutually exclusive reasons: (i) sequencing or assembly errors may hinder both *ab initio* and homology based predictors to deduce a correct ORF; (ii) transposon activity may have truncated gene models; (iii) we have insufficient evidences from ab initio predictions or EST matches to provide a complete gene model. At least 66% of the sorghum genome is composed of repeated elements, including transposable elements (Phytozome, Oct. 2008). This appears to be an underestimation of the total amount of repeated elements as there are possibly more diverged or fossil TEs, as well as lower copy repeated elements that have not been described yet.

Due to its outstanding characteristics as a crop, its close relationship with other important crops such as maize and sugarcane, and the vast genetic and genomic information available for this species, sorghum is poised to be an essential model plant to conduct comparative analyses and to generate cutting-edge knowledge for plant breeding and evolutionary genomics in the cereals.

REFERENCES

Agrios, G. 2005. Plant pathology. New York:Elsevier Academic Press.

Bakker, E. G., M. B. Traw, C. Toomajian, M. Kreitman, and J. Bergelson. 2008. Low levels of polymorphism in genes that control the activation of defense response in Arabidopsis thaliana. Genetics **178**:2031-2043.

Bhattramakki, D., J. Dong, A. K. Chhabra, and G. E. Hart. 2000. An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. Genome **43**:988-1002.

Bishop, J. G., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A **97**:5322-5327.

Boivin, K., M. Deu, J. F. Rami, G. Trouche, and P. Hamon. 1999. Towards a saturated sorghum map using RFLP and AFLP markers. Theoretical & Applied Genetics **98**:320-328.

BOSTID. 1996. Lost Crops of Africa: Grains. Washington, D.C.:National Academy Press.

Bowers, J. E., C. Abbey, S. Anderson et al. 2003. A high-density genetic recombination map of sequence-tagged sites for *Sorghum*, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165**:367-386.

Chittenden, L. M., K. F. Schertz, Y. R. Lin, R. A. Wing, and A. H. Paterson. 1994. A detailed RFLP map of *Sorghum bicolor x S. propinquum*, suitable for high-density mapping, suggests ancestral duplication of *Sorghum* chromosomes or chromosomal segments. Theoretical and Applied Genetics **87**:925-933.

Crasta, O. R., W. W. Xu, D. T. Rosenow, J. Mullet, and H. T. Nguyen. 1999. Mapping of post-flowering drought resistance traits in grain sorghum: Association between QTLs influencing premature senescence and maturity. Molecular & General Genetics **262**:579-588.

Diamond, J. 1999. Guns, Germs, and Steel: The Fates of Human Societies:W. W. Norton & Company.

Dufour, P., M. Deu, L. Grivet, A. D'Hont, F. Paulet, A. Bouet, C. Lanaud, J. C. Glaszmann, and P. Hamon. 1997. Construction of a composite sorghum genome map and comparison with sugarcane, a related complex polyploid. Theoretical & Applied Genetics **94**:409-418.

Elias, M., L. Penet, P. Vindry, D. McKey, O. Panaud, and T. Robert. 2001. Unmanaged sexual reproduction and the dynamics of genetic diversity of a vegetatively propagated crop plant, cassava (*Manihot esculenta* Crantz), in a traditional farming system. Molecular Ecology **10**:1895-1907.

Engels, J., J. Hawkes, and M. Worede. 1991. Plant Genetic Resources of Ethiopia. Pp. 280. Cambridge University Press, New York.

FAO. 2005. FAOSTAT DATA. http://faostat.fao.org/default.aspx.

Goff, S. A., D. Ricke, T. H. Lan et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. *ssp japonica*). Science **296**:92-100.

Haussmann, B. I. G., D. E. Hess, N. Seetharama, H. G. Welz, and H. H. Geiger. 2002. Construction of a combined sorghum linkage map from two recombinant inbred populations using AFLP, SSR, RFLP, and RAPD markers, and comparison with other sorghum maps. Theoretical & Applied Genetics **105**:629-637.

Ingram, D. S. 1999. Biodiversity, plant pathogens and conservation. Plant Pathology (Oxford) **48**:433-442.

Jantasuriyarat, C., M. Gowda, K. Haller et al. 2005. Large-scale identification of expressed sequence tags involved in rice and rice blast fungus interaction. Plant Physiology (Rockville) **138**:105-115.

Kim, J. S., M. N. Islam-Faridi, P. E. Klein, D. M. Stelly, H. J. Price, R. R. Klein, and J. E. Mullet. 2005. Comprehensive Molecular Cytogenetic Analysis of Sorghum Genome Architecture: Distribution of Euchromatin, Heterochromatin, Genes and Recombination in Comparison to Rice. Genetics **171**:1963-1976.

Kong, L., J. Dong, and G. E. Hart. 2000. Characteristics, linkage-map positions, and allelic differentiation of Sorghum bicolor (L) Moench DNA simple-sequence repeats (SSRs). Theoretical & Applied Genetics **101**:438-448.

Lee, M. 1996. Comparative genetic and QTL mapping in sorghum and maize. Symp Soc Exp Biol **50**:31-38.

Martin, G. B., A. J. Bogdanove, and G. Sessa. 2003. Understanding the functions of plant disease resistance proteins. Annual Review of Plant Biology **54**:23-61.

McIntyre, C. L., S. M. Hermann, R. E. Casu et al. 2004. Homologues of the maize rust resistance gene Rp1-D are genetically associated with a major rust resistance QTL in sorghum. Theor appl genet **109**:875–883.

Menz, M. A., R. R. Klein, J. E. Mullet, J. A. Obert, N. C. Unruh, and P. E. Klein. 2002. A high-density genetic map of Sorghum bicolor (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Mol Biol **48**:483-499.

Ming, R., S. C. Liu, Y. R. Lin et al. 1998. Detailed alignment of saccharum and sorghum chromosomes: Comparative organization of closely related diploid and polyploid genomes. Genetics **150**:1663-1682.

Mondragón-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Research **12**:1305-1315.

Murray, S. C., W. L. Rooney, S. E. Mitchell, A. Sharma, P. E. Klein, J. E. Mullet, and S. Kresovich. 2008. Genetic Improvement of Sorghum as a Biofuel Feedstock: II. QTL for Stem and Leaf Structural Carbohydrates. Crop sci **48**:2180-2193.

Nielsen, R. 2005. Molecular signatures of natural selection. Annual Review of Genetics **39**:197-218.

Paterson, A. H., J. E. Bowers, R. Bruggmann et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**:551-556.

Paterson, A. H., B. A. Chapman, J. C. Kissinger, J. E. Bowers, F. A. Feltus, and J. C. Estill. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. Trends Genet **22**:597-602.

Peng, Y., K. F. Schertz, S. Cartinhour, and G. E. Hart. 1999. Comparative genome mapping of Sorghum bicolor (L) Moench using an RFLP map constructed in a population of recombinant inbred lines. Plant Breeding **118**:225-235.

Pereira, M. G., M. Lee, P. Bramel-Cox, W. Woodman, J. Doebley, and R. Whitkus. 1994. Construction of an RFLP map in sorghum and comparative mapping in maize. Genome **37**:236-243.

Pratt, L. H., C. Liang, M. Shah et al. 2005. Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and skotomorphogenesis from a milestone set of 16,801 unique transcripts. Plant Physiol. **139**:869-884.

Simmonds, N., and J. Smartt. 1999. Principles of Crop Improvement. Pp. 412. Wiley-Blackwell, Malden, MA.

Stotz, H. U., J. G. Bishop, C. W. Bergmann, M. Koch, P. Albersheim, A. G. Darvill, and J. M. Labavitch. 2000. Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors. Physiological & Molecular Plant Pathology **56**:117-130.

Sweeney, M., and S. McCouch. 2007. The Complex History of the Domestication of Rice. Ann Bot **100**:951-957.

Tao, Y. Z., D. R. Jordan, R. G. Henzell, and C. L. McIntyre. 1998. Construction of a genetic map in a sorghum recombinant inbred line using probes from different sources and its comparison with other sorghum maps. Australian Journal of Agricultural Research **49**:729-736.

Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. Science **290**:2114-2117.

Whitkus, R., J. Doebley, and M. Lee. 1992. Comparative genome mapping of Sorghum and maize. Genetics **132**:1119-1130.

Xu, G. W., C. W. Magill, K. F. Schertz, and G. E. Hart. 1994. A RFLP linkage map of Sorghum bicolor (L.) Moench. TAG Theoretical and Applied Genetics **89**:139-145.

Yu, J., S. Hu, J. Wang et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science **296**:79-92.

CHAPTER TWO

POSITIVELY SELECTED DISEASE RESPONSE ORTHOLOGOUS GENE SETS IN THE CEREALS IDENTIFIED USING *SORGHUM BICOLOR* L. MOENCH EXPRESSION PROFILES AND COMPARATIVE GENOMICS[1]

Abstract

Disease response genes diverge under recurrent positive selection as a result of a molecular arms-race between hosts and pathogens. Most of these studies were conducted in animals, and few defense genes have been shown to evolve adaptively in plants. To test for adaptation in the molecules mediating disease resistance in the cereals we first combined information from the expression pattern of *Sorghum bicolor* genes and from divergence to the full genome of rice to identify candidate disease response genes. We then used evolutionary analyses of orthologous gene sets from several grass species, to determine whether the disease response genes show signals of positive selection and the residues targeted. We found 140 divergent genes upregulated under biotic stress in *S. bicolor* by evaluating the relative abundance of ESTs in different libraries and comparing them to rice genes. For 10 of these genes, we found sets of orthologs including sequences from rice and three other cereals; 6 genes showed a pattern of substitution that was consistent with positive selection. Three of these genes, a thaumatin, a peroxidase and a barley *mlo* homolog, are known antifungal proteins. The other three genes with evidence of positive selection were a MADS box transcription factor, an eIF5 translation initiation factor and a gene of unknown function but with evidence of expression during stress. Permutation

[1] Zamora, A., Q. Sun, M. T. Hamblin, C. F. Aquadro, and S. Kresovich. 2009. Positively selected disease resistance orthologous gene sets in the cereals identified using Sorghum bicolor L. Moench expression profiles and comparative genomics. Molecular Biology and Evolution **26**:2015-2030.

analyses, using different ortholog and paralog sequences, consistently identified 5 positively selected codons in the peroxidase, a member of a cluster of genes and a large gene family. We mapped the positively selected residues onto the structure of the peroxidase and thaumatin and found that all sites are on the surface of these proteins and several are close to biochemically determined active sites. Identifying new positively selected plant disease resistance genes and the critical amino acid sites provides a basis for functional studies that may increase our understanding of their underlying molecular mechanisms of action. Additionally, it may lead to the identification of individuals having variation at functionally important sites, as well as eventually using this information in the rational design and engineering of proteins involved in plant disease resistance.

Introduction

Evolutionary theory predicts that the long term interaction between host and pathogen populations should lead to an arms race, preventing organisms in either population from achieving an optimal genotype (Maynard Smith 1998). This competition for increased fitness in the face of constant change is expected to lead to recurrent events of positive selection whenever advantageous mutations appear that increase the capacity for disarming the other member of the antagonistic couple.

The unambiguous identification of positive selection is one of the most important endeavors in modern population genetics and comparative genomics (Nielsen 2005). Recently, numerous studies have identified positively selected disease response genes (DRGs) in several species of animals (Nielsen 2005; Schlenke and Begun 2005). However, although a large number and diversity of angiosperm DRGs have been described (Hammond-Kosack and Jones 1997; Michelmore and Meyers 1998; van Loon, Rep, and Pieterse 2006), only those from three categories have been shown to

have evolved under positive selection by looking at the pattern of nucleotide substitutions in orthologous genes. The first category corresponds to putative pathogen recognition proteins of the nucleotide binding site and leucine-rich repeat domain type (NBS-LRR), and is divided into two sub-categories depending on the domain present at the N-terminus, namely, the TIR (Toll-Interleukin 1 Receptor) or the CC (coiled-coil) domain, *i.e.* the classical, dominant, disease resistance (R) genes (Mondragón-Palomino et al. 2002). The second category consists of pathogen wounding enzymes, namely chitinases (Bishop, Dean, and Mitchell-Olds 2000; Tiffin 2004) and glucanases (Bishop et al. 2005), that degrade fungal cell walls and therefore act as pathogen-wounding proteins, but have traditionally been called pathogenesis related proteins or PRPs due to their synthesis in plants as a response to microbial attack (Bowles 1990; Sticher, Mauch-Mani, and Metraux 1997). The third category consists of polygalacturonase inhibitor proteins (PGIPs), which bind pathogen-derived enzymes that degrade the pectic oligomers in the cell wall, restraining their action (Stotz et al. 2000).

While these types of plant DRGs include numerous genes (e.g. in rice the number of canonical R genes is ca. 600 (Goff et al. 2002; Yu et al. 2002)) and are essential for the functioning of several disease resistance mechanisms, they represent two extremes of the process, viz. the recognition of the pathogen at the beginning of infection, and the molecular clash with the pathogen in a direct and often highly specific manner. Many genes that are important in these as well as other stages of defense, such as transcription and translation factors, have not been assessed for signals of historical molecular adaptation in plants. Interestingly, Bakker et al. (2008) studied recently the genetic diversity at 27 loci in *Arabidopsis* involved in the Salicylic Acid, Jasmonic Acid and Ethylene disease signaling transduction pathways and found that most are under strong purifying selection, with only hints of positive selection in 8 genes. Such

an unexpected result, contrasting with the pattern of polymorphism found for R genes in the same panel (Bakker et al. 2006), is intriguing and warrants further investigation of molecular evolution in genes from signal transduction pathways and other downstream elements of disease resistance. It is essential to determine whether other types of plant DRGs, known or yet un-annotated as such, have evolved under positive selection, particularly considering the strong and constant pressure imposed by pathogens, or whether there are unknown constraints to positive selection in the genome of plants that have lead to the current dearth of examples of adaptive evolution in plant genes downstream of NBS-LRR receptors.

Evidence for adaptive evolution can be detected by maximum likelihood codon-based analyses of the ratio of non-synonymous substitutions per non-synonymous site to that of synonymous substitutions per synonymous site. These tests have revealed numerous genes in many organisms for which particular domains or even single codons have evolved under positive selection (Nielsen and Yang 1998; Yang et al. 2000; Nielsen 2005). This method is most powerful provided there are sets of sequences from orthologous genes available for several species (Nielsen and Hubisz 2005) as is increasingly the case in the grasses.

The recent exponential increase in the availability of sequences from various cereal species in public databases, along with information on their expression patterns, provides an opportunity to identify and compare orthologous genes -which may also have a common function- across the grasses. Information available from the annotated rice genome provides a framework for correctly identifying, organizing and comparing the orthologs from other grass species and for transferring useful information across the cereals (Jaiswal et al. 2002). Additionally, the phylogeny of cereals and a few other grasses has been thoroughly studied using numerous morphological and biochemical characters (Barker et al. 2001), therefore a reliable consensus phylogeny

for subfamilies and tribes exists that can be used as a null model to study the pattern of substitutions in the genes of interest in an evolutionary framework. When comparing genes from the cereal species, the maximum likelihood approaches to modeling gene evolution have the additional benefit of looking at millions of years of gene evolution along the different lineages within Poaceae, since the most recent common ancestor between rice and sorghum (the most diverged lineages evaluated here) existed at least 50 million years ago (Barker et al. 2001; Prasad et al. 2005). More importantly, these methods are less sensitive to problems caused by changes in the patterns of genetic diversity resulting from non-equilibrium demography, which can mimic any and all of those generated by selection (Nielsen 2005).

A useful pathosystem for identifying candidate disease resistance genes in the grasses is that of *Sorghum bicolor* and its fungal pathogen *Colletotrichum sublineolum,* the causal agent of anthracnose. Different strains of anthracnose affect the stalk, foliage, panicle and grain of sorghum, and losses of up to 50% have been reported. This fungal pathogen has high genetic diversity (Rosewich et al. 1998) and its ability to change rapidly through meiosis and parameiosis, generating strains with increased pathogenicity (Souza-Paccola et al. 2003), suggests it may have had an important role as a selective factor and consequently as a driving force behind rapid gene evolution in its host. Analysis of the inheritance of anthracnose resistance (Mehta et al. 2005) suggests the involvement of several major genes (so far unknown), as well as a complex genetic basis also suggested by the variation for resistance in landraces from several locations in Africa (Erpelding and Prom 2006). Furthermore, different but closely related species, *Colletotrichum falcatum* and *C. graminicola,* infect sugarcane (Suman et al. 2005) and maize (Crouch, Clarke, and Hillman 2006), respectively, suggesting that these fungal pathogens might have co-evolved with their host plants for perhaps several millions of years, since sorghum, sugarcane and maize

are members of the tribe Andropogoneae (Poaceae; Figure 2.1). Such antagonistic coevolution would have left its mark on the genomes and the molecules involved, increasing the probability of identifying genes that have evolved under selection in the cereals due to the pressure imposed by this pathogen.

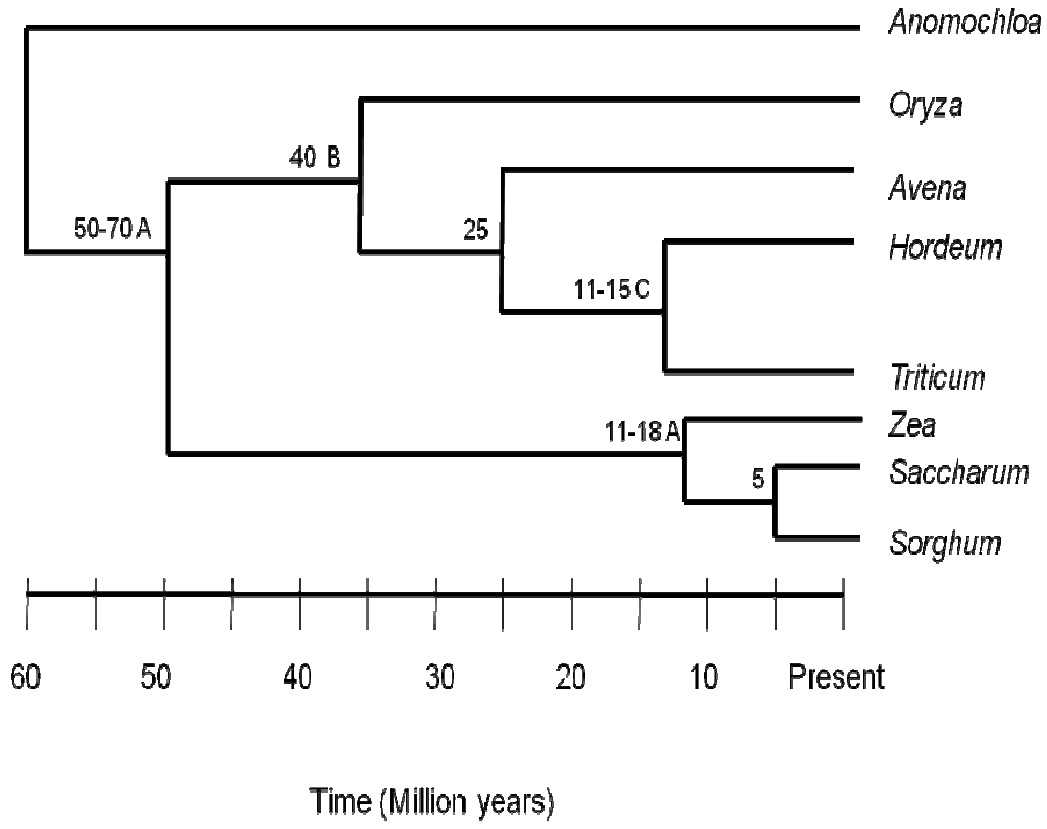


Figure 2.1. Phylogeny of the main cereal lineages showing the divergence times in millions of years (MY) and the references used. A. (Rice Chromosome 3 Sequencing 2005), B. (Ramakrishna et al. 2002); C. (Paterson, Bowers, and Chapman 2004).

In addition to the recently sequenced genome (Phytozome), there are over 200,255 ESTs available for *Sorghum bicolor,* from 20 different and non-normalized EST

libraries. Two of these libraries were made after inoculating either resistant or susceptible sorghum plants with *C. sublineolum,* the anthracnose pathogen (Pratt et al. 2005), producing incompatible and compatible reactions, respectively. These two libraries have been used to conduct EST-based differential expression analyses using custom designed software and to identify genes upregulated as a result of inoculation (Pratt et al. 2005). In this paper we describe the identification of biotic stress induced genes from *Sorghum*, some of which were not previously known to be involved in response to anthracnose or other (fungal) diseases. We also generate sets of putative orthologous genes in the cereals and demonstrate that several show statistical evidence for recurrent positive selection. Finally we identify the position of positively selected amino acid residues in the tertiary structure of thaumatin and peroxidase, both previously described pathogenesis related proteins.

Material and methods

Identification of biotic stress upregulated genes in Sorghum bicolor

In *S. bicolor,* the Biotic Stress subgroup of EST libraries is comprised of 3 EST libraries called the Pathogen induced Incompatible (PI1, resistant reaction, 9533 ESTs), the Pathogen Infected Compatible (PIC1, susceptible reaction, 10209 ESTs) and the Salicylic Acid treated library (SA1, 5801 ESTs) (see for a detailed description of the 20 *Sorghum bicolor* EST tissue and treatment specific libraries and analysis software). We used the program MAGIC Gene Discovery (Laboratory for Genomics and Bioinformatics, University of Georgia) to identify uniscripts constructed using ESTs that came primarily from these 3 libraries. A uniscript is defined here as a unique splicing form of a gene model made from ESTs that can come from only one (tissue or treatment specific genes) or many EST libraries (e.g. house-keeping genes). We selected uniscripts having more than 50% of ESTs expressed in the biotic stress

subgroup of libraries and with at least 3 ESTs composing the uniscript. A believability score was generated by Pratt (2005) for each of the uniscripts, based on the method of Stekel, Git, and Falciani (2000). This believability score reflects the probability that a gene is upregulated in a particular library and we report it for the genes used in this study. We also used MAGIC to obtain a control set of 700 constitutively expressed uniscripts, having a library ratio of 0.05 (no more than 5% of ESTs from any of the 20 libraries) and more than 20 ESTs.

Annotation of the upregulated genes using gene ontology

We used the transeq program (EMBOSS; http://www.ebi.ac.uk/emboss/transeq/) to translate all the 773 pathogen induced uniscripts, as well as the control set, in all six frames and compared these translations to the rice proteins (TIGR) using BLASTp to obtain the most likely orthologs in rice, and to identify highly expressed and divergent genes. In order to identify known protein domains in the uniscripts, we used parallel InterPro Scan, implemented at the Computational Biology Service Unit (CBSU, Cornell University) and compared this annotation to that precalculated by the MAGIC gene discovery tool, which compared all the uniscripts to the PIR database using TBLASTX. We then obtained the Gene Ontology (GO Slim) annotation for the rice genes from the TIGR website, using the LOC_Os identification codes in Batch download.

Identification of orthologous genes in other grass species.

To identify putative orthologous genes in other cereal species, particularly in sugarcane, maize, barley, wheat and rice, we used parallel-BLASTn (P-BLASTn), as implemented in the computer cluster of the Computational Biology Service Unit (Cornell University), and conducted a similarity search with default parameters using the set of *Sorghum* biotic stress upregulated genes (n=773) as a query, and all the

nucleotide Genbank sequences (nr) as database. To identify uniscripts with hits to the species of interest, we used SQL queries specifying the scientific names of the acceptable target species. We used SQL queries to select those uniscripts that had an E value suggestive of moderately rapid evolution (where the best hit was in the interval: $10^{-50} > E > 10^{-5}$) and then used Microsoft Excel to parse the SQL results and to identify those uniscripts with hits to three or more species having mRNAs and particularly those having prior information suggesting a function in disease resistance.

To corroborate best hits and reciprocal best hits, we used the sequences from all species in the selected sets of orthologous genes in subsequent BLASTn searches using other databases such as the EST_others and genome survey sequences (GSS). Good hits to ESTs, NCBI unigenes and genomic sequences such as High CoT and methylation filtration sequences were downloaded using Batch download from NCBI and used to construct strict contigs in Sequencher (Codon Codes, Ann Arbor, Michigan) with identity greater than 95% and overlap of more than 20 bases, allowing for large gaps to align the genomic sequences to the expressed sequences. These contigs were then used as input in FGenesH (Salamov and Solovyev 2000) for *ab initio* gene prediction and translation of the ORFs into protein sequences. These predicted genes were compared to the predicted genes in rice, which were obtained also by FGenesH (Goff et al. 2002; Yu et al. 2002) and have been corroborated in many cases by full length mRNA analyses. This strategy was initially used to identify full ORFs in sorghum, but since this work was conducted while the sorghum genome was being sequenced, we used the early genome drafts to corroborate our results and to obtain further gene models.

In order to select the most likely ortholog from each species, we required that all the exons aligned well from the beginning to the end with few gaps. If more than one such

gene from a species was found in a cluster of orthologs, we chose the one that was the closest match to sorghum over the entire length of the coding region.

Sequences that appeared in unexpected positions in the species phylogeny, for instance, a maize gene with greater divergence to sorghum than that of the rice ortholog, were considered old, divergent paralog genes. This strategy assumes that most of the substitutions are neutral and therefore the genealogy reflects the phylogenetic relationships of the species, while those sites that are under selection are non-informative or are very few compared to the neutral sites and therefore do not affect the overall topology.

We re-aligned the nucleotide sequences based on the protein sequence alignments using protal2dna (K. Schuerer and C. Letondal, Pasteur Institute bioinformatics) or tranalign (EMBOSS package). These nucleotide sequence alignments were used to identify the genes and the particular codons under positive selection using the maximum likelihood methods described by Nielsen and Yang (1998) and Yang et al. (2000) and implemented in the codeml algorithm from the PAML package (version 3.13), as well as by the Fixed Effects Likelihood (FEL) and Random Effects Likelihood (REL) models described by Kosakovsky-Pond et al. (2005) and implemented in the online version of the HyPhy package available at http://www.datamonkey.org. Regions with gaps were not included in the analyses of these orthologs and we refer to these sets as NG (No Gaps).

Effect of including paralogs in the determination of adaptive evolution

To assess the effects of including paralogs in the determination of adaptive evolution, we generated two additional sets of peroxidase genes (Paralogs Set 1 and 2, PS1 and PS2). We downloaded all the peroxidases from sorghum and other cereal species with high to moderate matches to the upregulated sorghum gene, and made phylogenetic trees using the UPGMA algorithm with 2000 bootstrap replications (as implemented in Megalign). We used all the obtained paralogs from all the species together in order to identify those genes from different species that clustered together in a subtree resembling the species phylogeny as an indication of orthology, as well as paralogs showing different levels of divergence.

Positive selection and proportion of sites under positive selection in the candidate genes.

We used the orthologous gene sets to construct likelihood ratio tests to determine whether these genes have evolved under selection. We first compared a null model (M0) with a single dN/dS (ω) rate to one with more classes (M3) to test whether the latter model fits the data better. Then we made a likelihood ratio test for the case where the reduced model includes ten classes of ω within the beta distribution, i.e. where classes range from strong negative selection to neutrality in the interval between 0 and 1 (M7), and the unrestricted model where there is an additional class of sites where ω can be greater than 1 (M8), a strong indication of positive selection.

Modelling of the tertiary structure of the candidate genes

Candidate genes thaumatin (2_8981) and peroxidase (2_7192) have close homologs for which the crystal structure of the protein has been determined, and we downloaded their PDB files (peroxidase 1H5D, thaumatin 1RQW) from the protein databank

(www.rcsb.org_pdb). We used Swiss Deep View and Modeller (in parallel at CBSU) to align the candidate gene protein product sequence to that of the homolog and to minimize the energy in the model. We then used Deep View to map the amino acid sites predicted to be under selection to the 3D model of the protein.

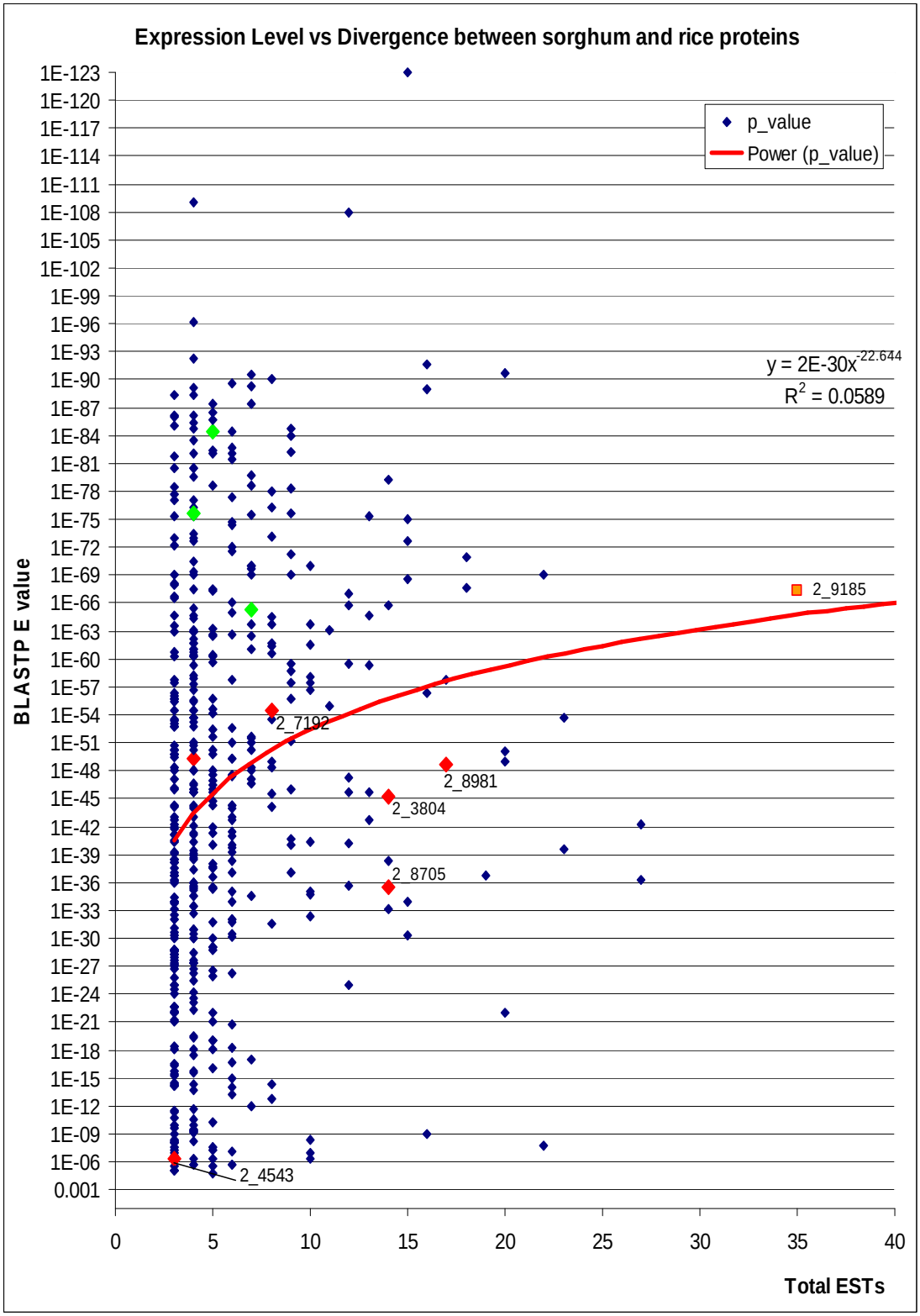
Results

Biotic Stress Upregulated genes in S. bicolor *include some highly expressed and divergent members*

Out of a total of 57,673 *S. bicolor* uniscripts, 773 showed increased level of expression in the biotic stress subgroup of EST libraries under the conditions defined in MAGIC Gene Discovery, *i.e.* those uniscripts have more than 50% of their ESTs belonging to this subgroup and more than 3 ESTs total. On average, these Biotic Stress Upregulated (BSU) uniscripts have a length of 739.9 basepairs (SD = 103.1bp), are made up of 7.1 ESTs (SD = 13.2), and have a biotic stress ratio of 0.67 (SD = 0.16). A few of these uniscripts (n = 9, 1.2%) appear to be contaminations of fungal origin, based on tBLASTx to the genome of *Neurospora crassa*, and they were eliminated from the analysis.

In order to annotate the BSU genes and to determine how divergent they are, we used BLASTp to compare them with rice proteins (TIGR v.3), and identified 505 sorghum BSU uniscripts with significant (E value < 10^{-5}) hits in rice (Fig. 2.2 and Figure 2.3). The mean divergence, in terms of E value, was 10^{-45}, with a standard deviation of E = 10^{-23}. The best fit line shows a subtle increase in the level of expression related to evolutionary conservation, particularly for the lowest categories of expression (3 to 6 ESTs, Fig. 2.2).

Figure 2.2. Expression level of biotic stress upregulated uniscripts as measured by the total number of ESTs by uniscript, and divergence between sorghum and rice proteins. Red diamonds: Disease response gene candidates with sites evolving under positive selection (Not labeled: 2_11684). Green diamonds: Conserved genes evolving under purifying selection (2_12005, 2_6529, 2_11011). Orange square: Conserved gene for lipid transfer protein, highly upregulated during biotic stress and with prior information of a role in disease resistance (see text).


Expression Level vs Divergence between sorghum and rice proteins
BLASTP E value
1E-123
1E-120
1E-117
1E-114
1E-111
1E-108
1E-105
1E-102
1E-99
1E-96
1E-93
1E-90
1E-87
1E-84
1E-81
1E-78
1E-75
1E-72
1E-69
1E-66
1E-63
1E-60
1E-57
1E-54
1E-51
1E-48
1E-45
1E-42
1E-39
1E-36
1E-33
1E-30
1E-27
1E-24
1E-21
1E-18
1E-15
1E-12
1E-09
1E-06
0.001
0
5
10
15
20
25
30
35
40
Total ESTs
p_value
Power (p_value)
y = 2E-30x^-22.644
R^2 = 0.0589
2_9185
2_7192
2_8981
2_3804
2_8705
2_4543

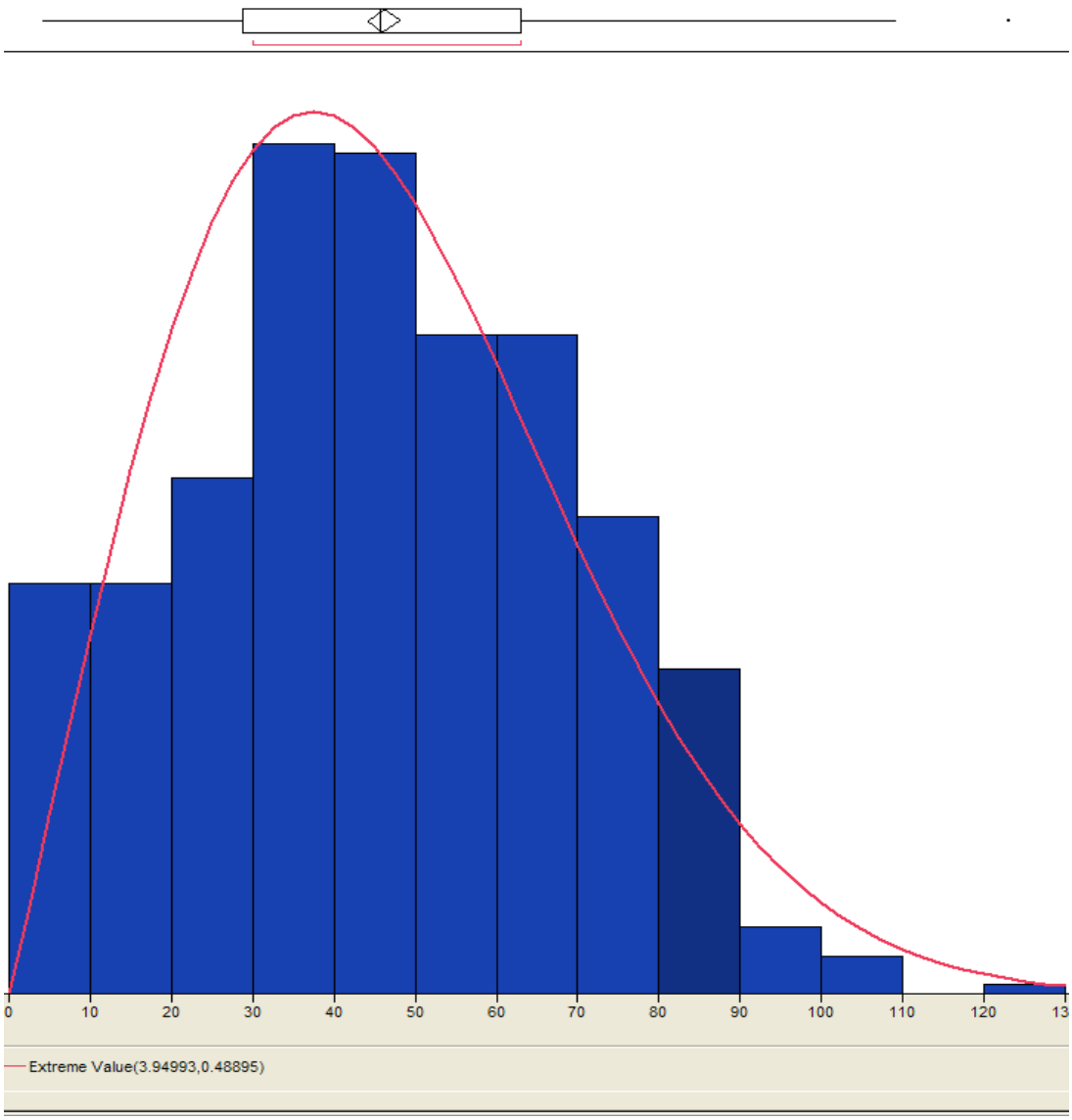


Figure 2.3. Distribution of BLASTp E values (-log transformed) of translated sorghum uniscripts vs. rice proteins. The distribution of E values has a significantly good fit to the Extreme Value distribution, as can be expected due to the use of BLAST. Mean = 46.15, Standard Deviation = 23.35.

However, overall, no significant correlation exists between the level of expression and the degree of conservation for this set of sorghum BSU uniscripts (Correlation coefficient = -0.026). Most genes (92%) had less than 10 ESTs, regardless of their divergence, and most conserved genes (BLASTp, $E< 10^{-45}$) show a low level of expression. Interestingly, however, there are several genes that show high expression and divergence (lower right quadrant of Fig. 2.2), suggesting that conditionally expressed disease response genes may include some members that have diverged under positive selection.

A comparison of the Gene Ontology (GO) annotation for the best hit in rice for all the 505 sorghum BSU uniscripts (Fig. 2.4) and a set of 700 constitutively expressed genes (Fig. 2.5), revealed that while kinase activity is the main molecular function identified in both groups, there was a 6% increase in the BSU genes. Two other main categories, DNA binding and protein binding, increased as well (both 3%), while catalytic activity, hydrolase activity and transcription activity all decreased substantially (14%, 8% and 7% less, respectively). Many categories are exclusively present in the BSU gene set, including calmodulin binding (3%), oxygen binding (3%), carbohydrate transporter (2%), lipase (2%), lipoxygenase (1%) and calcium and calmodulin dependent protein kinase activity (1%). Of these 505 BSU genes, there were several with annotation strongly suggesting a role in disease response, including NBS-LRR genes (n = 7), NB-ARC domain genes (n = 5), lectins (n =3), peroxidases (n = 2), thaumatins (n = 2), and one gene of each of the following: Enhanced Disease Resistance (EDR1) homolog; chitin inducible gene; Pathogenesis Related Protein 1b, and MLO-homolog. Interestingly, polygalacturonase inhibitor proteins (PGIPs) were not found in the BSU gene set, but were present in the constitutively expressed control gene set.

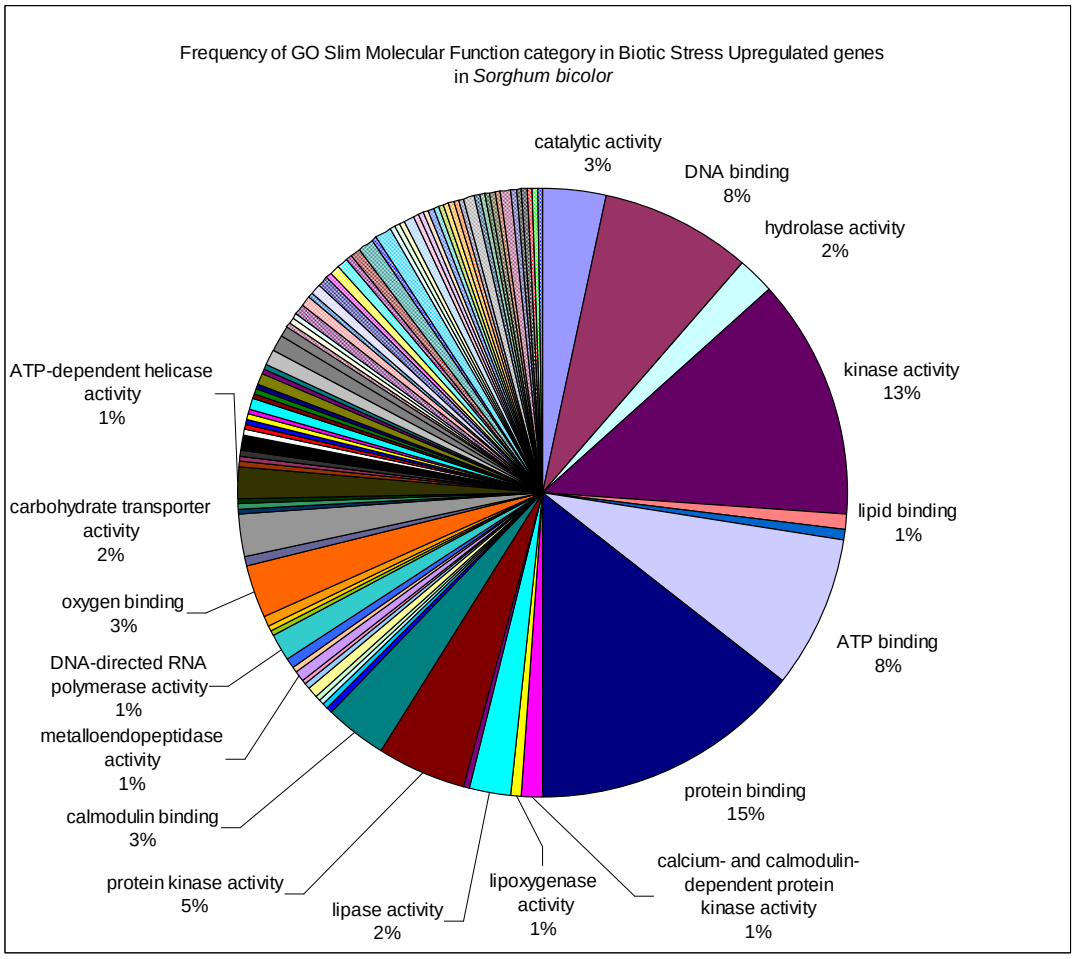


Figure 2.4. Gene ontology annotation (GO Slim, molecular function category) of the *Sorghum bicolor* biotic stress upregulated uniscripts based on the annotation of their orthologs in rice.

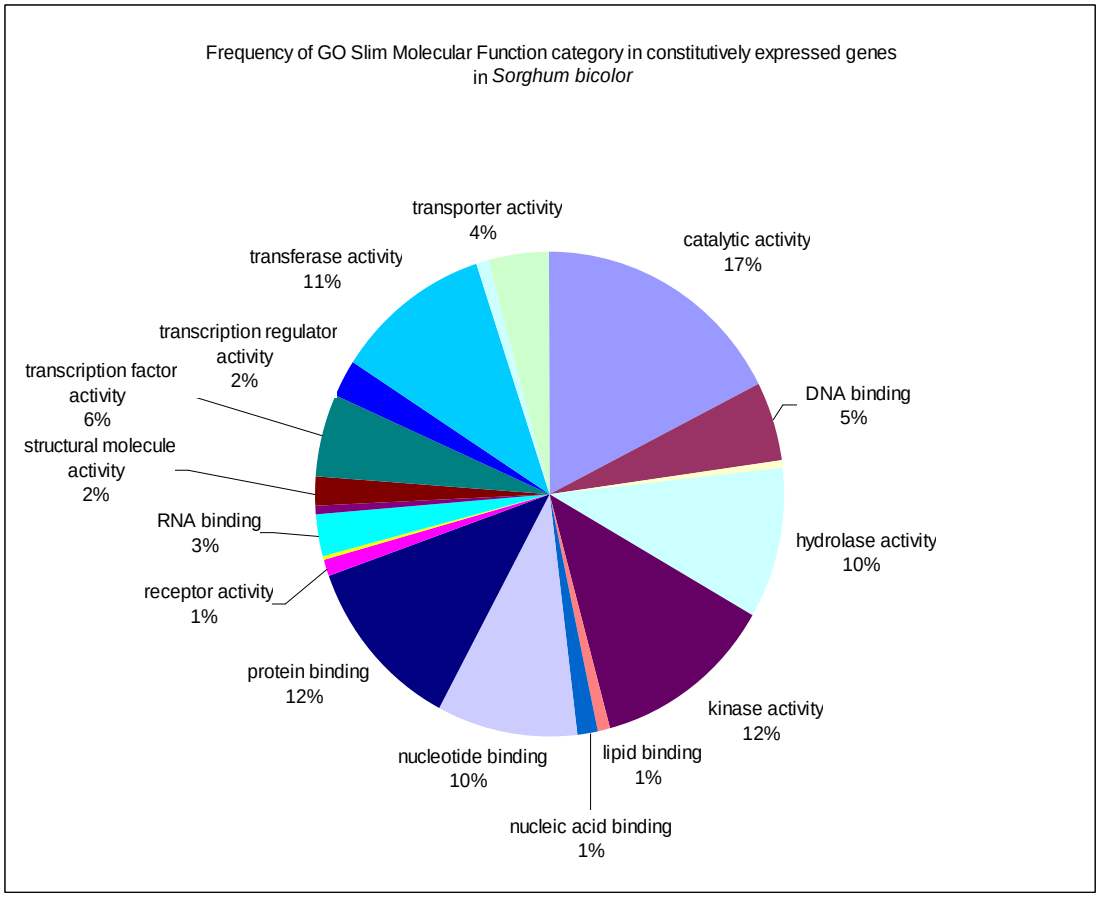


Figure 2.5. Gene ontology annotation (GO Slim, molecular function category) of the *Sorghum bicolor* constitutively expressed control set of uniscripts, based on the annotation of their orthologs in rice.

Identification of orthologous gene sets in the grasses

In order to test for signals of selection in the BSU uniscripts, we attempted to identify orthologs in other cereal species for for each of the 773 BSU uniscripts (see Methods), and used codon-based maximum likelihood tests of neutrality and positive selection. From these sets, we selected for further analysis genes that met the following criteria:

1) Orthologs were available from at least four Poaceae species.

2) Best BLASTn hits showed moderate divergence, *i.e.* BLASTn E values greater than 10^{-50} but smaller than 10^{-5}. Genes with E values smaller than 10^{-50} were considered too conserved, and therefore unlikely to have many sites under positive selection, as previous analyses indicated (A. Zamora, unpublished results). Additionally, genes with E values greater than 10^{-5} are usually too divergent, making their putative orthologs unreliable and very difficult to align due to large indels or numerous substitutions, and were not considered further.

3) Matches to partial or full-length mRNAs of other species were found. This condition expedited the identification of other sequences needed to reconstruct the entire coding sequence, critical for the necessary confirmation of orthology via reciprocal best hits.

Application of these criteria yielded 140 genes for further analysis. For each of these candidates, we attempted to reconstruct the full-length ORFs by additional BLASTn searches against the Genome Survey Sequences (GSS) and EST_others databases, *i.e.* genomic and expressed sequences (NCBI, 06/2005). Moreover, these additional sequences allowed us to empirically verify the gene prediction made by FGENESH by comparing the genomic and expressed sequences. Evidence of a role in plant defense in any of the species was also used to assign priority in the reconstruction of the set of orthologs, since this would indicate a conserved function in

different species and *prima facie* evidence of a correct annotation as a disease response gene in sorghum and other cereals. Only sets with complete ORFs were used to test for positive selection, as this is one of the conditions needed to assure orthology and a good alignment. We were able to reconstruct a total of 10 sets of orthologous ORFs using this strategy.

Evolutionary analysis of the orthologous gene sets identifies 6 genes under positive selection

The 10 sets of orthologues were tested for evidence of positive selection using various models in the PAML package (Table 2.1). Six of the ten genes were moderately divergent also at the protein level (BLASTp $E<10^{-50}$) when comparing sorghum to rice and all of them have sites predicted to be under positive selection (Table 2.1). On the contrary, four genes were not as divergent when the complete peptide was used and had BLASTp E values greater than 10^{-50} when compared to rice proteins (Table 2.2).

Although three of the more conserved genes, 2_12005, 2_6529 and 2_11011, have 50% of their ESTs coming from the Biotic Stress group of EST libraries (Fig. 2.6), their annotation does not suggest a role in disease resistance mechanisms and were found to be under strong purifying selection only (Table 2.1). Another moderately conserved gene, 2_9185, appears to be evolving primarily under purifying selection as well. However this gene is highly expressed (Figure 2.2) and 74% of its ESTs come from biotic stress libraries (Fig. 2.6). Additionally, this gene belongs to a family that has been implicated in disease resistance (Langlois-Meurinne, Gachon, and Saindrenan 2005; Wisser et al. 2005).

Table 2.1. Maximum likelihood estimations of omega (=dN/dS) for the Biotic Stress Upregulated and moderately divergent candidate genes

Gene	GO function	Species	M0 *vs.* M3		M7 *vs.* M8		
			ω	prob	Ω	prob	P1
2_8981	Thaumatin-like protein	Sb,So,Zm,Ta,Hv, As,Sc, Os	**1.77**	**3.3E-94**	**2.17**	**0.0028***	2.7
2_11684	Mlo 7 transmembrane	Sb, Zm,Hv,Ta,Os	**3.08**	**1.0E-42**	**3.72**	**0.0002***	0.69
2_4543	MADS-box transcription factor	Sb,Zm,Ta,Pm, Os, Bd	**2.07**	**5.5E-20**	**2.26**	**0.0011***	1.8
2_7192	Peroxidase	Sb,Zm,Ta,Hv,Os	**3.35**	**9.4E-51**	**4.30**	**2E-06***	2.7
2_8705	Putative fungal killer protein	Sb,Sof,Zm,Hv, Ta,Os, Bd	**3.19**	**1.7E-35**	**2.99**	**6.8E-10***	7.4
2_3804	EIF gamma	Sb,Zm,Ta,Hv,Os	**1.73**	**6.2E-52**	**3.17**	**0.001***	0.56
2_12005	EF hand domain	Sb,Sof, Zm,Ta,Hv,Os	0.78	**3.8E-53**	1.23	0.40	0
2_6529	Phosphoglucomutase	Sb, Zm,Os	1.02	**1.6E-10**	0.97	0.96	0
2_11011	Phosphoglycerate mutase	Sb,Zm,Os	0.49	**1.2E-85**	1.47	0.21	0
2_9185	Glycosyl transferase	Sb, Sof,Zm,Ta,Os	0.4	**8.7E-08**	0.4	0.7	0

M0: one ratio; M3: discrete including a class with positive selection; M7: 10 classes within beta; M8: 10 beta plus one under positive selection. ω = dN/dS for sites under positive selection. prob: probability of LRT in Chi square. P1: proportion of sites with posterior probability >0.90 of belonging to ω>1. Sb: *Sorghum bicolor*; Sof: *Saccharum officinarum*; Zm: *Zea mays*; Ta: *Triticum aestivum*; Hv: *Hordeum vulgare*; Os: *Oryza sativa*; Pm: *Panicum maximum*; Sc: *Secale cereale*; As: *Avena sativa;* Bd: *Brachypodium dystachion.* * Genes showing statistically significant evidence of positive selection for protein differentiation, after Bonferroni correction.

Table 2.2. Uniscripts tested for positive selection in ML dN/dS ratio tests and their values for the parameters used in the identification of interesting candidates

Uniscript	Total ESTs	Biotic Stress ratio	Match length aa	E value	Rice best hit
Positively selected genes					
2_8705	14	0.714	110	3E-36	LOC_Os05g38040.3expressed protein
2_8981	17	0.706	157	2E-49	LOC_Os12g43380.1Thaumatin-like protein precursor
2_4543	3	1	66	4.00E-07	LOC_Os01g69850.1SRF-type transcription factor family
2_7192	8	0.625	180	3E-55	LOC_Os07g48020.1Peroxidase 2 precursor
2_11684	5	0.5	169	6.00E-50	LOC_Os03g03700.1 MLO protein homolog 1
2_3804	14	0.575	118	6.00E-46	LOC_Os06g48350.1Eukaryotic translation initiation factor 5
Average	10	0.69	133.3	6.7E-08	
SD	5.8	0.17	43.2	1.6E-07	

Table 2.2 (Continued)

Non positively selected genes					
2_12005	5	0.6	157	4.00E-85	LOC_Os08g04630.1EF hand family protein
2_6529	7	0.571	131	6.00E-66	LOC_Os10g11140.1Phosphoglucomutase
2_11011	4	0.75	149	4.00E-76	LOC_Os03g21260.12,3-bisphosphoglycerate-independent phosphoglycerate mutase
2_9185	35	0.743	126	4E-68	LOC_Os02g38140.1 beta 1,4 N-acetylglucosaminyltransferase
Average	12.75	0.67	140.7	1.5E-66	
SD	14.88	0.09	14.7	2.9E-66	

All these uniscripts had multiple sequences from different species and were used in codon based ML tests of neutrality and positive selection, those in bold have amino acid sites that have evolved under positive selection.

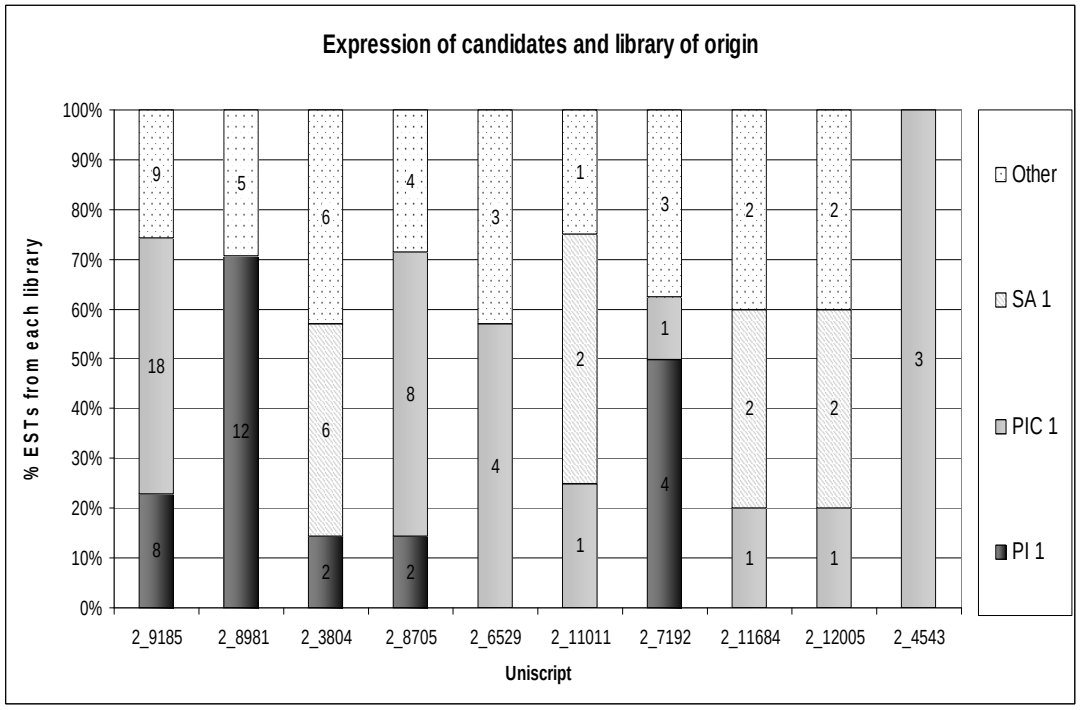


Figure 2.6. Uniscripts having multiple sequences available at NCBI for different cereal species. All these have at least 3 ESTs per uniscript, of which at least 50 percent come from the Biotic Stress subgroup (PI1 from resistant plants, PIC1 from susceptible plants and SA1 plants treated with salicylic acid).

The comparison of model M3 with M0 (Table 2.1) shows that all ten genes have variation for dN/dS (ω) ratios among their codons, i.e., selective constraints are heterogeneous, as expected. For the four conserved genes, model M8, which allows for some sites to have ω greater than 1.0, was not better than M7, indicating that all amino acids belong to classes ranging from strict purifying selection to neutrality. For the six moderately divergent genes, the M8 model was significantly better than M7, providing strong evidence for positive selection at some codons. We also used the FEL and REL algorithms from HyPhy (Kosakovsky Pond and Frost 2005) and found evidence of positive selection only for the peroxidase set. Two sites, 105E and 117Q had dN-dS values of 2.24 and 1.15, respectively. These two sites were also identified using codeml as positively selected with posterior probabilities of 0.95 and 0.99, but three other sites identified consistently and with high posterior probability (>0.90) were not identified as such by HyPhy. All other sets of genes were found to have sites evolving only under significative negative selection by HyPhy.

Frequency of adaptive events in plant disease response genes

Out of the six genes that showed evidence of positive selection, an average of 5.3 codons per gene (SD = 3.0) had posterior probabilities greater than 0.90 of belonging to the class of sites with $\omega > 1$. Thus, taking gene size into account, 2.34% of the codons per gene are estimated to have evolved adaptively in this set of genes. In total there are 32 codons with $\omega > 1$, and an average of 4.2 different amino acids per site. Assuming that a single non-synonymous substitution lead to the observed residues, there are a total of 134.4 fixed substitutions for the six genes, which divided by the sum of the length of the branches in the phylogeny (172-253MY, Figure 2.1)(Ramakrishna et al. 2002; Paterson, Bowers, and Chapman

2004; Rice Chromosome 3 Sequencing 2005), gives an estimated 0.53-0.78 adaptive substitutions per million years, i.e. approximately 5 to 8 adaptive mutations becoming fixed every 10 million years across all species used in the analysis. Also, if there are 2.34% adaptive codons per gene, 4.2 substitutions/codon, 300 amino acids in the protein and ca. 12.000 genes involved in disease resistance, there should be approximately 1.39 to 2.05 adaptive mutations fixed every one thousand years in any of the lineages in the phylogeny (Fig. 2.1). This analysis is only taking into account the replacement substitutions and not silent substitutions, which could also be adaptive (Komar, Lesnik, and Reiss 1999; Kimchi-Sarfaty et al. 2007). Additionally, there are many insertions and deletions in the multiple species alignments and it is likely that some of them could have a functional effect. Finally, there is a chance that changes in regulatory regions, which haven't been covered here, could also have an important effect in increasing the adaptive mutation rate to levels similar to those of pathogens (Herring et al. 2006; Taubes 2008).

Characteristics of the positively selected BSU genes

Three of the genes identified here as evolving adaptively are known disease response genes. Thaumatins, peroxidases and barley *mlo* homologs have been previously shown to undergo changes in expression due to pathogen attack and to have a role in disease resistance (see discussion). Although both thaumatin (2_8981) and peroxidase (2_7192) genes are mainly expressed in resistant plants (Pathogen Incompatible, PI1 library; Figure 2.6), the former has a higher level of expression, which may suggest rapid response to pathogen stimuli or genetically determined upregulation during early seedling development. In fact, for uniscript 2_8981, 12 ESTs came from the PI1 library and 5 from other unrelated libraries (Steckel reliability value, R = 28.2 (Pratt et al. 2005)), indicating a low to moderate

level of constitutive expression that increases significantly after inoculation with the pathogen. Since these ESTs were found only in challenged but healthy plants, it is arguable that they might play a significant role in resisting the pathogen's attack. Although there are many paralogs for both thaumatins and peroxidases in the genome of sorghum, only two of each of these genes appear to be upregulated in the available libraries.

The *mlo* homolog (2_11684) was the only one of these known disease response genes expressed in SA treated plants. In the genome of *S. bicolor*, 2_11684 is a member of a small family with 12 genes. This gene is a singleton in a chromosomal region, just like other seven members in the family, while other 4 genes are clustered together (Phytozome; Sbi_0.29522). Remarkably, 2_11684 has a large number of small internal exons surrounded by large introns, suggesting the possibility of alternative splicing. The other three positively selected genes have not been previously implicated in disease responses and we present additional information on them next.

2_8705: SESPY, a new putative disease response gene

This single copy gene is highly upregulated in *Sorghum bicolor* after inoculation with *C. sublineolum* and is very divergent with respect to its rice ortholog (Fig. 2.2). Ten of its 14 ESTs come from the Biotic Stress libraries, particularly from susceptible plants (Fig. 2.6). Maximum likelihood analysis at this gene identified four main regions of high amino acid sequence conservation, where most codons evolve under purifying selection (Fig 2.7). Two highly conserved regions have the sequence SESPYFGSSVHYG (located at residues 100-110) and ATRGDWWQGSLYY at the C-terminus of the protein. However, there are 9 hyper-variable sites (posterior probability > 0.90), making SESPY the gene with the

most sites predicted to have evolved under selection in our analysis. Positions 51 and 93, for instance, occur amidst regions of strict conservation across all species of grasses strongly suggesting a functional role.

Figure 2.7. Multiple protein sequence alignment of *S. bicolor* 2_8705 uniscript and orthologs in other grass species, from position 47 to 126 out of 162 amino acids. Positions with posterior probability greater than 0.90 of having evolved under positive selection are 51, 85, 93, 120,124 and 126 (in boxes). Positions 68 to 83 were not included in the analysis.

.

2_4543: SRF type transcription factor

This gene is a member of the MADS-domain family, which encode eukaryotic transcription factors with sequence specific DNA binding characteristics. It belongs to the Myocyte Enhancer Factor 2 (MEF2)-like subfamily of MADS genes, the plant specific MIKC-type, and includes a K-box domain. This family of eukaryotic transcription factors include a DNA-binding and dimerization domain and has been shown to be important in homeotic regulation in plants (Marchler-Bauer et al. 2009) (NCBI, Conserved Domain Database, 07/2007). The rice gene Os01g0922800 (RefSeq peptide NP_001045235.1) is the ortholog of *Sorghum bicolor* 3g044170 gene model (Phytozome) that corresponds to the uniscript 2_4543. There are many paralogs in rice and *Arabidopsis* and the few whose functions are known are involved in flower organ specification (Arora et al. 2007). Arora et al. (2007) studied the expression profiles of the MADS-box genes in rice and *Arabidopsis* and found that most of the basic structures and functions are conserved, but there is evidence of new function acquisition, duplication and sub-functionalization. There is no evidence in those model genomes of involvement of the ortholog of 2_4543 in disease response, and thus, it may be that the three ESTs found only in the PIC1 library were there just by chance. This gene is overall the most divergent of those presented here (Table 2.2) and although it shows high conservation in the amino terminal portion of the protein that corresponds to the DNA binding MADS box domain, the maximum likelihood tests suggest there are 3 codons that have changed repeatedly through non-synonymous substitutions along the protein. The carboxyterminal portion of this protein is hypervariable and was not included in the maximum likelihood analysis due to the difficulty of obtaining a satisfactory alignment and to the false positive errors this can generate. Such variability could reflect either a lack of constraint, positive selection or both.

2_3804: Eukaryotic translation initiation factor (eIF5)

In *S. bicolor,* 57.5% of the 14 ESTs that make up this uniscript come from the Biotic Stress libraries (Table 2.2) and most of them were found in the salicylic acid (SA1) library, suggesting a change in expression response due to this plant defense signal molecule (Fig. 2.6). 2_3804 is an eukaryotic translation initiation factor-5 and contains a putative zinc binding C4 finger and the InterPro domains Armadillo-type fold (InterPro:IPR016024) as well as translation initiation factor IF2/IF5, N-terminal (InterPro:IPR016189).

Positively selected sites map to surface of the thaumatin and peroxidase 3D structures

Crystal structures have been determined for protein homologues of two of the genes identified here as having evolved under positive selection, namely the thaumatin (PDB: 1RQW) and peroxidase (PDB: 1H5D) genes. Figures 2.8 and 2.9 show models of the sorghum proteins based on these available structures (see Methods). The Root Mean Square (RMS) is 0.00Å (using 150 α carbons) for thaumatin and 0.10 Å (with 264 α C) for peroxidase, which means that the protein sequences aligned very well to their respective homolog's structure, since the alpha carbons of both are very close in space. Moreover, the final total energy of the models was -7080.58 KJ/mol for thaumatin and -5223.24 KJ/mol for peroxidase, and there were no amino acids clashing with each other. We used these structures to identify the position of the positively selected amino acid residues in the tertiary structure. In both cases all the amino acid sites identified as being under positive selection are located on the surface of the protein, where they may be in contact with the solvent and other molecules. In the case of the thaumatin protein, most of the positively selected residues are located in the loops of the structure and on the

fringe (Fig. 2.8 A,B), towards the concave side of the protein (Fig. 2.8 C). No positively selected sites were found either on the central concave portion or on the opposite side of the protein (Fig. 2.8 D). The core of thaumatin-like proteins is made up by the central portion of the peptide sequence and folds into 11 beta-sheets (De Vos et al. 1985). Shatters et al. (2006) found that this core is highly conserved and that most of the variation observed across very divergent species is found in the surface exposed loops, which is consistent with our determination of the position of putatively selected residues in these regions of the protein (Figure 2.8).

The structure and biochemistry of the horse radish peroxidase C has been extensively studied and the location of the active site pocket and key functional residues are known (Gajhede et al. 1997). A hypervariable region between alpha helices F and G, the F' and F'' helices, identified in Class III peroxidases (higher plant peroxidases, (Gajhede et al. 1997) (corresponding to residues 153-180 in our model) appears to be conserved in our set of species (Fig. 2.9). However, we found a hypervariable region, the E helix, containing three amino acid sites with posterior probabilities greater than 0.90 of being under positive selection and in which conserved and variable residues alternate. This helix is located to the side of the active site and several of these amino acid sites interact directly with the heme group (Gajhede et al. 1997). Additionally, in the peroxidase (Fig. 2.9), seven of the eight positively selected sites are located within 20Å of the oxygen atoms in the propionate side chain of the tetrapyrrole or heme ring (in red), and primarily towards one side of the protein.

Figure 2.8. Modeled *Sorghum bicolor* 2_8981 thaumatin using Swiss Deep View and P-Modeller. A: front, B: left side, C: right side and D: back, views of the same protein model. All the amino acid residues inferred to have evolved under positive selection (yellow) occur in the surface of the protein, exposed to the solvent and available for contact with other proteins, as opposed to buried in the core. Thaumatins are abundant and easily crystallized, therefore several x-ray high resolution structures are available and were used to model 2_8981.

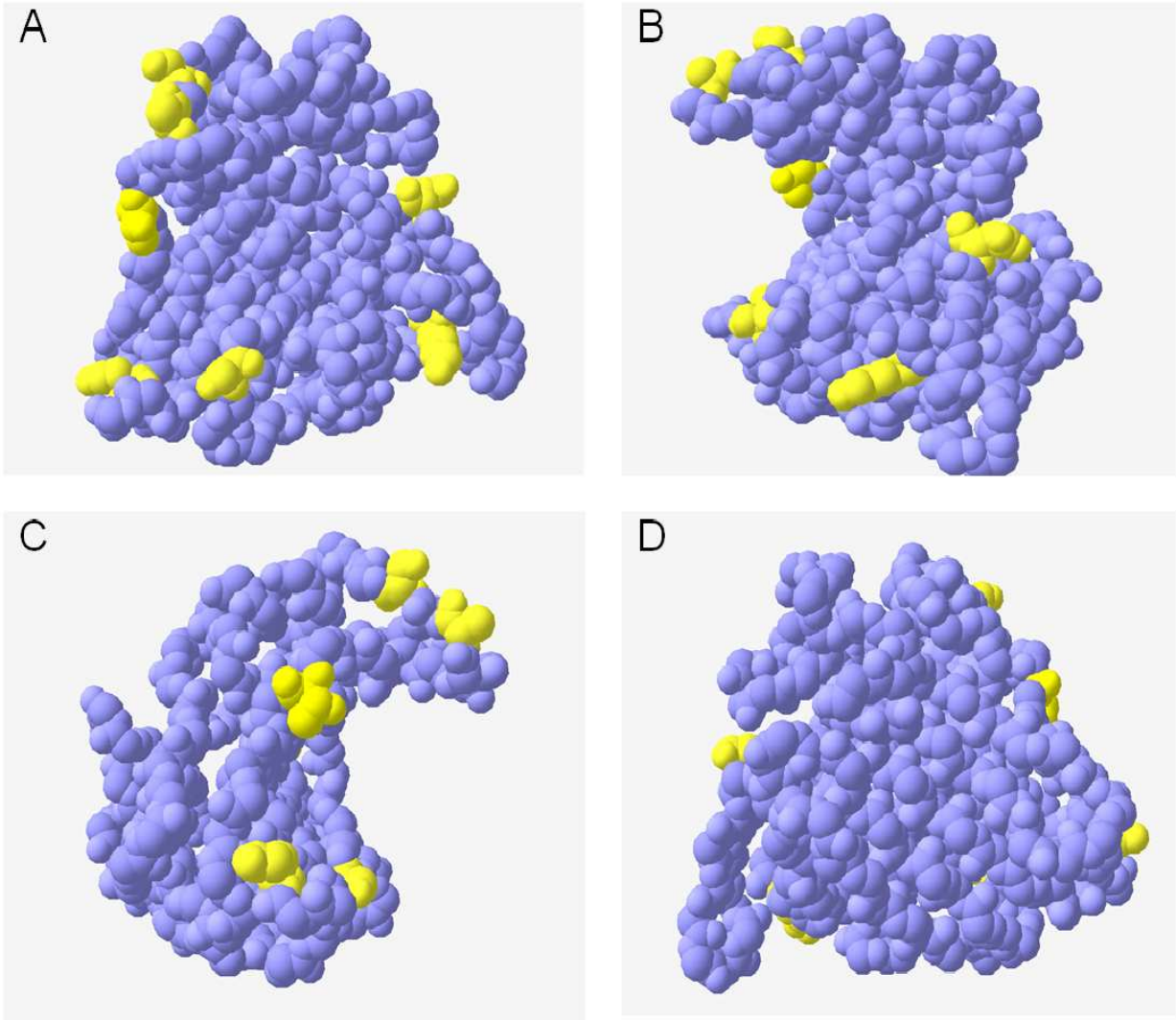
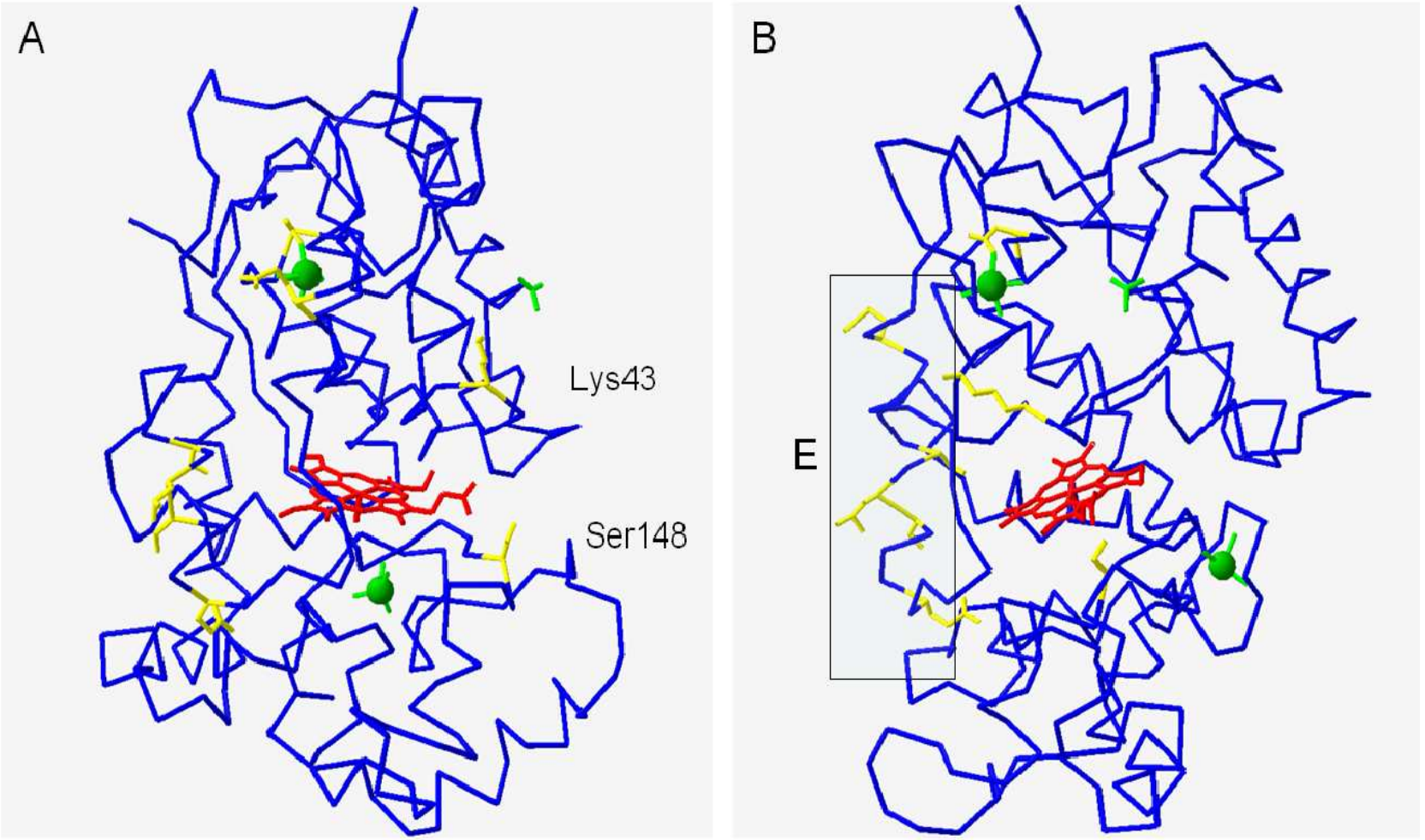

A
B
C
D

Figure 2.9. *Sorghum bicolor* 2_7192 peroxidase modeled with Swiss Deep View and P-Modeller using 1H5D including HEM. All the codon sites with posterior probabilities greater than 0.9 of having evolved under positive selection (side chains in yellow) occur on the surface of the protein. Two of the sites (Lys43, Ser148) are located at the edge of the active site cleft and several others are located 90 degrees from the active site in alpha helix E (box) that may be in contact with the tetrapyrrol ring of this metalloprotein (in red). The green balls are calcium ions. An acetate molecule is also shown (light green stick structure). **B** shows the same protein with a 90 degree clockwise rotation with respect to **A**.


A
Lys43
Ser148
B
E

Four residues, Gln117, Ser121, Leu122 and Ser124, are located in alpha helix E (Fig. 2.9B) located 90º to the side of the catalytic site and some amino acids in this alpha helix may be in contact with and help coordinate the *heme* group, due to the positioning of the *heme* and in particular due to the proximity of that helix to the methyl and vinyl side chains of one of the pyrrole rings. Remarkably, in the sorghum peroxidase, two positively selected sites, Lys43 and Ser148, are located at the edge of the active site cleft, 8Å and 6Å, respectively, from the oxygen atoms in the propionate side chain, suggesting that they may be important in the catalytic activity of the protein (Fig. 2.9). Moreover, an additional positively selected site is 7Å from one of the calcium ions, essential for catalysis in peroxidases.

Assessing the effects of including paralogs instead of orthologs in the codon based evolutionary analysis

Two of the genes identified as having evolved under positive selection are members of gene families that include copies showing high similarity, a feature in common with many other plant genes. Peroxidases and thaumatins have paralogs in multiple different positions in the genomes of plants, and can be arranged as clusters or single copy genes. Thus, unknowingly using paralogs in the evolutionary analyses could cause over or under-estimations of the number of genes found to be under positive selection, as well as of the residues involved in that process. Including paralogs could also result in the identification of signals of adaptation due to sub-functionalization, instead of adaptation (in the form of increased resistance to the pathogen) due to changes in true orthologous genes, the main goal of this study. In order to assess these possibilities, we first used the information available from the rice and sorghum genome projects (Gramene; Phytozome) to determine whether the genes used from these two species were located in syntenic positions

Figure 2.10. Comparison of the orthologous regions in the genomes of rice (orange) and sorghum (green), Gramene, October 14, 2008. This is a comparison of rice chromosome 7 and sorghum chromosome 2. It shows that each one of the 5 *S. bicolor* peroxidases in this cluster is the best reciprocal ortholog with each of the 7 paralogs in rice. To the sides of the cluster of paralogs, there are three non-peroxidase genes showing a one to one relationship or orthology, as measured by reciprocal best hits and synteny. This confirms that the Sb peroxidase and the Os peroxidase used are in fact the reciprocal best hits and that they are located in syntenic positions.

.

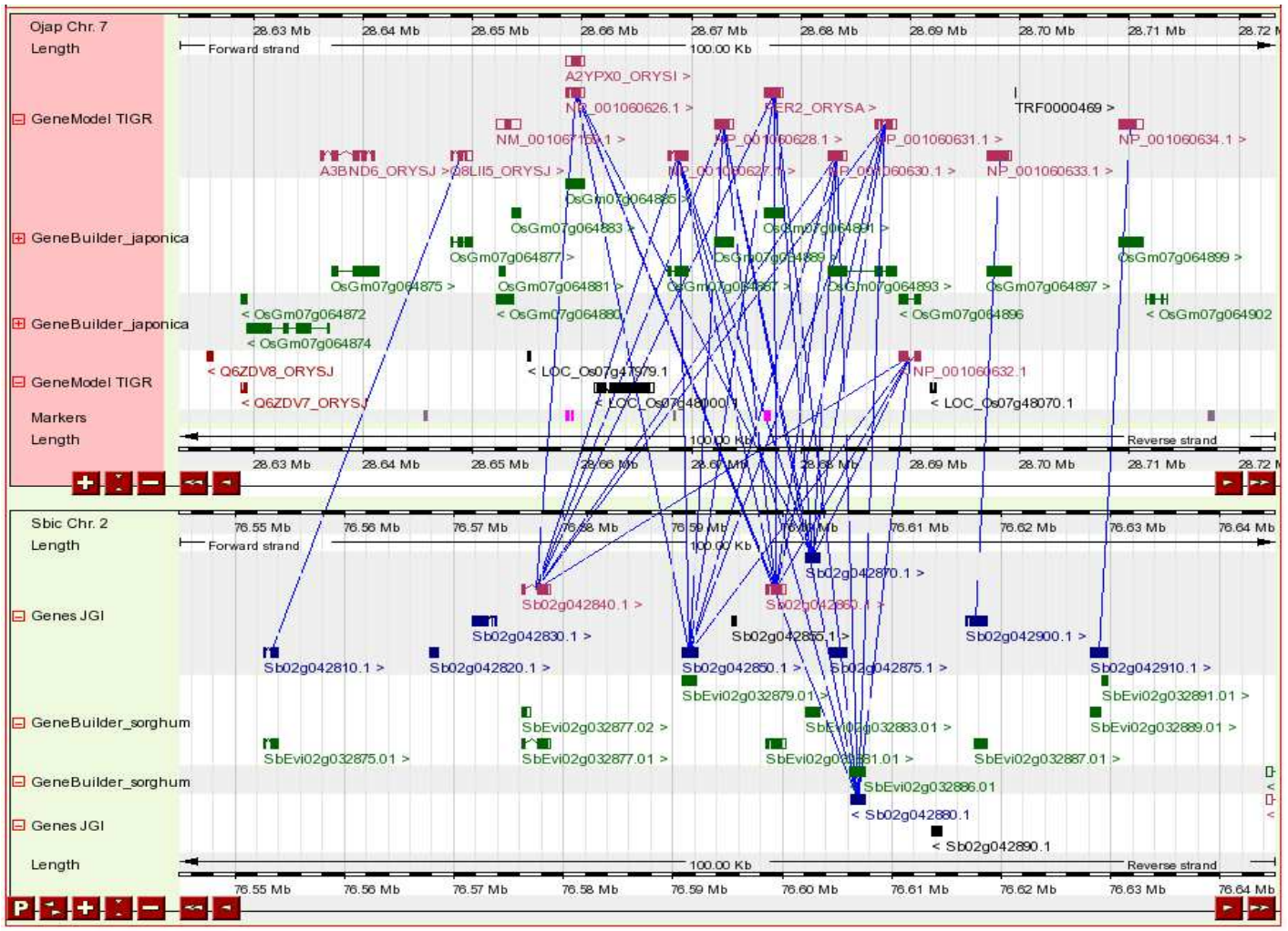
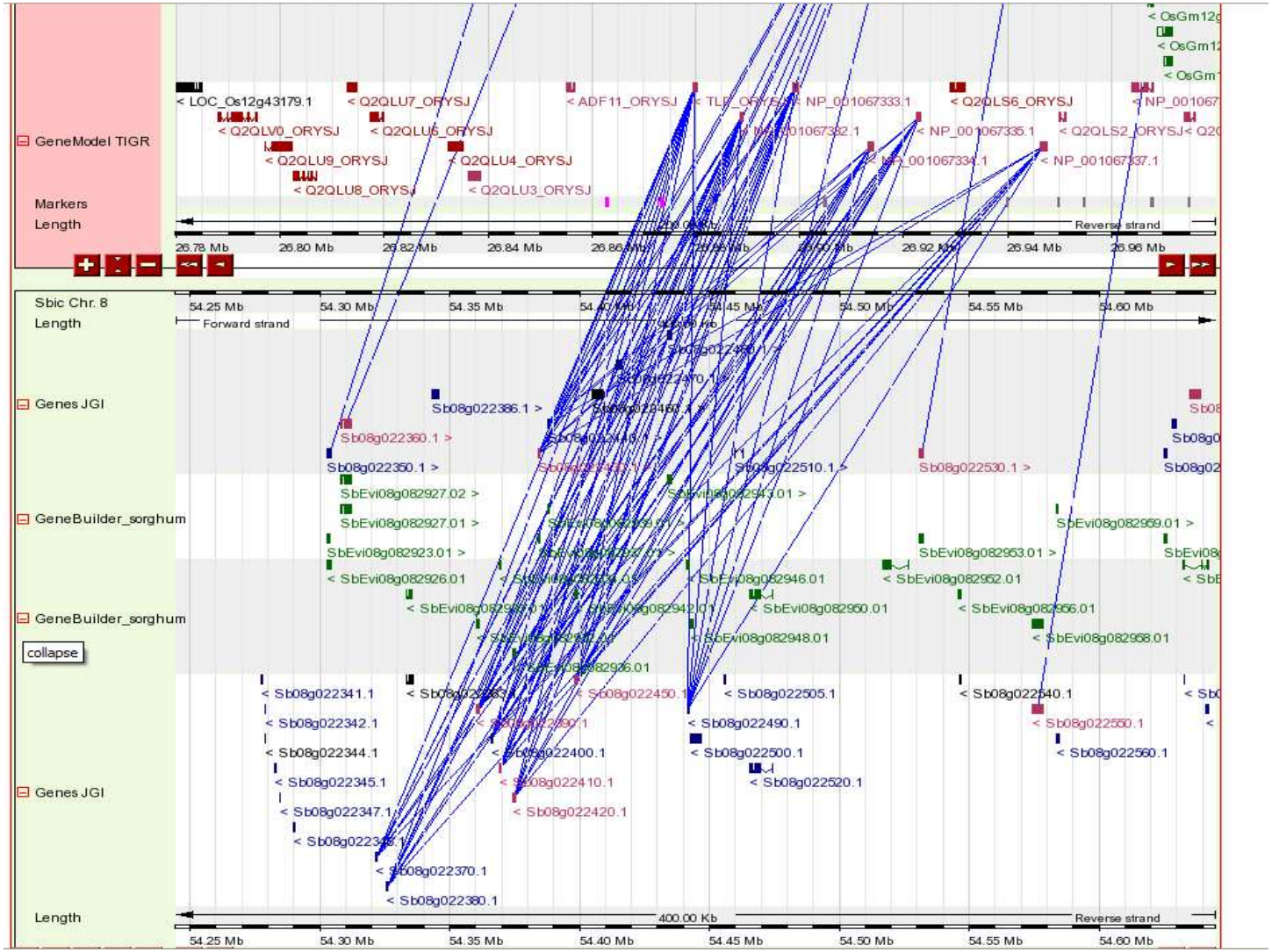

Ojap Chr. 7
Length
Forward strand
100.00 Kb
GeneModel TIGR
GeneBuilder_japonica
GeneBuilder_japonica
GeneModel TIGR
Markers
Length
Reverse strand
Sbic Chr. 2
Length
Forward strand
Genes JGI
GeneBuilder_sorghum
GeneBuilder_sorghum
Genes JGI
Length
Reverse strand

Figure 2.11. Thaumatin clusters in rice (orange) and sorghum (green) . This snapshot from the September 08 release of Gramene, is showing the comparison of rice chromosome 12, on top, and sorghum chromosome 8, below, and it shows that the gene to the right (Sb08g22550), for example, is a single copy gene in sorghum and it has a single ortholog in rice. However, each one of these 10 sorghum paralogs located in this cluster has a line to 7 genes in rice, and vice-versa. This indicates that the closely related paralogs in these clusters in the genomes of sorghum and rice have a many to many relationship of orthology


GeneModel TIGR
Markers
Length
Reverse strand
26.78 Mb
26.80 Mb
26.82 Mb
26.84 Mb
26.86 Mb
26.88 Mb
26.90 Mb
26.92 Mb
26.94 Mb
26.96 Mb
< LOC_Os12g43179.1
< Q2QLU7_ORYSJ
< ADF11_ORYSJ
< NP_001067333.1
< Q2QLS6_ORYSJ
< Q2QLV0_ORYSJ
< Q2QLU6_ORYSJ
< NP_001067335.1
< Q2QLS2_ORYSJ
< Q2QLU9_ORYSJ
< Q2QLU4_ORYSJ
< NP_001067337.1
< Q2QLU8_ORYSJ
< Q2QLU3_ORYSJ
Sbic Chr. 8
Length
Forward strand
54.25 Mb
54.30 Mb
54.35 Mb
54.40 Mb
54.45 Mb
54.50 Mb
54.55 Mb
54.60 Mb
Genes JGI
Sb08g022386.1 >
Sb08g022360.1 >
Sb08g022350.1 >
Sb08g022510.1 >
Sb08g022530.1 >
GeneBuilder_sorghum
SbEvi08g082927.02 >
SbEvi08g082927.01 >
SbEvi08g082923.01 >
SbEvi08g082959.01 >
SbEvi08g082953.01 >
< SbEvi08g082926.01
< SbEvi08g082946.01
< SbEvi08g082952.01
< SbEvi08g082950.01
< SbEvi08g082956.01
< SbEvi08g082948.01
< SbEvi08g082958.01
GeneBuilder_sorghum
collapse
< Sb08g022341.1
< Sb08g022342.1
< Sb08g022344.1
< Sb08g022345.1
< Sb08g022347.1
< Sb08g022450.1
< Sb08g022505.1
< Sb08g022490.1
< Sb08g022500.1
< Sb08g022520.1
< Sb08g022540.1
< Sb08g022550.1
< Sb08g022560.1
< Sb08g022400.1
< Sb08g022410.1
< Sb08g022420.1
< Sb08g022370.1
< Sb08g022380.1
Genes JGI
Length
400.00 Kb
Reverse strand
54.25 Mb
54.30 Mb
54.35 Mb
54.40 Mb
54.45 Mb
54.50 Mb
54.55 Mb
54.60 Mb

We then implemented a paralog permutation strategy to evaluate the robustness of the analysis to the use of increasingly divergent paralogs. In both peroxidase and thaumatin, the genes used in the previously described analyses are located in the corresponding regions in the genomes of rice and sorghum (Figures 2.10 and 2.11). Both kinds of genes are arranged locally as small clusters in both genomes: peroxidases in chromosomes 7 (28.67Mb) and 2 (76.6Mb) of japonica rice and sorghum, respectively; and thaumatins in chromosomes 12 (26.8Mb) and 8 (54.4Mb) in the same order. Within these clusters, each of the copies in rice are the best reciprocal ortholog, according to BLAST searches (precomputed at Gramene), to all the copies in the sorghum cluster, and vice-versa, i.e, members of a cluster are very similar to each other and equally divergent to each one of the copies in the other species (Figure 2.10 and 2.11).

Interestingly, however, a phylogeny of the peroxidase genes from several cereal species (Figure 2.12), including genes located in clusters found in syntenic positions in sorghum and rice (including 2_7192, the sorghum biotic stress upregulated gene) shows that the genes at both flanks of the clusters are more similar to its ortholog in the other species, than to genes within its own genome, i.e. the genes at the edges of the cluster have a 1 to 1 relationship of orthology, while those in the center of the cluster vary in their similarity. Genes in positions 1 and 2 from left in both species' clusters, as well as those in the right flank of the cluster have a single ortholog. The sorghum gene that corresponds to the 2_7192 uniscript is located in the center of the cluster (Sb02g042860.1) and it has the same position, third from left, as the rice gene used in the initial comparison (Os_NP_001060628.1). However, this rice gene clusters only with its neighbor, Os_NP_001060629.1, suggesting a recent duplication or concerted evolution. The next genes from left, Sb02g042870.1, in sorghum, and Os_NP_001060631.1, in rice, appear in distant positions in the

phylogeny. Additionally, there are two gene fragments or pseudogenes in sorghum, (Sb02g042855.1 and Sb02g042875.1) amidst the complete homologous ORFs. Hence, the sorghum and rice genes used in codeml evolutionary analyses are the correct orthologs based on reciprocal BLAST best hits, synteny and genealogy.

Since the method used to identify the orthologs in these two species was the same used to identify orthologs in other cereal species, it may be safe to argue that there is a low probability that we used paralogs in the analyses. Nevertheless, it is a possibility since we didn't have enough information to determine synteny for the genes of other species. Therefore, in order to assess the effects of using of paralogs on the number of positively selected sites, we conducted a series of permutations with increasingly divergent paralogs and compared the results. To do this, we used the peroxidase phylogeny (Figure 2.12) to generate two different data sets by first, fixing the orthologous genes for sorghum and rice, and changing the other 3 sequences (Paralogs Set 1, PS1, Supplementary Data: Alignments) and, second, leaving only the sorghum gene unchanged (Paralogs Set 2, PS2). The first data set (PS1) included the most divergent sequences from each species, as indicated by their long branches, but found in the position in the genealogy corresponding to the species phylogeny (Figure 2.12). The second data set (PS2) included more divergent paralogs, located in outer branches of the phylogeny, which do not correspond to the species tree and that can be expected to have diverged faster or for a longer time (Figure 2.12). We used the full length alignment of the coding regions in all the replicates for the maximum likelihood analyses for these two sets, without eliminating the regions with gaps, but we visually inspected and modified the Clustal W alignments to improve them. We then compared the results of these two sets with those from the set of orthologs without gaps (No Gaps (NG), average of 7 different runs, Supplementary Data: Comparison of codeml results for 2_7192).

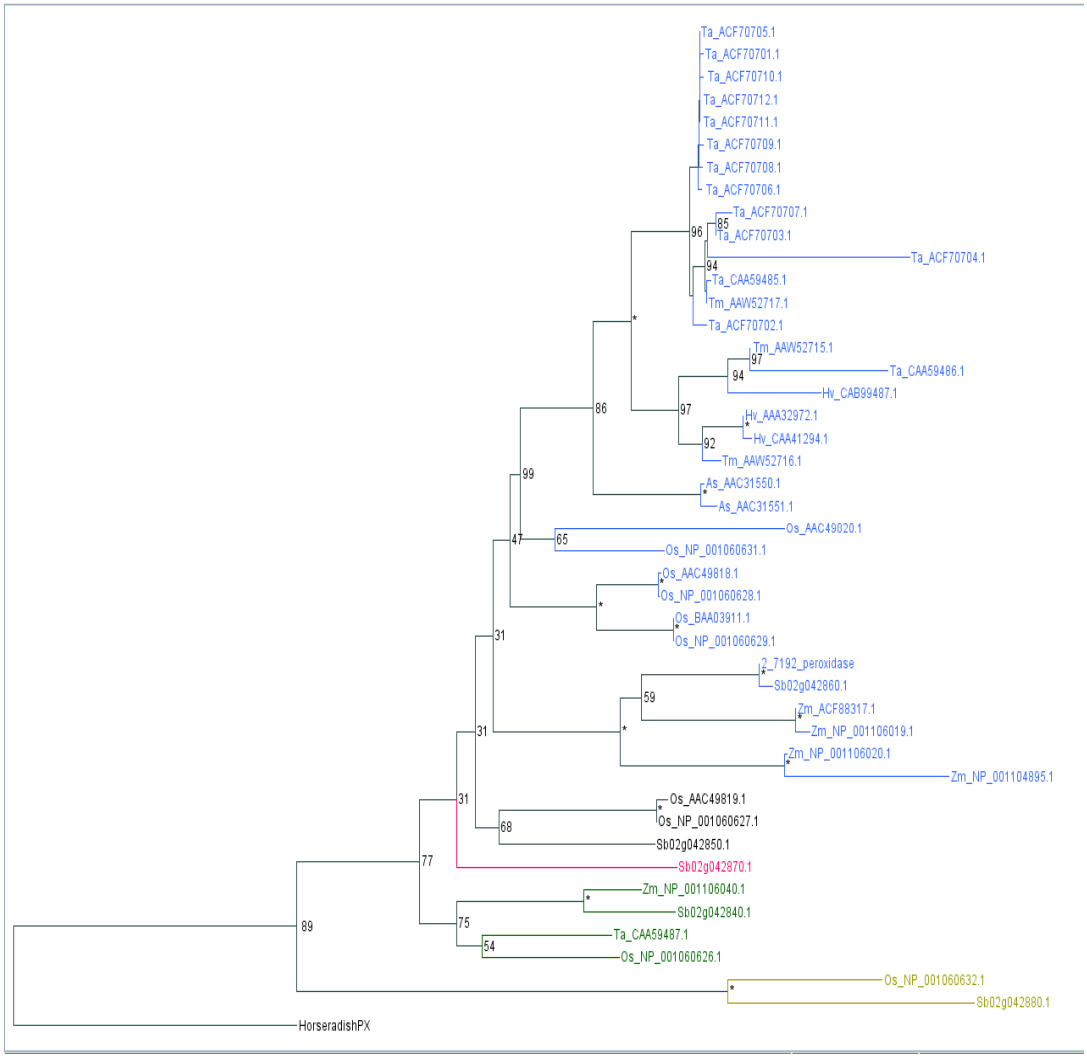


Figure 2.12. Paralogous and orthologous peroxidases from cereals, including those from the *S. bicolor* and *O. sativa* orthologous clusters shown in Figure 2.6. Numbers at the nodes are the bootstrap values for 2000 replications (*: 100). Ta: *Triticum aestivum*; Tm: *Triticum monococcum*; Hv: *Hordeum vulgare*; As: *Avena sativa*; Os: *Oryza sativa*; Zm: *Zea mays*; Sb: *Sorghum bicolor*.

Remarkably, there is a greater similarity between rice and sorghum at the genes located at the left (Os NP_00106026 and Sb02g042840, green) and right side (Os NP_001060632 and Sb02g042880, gold) of the cluster (Fig. 2.10). This suggests that genes at the sides of the cluster are more conserved than those in the middle, where there is evidence of recent duplication in the rice cluster (NP_001060628 and NP_001060629<PER2_ORYSA>) and a pseudogene in sorghum (Sb02g042855).

The genes used in this phylogeny of 36 peroxidase genes from several cereal species and varying in similarity to the sorghum peroxidase (2_7192) full coding sequence (BLASTn cutoff, $E = 10^{-60}$) came from the genomes of *Triticum aestivum*, *T. monococcum*, *Hordeum vulgare*, as well as several divergent peroxidases from rice, different from those located in the cluster of orthologs to 2_7192 (Figure 2.10). The phylogeny of these genes shows, first, that multiple copies of each species are clustered together, suggesting concerted evolution, and second, that the topology of a subtree containing these clusters resembles the species phylogeny, indicating that those are the likely orthologs to 2_7192 (Figure 2.1). However, two differences with respect to the species tree exist. One difference is that *Zea mays* has two branches, although this was expected since the genome of maize suffered a duplication about the same time since the divergence of the maize-sorghum lineages. The other is that there is a subtree containing genes of *T. aestivum, T. monococcum* and *H. vulgare* arranged in two branches, which suggests the maintenance of similar but old alleles in these species of the Triticeae, perhaps due to balancing selection. The analysis of Paralogs Set 1 (PS1) showed that there were 13 and 4 putative positively selected sites with posterior probability greater than 0.5 and 0.9, respectively, compared to the 17 and 6 sites identified in the orthologs without gaps (No Gaps, NG) analysis. The slight difference is not statistically

significant (Fisher exact test, p>0.05), suggesting that the choice of sequences didn't affect the number of sites identified to be under positive selection. Remarkably, however, there were 5 sites that were identified as being under positive selection in both the NG and PS1 sets, even though the sequences and alignments used were different. Four of these 5 sites had posterior probability >0.88 in the No Gaps set and they include K43 and S148, located at the edges of the active site cleft (Fig. 2.9A), as well as Q117 and L122 located in the hypervariable E helix (Fig. 2.9B). Additionally, S246 had posterior probabilities of 0.73 and 0.95 in the NG and PS1 sets, respectively and is located in the L alpha helix (amino acid numbers refer to the No Gaps alignment).

For the most divergent paralogs (PS2), the number of positively selected sites decreased to 11 and 1 sites with posterior probabilities >0.5 and >0.9 respectively, as the average of 4 replicates with the same sequences but with different unrooted trees provided by the user. Surprisingly, there were 5 sites identified in common both with the NG set and 4 with PS1. These sites are P40, K43, Q117, L122 and S148. The amino acid residue P40 had a posterior probability of 0.74 in NG, but had an average of 0.998 in PS2, which makes it interesting, particularly due to its close position to the active site cleft. All the other residues had posterior probabilities ranging from 0.5 to 0.77 (Supplementary Data: Comparison of codeml results, Paralog Set 2).

All sites except P40, would be considered doubtful in PS2 due to their low posterior probability, were not for the fact that 5 of them were also found in the other analyses with different sequences and alignments. In total, there were 3 sites (105E, 121S, 124S) with posterior probability >0.9 in the NG set not found in either PS1 or PS2, and one site (227S) in PS1 not identified as under selection in NG. Interestingly, the three sites with posterior probability >0.9 found exclusively in the

NG set (105E, 121S, 124S) are located in the hypervariable E helix, where other two residues (117Q and 122L) were identified as positively selected in the three sets of sequences (NG, PS1 and PS2).

Discussion

In this study we identified candidate defense response genes by combining information from their expression level, proportion of sequences obtained from the Biotic Stress subgroup of EST libraries, inter-specific divergence between sorghum and rice, and prior annotation of the genes in the species involved. We then assessed whether these candidate or known disease response genes showed a pattern of substitutions consistent with adaptive evolution. We did this under the hypothesis that millions of years of antagonistic coevolution between cereal-hosts and fungal-pathogen populations may have left a signal of positive selection in these genes. Our strategy for the identification of disease response genes took advantage of the increasing abundance of genomic sequences across the grasses, as well as expression data from one or several different species. This allowed us to identify 773 genes whose expression patterns may change in response to the fungal pathogen *Colletotrichum sublineolum*, and to test a subset of ten genes for evidence of adaptive evolution. Using this approach we demonstrate for the first time that six classes of disease response genes identified here evolve adaptively, including three previously known pathogenesis-related proteins and three genes not implicated before in disease response, all of which are particularly interesting candidates for functional validation studies.

The combination of expression profiles and inter-specific divergence is a useful method in the identification of adaptively evolving disease response genes

In our initial screen, we selected genes having 50% of their ESTs coming from the Biotic Stress group of libraries and a minimum of 3 ESTs, on the assumption that genes involved in disease resistance would be enriched in this set. The comparison of the annotation for this set of genes with a control set composed of constitutively expressed *S. bicolor* genes, revealed that there are several functional categories of genes that show significantly different patterns of expression, in particular, catalytic activity, hydrolase activity and transcription were substantially reduced, suggesting a decrease in basal metabolic functions and gene expression during active defense or pathogenesis. On the contrary, kinase activity increased, suggesting active signal transduction and amplification of the signal of pathogen attack, events typical of hypersensitive responses and pathogenesis (Frye, Tang, and Innes 2000; Popescu et al. 2009). In sugarcane, kinases are one of the most abundant domains found in ESTs, many of these are found in genes involved in disease resistance and a few kinase containing genes were exclusive to biotic-interaction libraries (Vettore et al. 2003). Other gene ontology categories were present only in the BSU set, including genes with calmodulin binding activity, involved in calcium signaling which is one of the first events in pathogen detection (Heo et al. 1999); oxygen binding, probably associated to the reactive oxidative burst (McDowell and Dangl 2000); and lipases, possibly involved in the generation of salicylic acid or jasmonic acid precursors, important in signal transduction molecular mechanisms and in the establishment of systemic acquired resistance (Salzman et al. 2005). Additionally, many of the uniscripts represent genes known to be involved in disease resistance or have domains typical of that kind of genes. Some of the most important categories of genes found include: leucine rich repeat (LRR) domain proteins, kinases,

transcription factors, and several kinds of pathogenesis related proteins (PRPs) such as chitinases, glucanases, peroxidases and thaumatins (Sticher, Mauch-Mani, and Metraux 1997).

By using BLASTn and BLASTp values between sorghum and rice orthologs as a measure of divergence, we were able to identify genes that have above average divergence over the 50 million years since these two species had a common ancestor (Paterson, Bowers, and Chapman 2004) (Figure 2.1), possibly due to recurrent events of positive selection, as expected in an arms race scenario. Among these genes, maximum likelihood analyses provided statistical support for six genes with selectively driven elevated levels of amino acid substitution, consistent with our hypothesis. The accuracy and power of the kind of analysis used here has been thoroughly studied (Anisimova, Bielawski, and Yang 2001; Anisimova, Bielawski, and Yang 2002) and compared to other tests (Wong et al. 2004). The codeml maximum likelihood method has been shown to outperform other methods in several different selective scenarios and the likelihood ratio tests used to compare the models have been determined to be conservative, showing very low type I error rates (Wong et al. 2004). Additionally, since the M7 model is very flexible and includes sites evolving in a range from strict purifying selection to neutrality, the comparison of M8 vs. M7 is a very stringent test of positive selection (Anisimova, Bielawski, and Yang 2001). In addition to the codeml analyses, we tested for evidence of positive selection for these genes using the Fixed Effects Likelihood and Random Effects Likelihood methods implemented in HyPhy. We only found evidence of positive selection for the peroxidase gene using the FEL method. The sites identified by the FEL method were the same as those identified by codeml, which gives strong support to the hypothesis that peroxidases evolve under positive selection due to their importance in basal and broad spectrum disease resistance

(Johrde and Schweizer 2008). Also, the fact that this was the only gene identified by FEL as having positively selected sites is not surprising since the peroxidase gene had the highest value of ω in codeml, and therefore the strongest signal of positive selection, even though only five taxa were used in the alignment. Additionally, all codon-based maximum likelihood methods for detecting positive selection have low power when the number of taxa is small, but the discrepancy between codeml and the methods implemented in HyPhy suggest either lower power in HyPhy or to higher false positives in codeml. Since there was an agreement in the sites identified in the peroxidase gene using different sets with codeml, but HyPhy found only the two sites with the highest posterior probability in codeml, it appears that the low number of sequences in the alignment resulted in a lower power for detecting positive selection in HyPhy. Both FEL and REL algorithms appear to have been optimized using much larger and divergent datasets from viral effector proteins and might not perform effectively with smaller and more conserved sets as those used in this study. However, as more plant genomes get sequenced, all of these methods will become very useful to develop databases of results of positive selection tests that can be actualized on the fly, and that can allow the rapid study of the cereal "selectome", just as has been done with vertebrate genome sequence data (Nickel, Tefft, and Adams 2007; Proux et al. 2008), and to identify interesting candidates for genetic, biochemical and physiological studies.

The permutation of paralogs allows the analysis of genes belonging to gene families, i.e. most of the genes in plants

The permutation analysis we performed for peroxidases was useful to identify amino acid residues that are consistently identified as having evolved under positive selection, even though the analyses were done with different sets of sequences and

with or without removing the gaps in the sequences. These permutations represent multiple independent biological replicates of the evolutionary experiment of positive selection and give confidence that the identified residues are functionally important. The position of the positively selected residues in critical sites of the crystal structure of the proteins further reinforces that possibility. The use of paralogs led to the identification of a few of the same sites, even in sets including very divergent paralogs (P40 in PS2), suggesting that those sites may have also evolved adaptively during the process of sub-functionalization. Importantly, the inclusion of paralogs does not lead to an increased number of false positives, as all the sites with posterior probabilities >0.9 in PS1 and PS2 were also found in the original set NG, and when we used more divergent paralogs in the analysis, the number of sites identified as positively selected decreased to one site with posterior probability >0.9. This may be due to the effect of having substitutions distributed in different positions in the paralog sequences as a consequence of different selective pressures and evolutionary histories. In multiple alignments of paralogs there are many substitutions located in different positions of the alignment for each sequence, instead of several substitutions in the same few positions repeatedly, as is the case in the orthologs (Fig. 2.7), and the maximum likelihood analysis doesn't identify those sites as positively selected. Therefore not only the number of positively selected sites predicted decreased when using very divergent paralogs, but the strength of the prediction was lower as well. Thus, using multiple highly divergent paralogs leads to low power to detect positively selected sites because such set of sequences would more closely resemble neutrally evolving sequences. Moreover, while several other sites had lower posterior probabilities, the fact that 5 sites were reliably identified in the three sets of sequences makes it unlikely that these sites are false positives. It is important to note, however, that poor alignments with the same

sequences can lead to the identification of many false positives, as has been reported in the literature.

Evidence of the importance in disease resistance of some of the positively selected BSU genes

Thaumatins can apparently increase the permeability of hyphal membranes but can also hydrolyse polymeric B-1,3-glucans (Grenier et al. 1999). Additionally, in potato, the overexpression of tobacco PR-5, a thaumatin-like protein, delayed the onset of symptoms after inoculation (Strange 2003) and increased the resistance to *Phytophthora infestans* (Sticher, Mauch-Mani, and Metraux 1997). Some thaumatins can be induced by osmotic stress or developmentally during cherry ripening (Fils-Lycaon et al.) and are abundant in the seed aril of *Thaumatococus. danielii*, where the thaumatin is a sweet protein (De Vos et al. 1985), which suggests a double functionality in seed dispersal and defense from fungi, or at least an ancestral role in defense and a neo-functionalization for use in seed dispersal. In groundnut, thaumatin-like proteins were induced by foliar application of *Pseudomonas fluorescens* and appear to have had a role in the control of leaf spot and rust diseases (Meena et al. 2000). Transgenic wheat plants expressing either a rice-derived thaumatin-like, as well as other line co-expressing chitinase and glucanase transgenes, both categories previously shown to evolve adaptively (Bishop, Dean, and Mitchell-Olds 2000; Bishop et al. 2005), showed increased resistance under greenhouse conditions, in the form of delayed onset of disease symptoms. However, these lines did not show any increased resistance to wheat scab under strong inoculation pressure in the field with *Fusarium graminarum*, suggesting an inefficient initial response to pathogen attack (Anand et al. 2003). Indeed, a large number of pathogenesis-related proteins form part of a quantitative

resistance mechanism, and act synergistically to deter pathogens (van Loon and van Strien 1999).

Peroxidases are enzymes that generate highly toxic molecules collectively referred to as reactive oxygen intermediates (ROIs), such as superoxide anion and hydrogen peroxide, essential in the initial part of the general defense mechanism known as the reactive oxygen burst. This mechanism, common to all eukaryotes, is the first line of defense since it does not require transcription or translation to occur since peroxidases can be stored in the peroxisome, and leads rapidly to programmed cell death, which isolates and kills the pathogen. ROIs may also be used to kill an unrecognized pathogen when other signal transduction pathways reveal that the cell is under attack (Johrde and Schweizer 2008). The *S. bicolor* 2_7192 peroxidase is expressed mainly as a result of pathogen attack, i.e. ESTs were observed only in the PI1 and PIC1 libraries but not in SA1 (Figure 2.6). In agreement with this result, in barley, fungal pathogens elicit the increased expression of several PRPs including peroxidase, while the level of salicylic acid (SA) remains low (Vallélian-Bindschedler, Métraux, and Schweizer 1998). Additionally, generation of ROIs occurs both in compatible and incompatible reactions, initially as a weak response, but a second more intense oxidative burst occurs exclusively in incompatible interactions 3-6 h after inoculation (Strange 2003). This is consistent with the observed upregulation of the 2_7192 peroxidase in *Sorghum bicolor,* where the gene is found mainly in the incompatible reaction (Figure 2.6).

In barley, the *mlo* locus confers broad spectrum resistance to several isolates of the fungal pathogen *Blumeria (=Erysiphe) graminis* f. sp. *hordei,* which causes barley mildew. Interestingly, this form of resistance is recessive in inheritance and has proven to be durable (Strange 2003). In tomato, a naturally occurring loss-of-function mutant of the barley *mlo* homolog also causes broad spectrum resistance to

powdery mildew (Bai et al. 2008). Büschges et al. have suggested that the Mlo wild type allele, which encodes a membrane anchored 60KD protein, acts as a negative regulator of defense and that a complete inactivation of this gene leads to constitutively active mechanisms of disease including the expression of multiple PRPs. Additionally, the recessive phenotype shows lesions in axenically grown plants, which indicates an extreme sensitivity to offensive environmental factors, including UV light. Multiple alleles for this gene have been generated by mutagenesis conferring different degrees of lesion mimic phenotype and disease protection. Furthermore, some recombinants of these alleles restore the wildtype ORF and the susceptible phenotype (Büschges et al. 1997). Although the sites identified by mutagenesis as being important for the function of this protein in barley do not correspond to the sites identified by us as being under positive selection, this was expected since the former sites have a large deleterious effect in the function of the protein. These mutations usually give a fitness penalty in the absence of the pathogen making it unlikely that they could become fixed in natural populations (Büschges et al. 1997).

The *S. bicolor* uniscript 2_3804 codes for a Eukaryotic translation initiation factor 5 (eIF5), a protein that makes critical connections with the 40S ribosome *in vivo*, forms a multifactor complex (MFC) with eIF1, eIF2, and eIF3 that stimulates Met-$tRNA_i^{Met}$ binding to 40S ribosomes and promotes scanning or AUG recognition (Valasek et al. 2003). eIF5-mediated GTP hydrolysis of the ternary complex (ribosome, eIF2, Met–tRNA and GTP) occurs during the binding of the 60S ribosomal subunit to form the functional 80S ribosome (Robaglia and Caranta 2006). No disease phenotypes have been associated with mutations in eIF5 and because of its essential function in translation it is possible that most induced mutations would be lethal. Accordingly, it is reasonable to expect that such a

protein could be an important target for pathogen molecules intended to subvert the host plant. In fact, translation initiation factors are specifically used by potyviruses (Robaglia and Caranta 2006) and naturally occurring variation at eIF4 results in recessive resistance to potyviruses in *Capsicum* (Yeam et al. 2007) and in *Hordeum* (Stein et al. 2005). There are examples of recessive inheritance of resistance to anthracnose in *Sorghum bicolor* (Mehta et al. 2005) and it would be interesting to determine if mutations in translation factors including eIF5 are the underlying cause.

Finally, the ortholog of the single copy SESPY gene (*S. bicolor* 2_8705) in rice (the full length cDNA AK121844.1, also described as the unigene Os11961) is under a Sheath Blight QTL, but there are also several NBS-LRR genes close to this sheath blight QTL (Wisser et al. 2005). An ortholog of this gene, identified through reciprocal BLAST comparisons, is also found in *A. thaliana*, AT5G59080 and is annotated as an unknown protein. Luhua et al. (2008) found that constitutive expression of this gene in *A. thaliana* conferred significantly enhanced tolerance to oxidative stress, but not to osmotic or salinity stress. Moreover, the maize ortholog, determined by a match to this gene in NCBI using 2_8705 as a query, has been implicated in stress resistance (Jia et al. 2006). Therefore the statistical support for adaptive evolution in this gene is strong and there is evidence originating from several species suggesting that this gene is involved in disease/stress resistance and may warrant further analysis.

Conclusions

Although purifying selection rules the evolution of the vast majority of amino acid sites of the studied genes, positive selection at a few critical residues seems to be an important component of the evolution of many genes involved in disease resistance, including those essential for pathogen or disease recognition, signal transduction

pathways, and direct attack towards the pathogen. Products of at least three out of six positively selected genes presented here appear to play some role in plant defense, and all are interesting candidates for functional validation through silencing (Kessler, Halitschke, and Baldwin 2004), biochemical tests of molecular interaction with pathogen effector molecules (Huitema et al. 2004; Tian et al. 2004; Tian, Benedetti, and Kamoun 2005; Tian et al. 2007), as well as directed mutagenesis of target sites identified through evolutionary and structural analyses. Such studies may reveal the specific molecular interactions that are involved, providing hypotheses for the nature of selective forces driving evolution at these loci. For instance, Bishop et al. (2005) showed that the positively selected sites found in plant polygalacturonase inhibitors correspond to sites proven to be functionally important in biochemical studies and that other sites predicted to be under positive selection may also be essential due to their position in the active site of the enzyme.

The determination that some of these genes have evolved under positive selection suggests that they may be targets for pathogen's virulence genes, *i.e.* pathogen derived molecules designed to avoid recognition or undermine host defenses. Alternatively, they can be parts of direct defense mechanisms that change rapidly to adapt to an ever changing pathogen. Thus, the identification of the conserved regions and particularly the derived mutations that made these genes more advantageous to the different grass lineages reveal the critical regions of the protein, as well as pathways that pathogens tend to attack. This information may help identify, conserve and use individuals from natural populations that possess functionally different alleles and to quantify the extant genetic diversity at functional sites. This information can also lead to the rational design and engineering of proteins involved in plant disease resistance.

The number of candidates identified in this study is rather small for a genome-wide scale analysis and this was due to the limitations of the data available for several of the large and repetitive genome species, particularly maize, sugarcane and wheat, at the time of analysis. Although there are numerous ESTs from these species, these data produce many incomplete transcripts for an unknown fraction of the total genes, making it difficult to determine orthology through reciprocal best hits in BLAST comparisons. Additionally, sequencing errors are also common in the kind of data used and it is not a trivial matter to differentiate them from very similar paralogs. Therefore in this study we preferred to err in the side of caution and be very careful to avoid including false positives caused by the mentioned limitations. Nevertheless, the fact that six of these ten genes are under positive selection is quite striking and provides new strong candidates for future functional analyses.

Another factor that could affect the identification of disease response genes using the strategy presented here is the degree of conservation. A few canonical Resistance genes (NBS-LRRs) appear to evolve more slowly and to maintain old alleles, likely due to the effects of balancing selection (McDowell and Simon 2006). Also, Bakker et al. (2008), conducted a molecular diversity survey of 27 defense response genes involved in the salicylic acid, jasmonic acid and ethylene pathways and found that most of these genes show purifying selection. Thus, by focusing on moderately to rapidly divergent genes, it is likely that we missed several legitimate and important disease response genes. In fact, there are several highly upregulated but conserved genes in sorghum (Figure 2.2) that could be important in disease response and warrant further investigation. It is possible that a molecular population genetic analysis of such conserved genes will show a similar scenario as

that described by Bakker et al. (2008), since low divergence should be correlated to low non-synonymous polymorphism, as a result of strict purifying selection. One of the genes described here as only having signals of purifying selection but with prior information of a role in defense, glycosil transferase (2_9185), was determined to be more conserved than the cutoff value we used in this study (BLASTp, $E=10^{-50}$) once we reconstructed the complete open reading frame. This gene was a false positive for divergence but may not be a false positive for disease resistance, even though its evolution is ruled by negative selection. Highly conserved genes, in which non-synonymous substitutions would lead to reduced fitness most of the time, such as those essential for defense signal transduction pathways (Bakker et al. 2008), are probably guarded by R genes in order to protect the individual plant at the expense of individual cells (Dangl and Jones 2001; Van der Hoorn, De Wit, and Joosten 2002).

In our study, however, we decided to identify candidate disease resistance genes, not only by the information on the literature, but based on high up-regulation under pathogen attack and particularly based on high divergence. Therefore this fundamental difference in our approach makes it more likely to identify genes under positive selection. Indeed, a few of the genes identified using this strategy and that showed high values of dN/dS ratio are known pathogenesis related proteins, namely, thaumatins and peroxidases. Other genes found here, may also have an important role in quantitative disease resistance and can now be functionally tested.

Additionally, we feel that this study demonstrates the feasibility of identifying genes that may be important in other kinds of adaptive process, e.g. drought tolerance in sorghum, and that it should be simple and useful to implement this kind of analyses in automated pipelines, for instance in Gramene, that will allow end users to identify functional polymorphisms and take full advantage of the rapid

advances in genome sequencing and the exponential increase in high quality sequence data, full length cDNAs, proteomic analyses, expression profiles and several other kinds of functional genomic studies currently being conducted in the cereals. As the number of full length coding sequences increases for the different cereal genomes, this type of analysis will be easier to do for thousands of genes and it is likely that signals of adaptive evolution will be found in many more DRGs involved in several aspects of disease resistance.

Funding

This work was supported by the National Science Foundation [0115903 to SK].

Aknowledgements

We thank Sharon Mitchell, Alex Casa and Peter Moffett, and three anonymous reviewers for valuable comments on the manuscript. We also want to give special thanks to Lee Pratt for the use of MAGIC and valuable comments during the development of this project. Many thanks to Jarek Pillardy, from CBSU, who provided suggestions and guidance with the bioinformatics analysis; Chenwei Lin helped with Perl scripts to parse the InterProScan output and Immanuel Yap for help with the use of GRAMENE. AZM is supported by the Costa Rica-USA Foundation for Cooperation (CRUSA) through a special Fulbright scholarship in biotechnology and by the Institute of Genomic Diversity (Cornell University, Ithaca, NY).

REFERENCES

Anand, A., T. Zhou, H. N. Trick, B. S. Gill, W. W. Bockus, and S. Muthukrishnan. 2003. Greenhouse and field testing of transgenic wheat plants stably expressing genes for thaumatin-like protein, chitinase and glucanase against *Fusarium graminearum*. Journal of Experimental Botany **54**:1101-1111.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2002. Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection. Mol Biol Evol **19**:950-958.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. Mol Biol Evol **18**:1585-1592.

Arora, R., P. Agarwal, S. Ray, A. K. Singh, V. P. Singh, A. K. Tyagi, and S. Kapoor. 2007. MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. BMC Genomics **8**:242.

Bai, Y., S. Pavan, Z. Zheng et al. 2008. Naturally occurring broad-spectrum powdery mildew resistance in a Central American tomato accession is caused by loss of *mlo* function. Molecular Plant-Microbe Interaction **21**:30-39.

Bakker, E. G., C. Toomajian, M. Kreitman, and J. Bergelson. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. Plant Cell **18**:1803-1818.

Bakker, E. G., M. B. Traw, C. Toomajian, M. Kreitman, and J. Bergelson. 2008. Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana*. Genetics **178**:2031-2043.

Barker, N. P., L. G. Clark, J. I. Davis et al. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Annals of the Missouri Botanical Garden **88**:373-457.

Bishop, J. G., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A **97**:5322-5327.

Bishop, J. G., D. R. Ripoll, S. Bashir, C. M. B. Damasceno, J. D. Seeds, and J. K. C. Rose. 2005. Selection on glycine beta-1,3-endoglucanase genes differentially inhibited by a *Phytophthora* glucanase inhibitor protein. Genetics **169**:1009-1019.

Bowles, D. J. 1990. Defense-Related Proteins in Higher Plants. Annual Review of Biochemistry **59**:873-907.

Büschges, R., K. Hollricher, R. Panstruga et al. 1997. The barley *mlo* gene: a novel control element of plant pathogen resistance. Cell **88**: 695-705.

Crouch, J. A., B. B. Clarke, and B. I. Hillman. 2006. Unraveling evolutionary relationships among the divergent lineages of *Colletotrichum* causing anthracnose disease in turfgrass and corn. Phytopathology **96**:46-60.

Dangl, J. L., and J. D. G. Jones. 2001. Plant pathogens and integrated defence responses to infection. Nature **411**:826-833.

De Vos, A. M., M. Hatada, H. Van Der Wel, H. Krabbendam, A. F. Peerdeman, and S. H. Kim. 1985. 3-Dimensional structure of thaumatin I: an intensely sweet protein. Proceedings of the National Academy of Sciences of the United States of America **82**:1406-1409.

Erpelding, J. E., and L. K. Prom. 2006. Variation for anthracnose resistance within the sorghum germplasm collection from Mozambique, Africa. Plant Pathology Journal **5**:28-34.

Fils-Lycaon, B. R., P. A. Wiersma, K. C. Eastwell, and P. Sautiere. 1996. A cherry protein and its gene, abundantly expressed in ripening fruit, have been identified as thaumatin-like. Plant Physiol. **111**:269-273.

Frye, C. A., D. Tang, and R. W. Innes. 2000. Negative regulation of defense responses in plants by a conserved MAPKK kinase. Proceedings of the National Academy of Sciences **98**:373–378.

Gajhede, M., D. J. Schuller, A. Henriksen, A. T. Smith, and T. L. Poulos. 1997. Crystal structure of horseradish peroxidase C at 2.15 Å resolution. Nat Struct Mol Biol **4**:1032-1038.

Goff, S. A., D. Ricke, T. H. Lan et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. *ssp japonica*). Science **296**:92-100.

Grenier, J., C. Potvin, J. Trudel, and A. Asselin. 1999. Some thaumatin-like proteins hydrolyse polymeric B-1,3-glucans. The Plant Journal **19**:473-480.

Hammond-Kosack, K. E., and J. D. G. Jones. 1997. Plant disease resistance genes. Jones, R. L. [Editor] **Annual Review of Plant Physiology and Plant Molecular Biology**:575-607.

Heo, W. D., S. H. Lee, M. C. Kim et al. 1999. Involvement of specific calmodulin isoforms in salicylic acid-independent activation of plant disease resistance responses. Proceedings of the National Academy of Sciences of the United States of America **96**:766-771.

Herring, C. D., A. Raghunathan, C. Honisch et al. 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat Genet **38**:1406-1412.

Huitema, E., J. I. B. Bos, M. Y. Tian, J. Win, M. E. Waugh, and S. Kamoun. 2004. Linking sequence to phenotype in *Phytophthora*-plant interactions. Trends in Microbiology **12**:193-200.

Jaiswal, P., D. Ware, J. J. Ni et al. 2002. Gramene: development and integration of trait and gene ontologies for rice. Comparative and Functional Genomics **3**:132-136.

Jia, J., J. Fu, J. Zheng et al. 2006. Annotation and expression profile analysis of 2073 full-length cDNAs from stress-induced maize (*Zea mays* L.) seedlings. The Plant Journal **48**:710-727.

Johrde, A., and P. Schweizer. 2008. A class III peroxidase specifically expressed in pathogen-attacked barley epidermis contributes to basal resistance. Molecular Plant Pathology **9**:687-696.

Kessler, A., R. Halitschke, and I. T. Baldwin. 2004. Silencing the jasmonate cascade: Induced plant defenses and insect populations. Science **305**:665-668.

Kimchi-Sarfaty, C., J. M. Oh, I.-W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman. 2007. A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science **315**:525-528.

Komar, A. A., T. Lesnik, and C. Reiss. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Letters **462**:387-391.

Kosakovsky Pond, S. L., and S. D. W. Frost. 2005. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Mol Biol Evol **22**:1208-1222.

Langlois-Meurinne, M., C. M. M. Gachon, and P. Saindrenan. 2005. Pathogen-responsive expression of glycosyltransferase genes UGT73B3 and UGT73B5 is necessary for resistance to *Pseudomonas syringae pv tomato* in *Arabidopsis*. Plant Physiol. **139**:1890-1901.

Luhua, S., S. Ciftci-Yilmaz, J. Harper, J. Cushman, and R. Mittler. 2008. Enhanced Tolerance to Oxidative Stress in Transgenic *Arabidopsis* Plants Expressing Proteins of Unknown Function. Plant Physiol. **148**:280-292.

Marchler-Bauer, A., J. B. Anderson, F. Chitsaz et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucl. Acids Res. **37**:D205-210.

Maynard Smith, J. 1998. Evolutionary genetics. Oxford:Oxford University Press.

McDowell, J. M., and J. L. Dangl. 2000. Signal transduction in the plant immune response. Trends in Biochemical Sciences **25**:79-82.

McDowell, J. M., and S. A. Simon. 2006. Recent insights into R gene evolution. Molecular Plant Pathology **7**:437-448.

Meena, B., R. Radhajeyalakshmi, T. Marimuthu, P. Vidhyasekaran, S. Doraiswamy, and R. Velazhahan. 2000. [Induction of pathogenesis-related proteins, phenolics and phenylalanine ammonia-lyase in groundnut by *Pseudomonas fluorescens*] Induktion von Pathogenese-assoziierten Proteinen, Phenolen und Phenylalaninammonium-Lyase in Erdnussen durch *Pseudomonas fluorescens*. Zeitschrift fuer Pflanzenkrankheiten und Pflanzenschutz **107**:514-527.

Mehta, P. J., C. C. Wiltse, W. L. Rooney, S. D. Collins, R. A. Frederiksen, D. E. Hess, M. Chisi, and D. O. TeBeest. 2005. Classification and inheritance of genetic resistance to anthracnose in sorghum. Field Crops Research **93**:1-9.

Michelmore, R. W., and B. C. Meyers. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research **8**:1113-1130.

Mondragón-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Research **12**:1305-1315.

Nickel, G. C., D. Tefft, and M. D. Adams. 2007. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. Nucl. Acids Res.:1-8.

Nielsen, R. 2005. Molecular signatures of natural selection. Annual Review of Genetics **39**:197-218.

Nielsen, R., and M. J. Hubisz. 2005. Evolutionary genomics: detecting selection needs comparative data. Nature **433**.

Nielsen, R., and Z. H. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.

Paterson, A. H., J. E. Bowers, and B. A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America **101**:9903-9908.

Popescu, S. C., G. V. Popescu, S. Bachan, Z. Zhang, M. Gerstein, M. Snyder, and S. P. Dinesh-Kumar. 2009. MAPK target networks in *Arabidopsis thaliana* revealed using functional protein microarrays. Genes & Development **23**:80-92.

Prasad, V., C. A. E. Stromberg, H. Alimohammadian, and A. Sahni. 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science **310**:1177-1180.

Pratt, L. H., C. Liang, M. Shah et al. 2005. Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and skotomorphogenesis

from a milestone set of 16,801 unique transcripts. Plant Physiol. **139**:869-884.

Proux, E., R. A. Studer, S. Moretti, and M. Robinson-Rechavi. 2008. Selectome: a database of positive selection. Nucl. Acids Res.:gkn768.

Ramakrishna, W., J. Dubcovsky, Y. J. Park, C. Busso, J. Emberton, P. SanMiguel, and J. L. Bennetzen. 2002. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics **162**:1389-1400.

Rice Chromosome 3 Sequencing, C. 2005. Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. Genome Research **15**:1284-1291.

Robaglia, C., and C. Caranta. 2006. Translation initiation factors: a weak link in plant RNA virus infection. Trends in Plant Science **11**:40-45.

Rosewich, U. L., R. E. Pettway, B. A. McDonald, R. R. Duncan, and R. A. Frederiksen. 1998. Genetic structure and temporal dynamics of a *Colletotrichum graminicola* population in a sorghum disease nursery. Phytopathology **88**:1087-1093.

Salamov, A. A., and V. V. Solovyev. 2000. *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Research **10**:516-522.

Salzman, R. A., J. A. Brady, S. A. Finlayson et al. 2005. Transcriptional Profiling of Sorghum Induced by Methyl Jasmonate, Salicylic Acid, and Aminocyclopropane Carboxylic Acid Reveals Cooperative Regulation and Novel Gene Responses. Plant Physiol. **138**:352-368.

Schlenke, T. A., and D. J. Begun. 2005. Linkage disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. Genetics **169**:2013-2022.

Shatters, R. G., L. M. Boykin, S. L. Lapointe, W. B. Hunter, and A. A. Weathersbee. 2006. Phylogenetic and structural relationships of the PR5 gene family reveal an ancient multigene family conserved in plants and select animal taxa. Journal of Molecular Evolution **63**:12-29.

Souza-Paccola, E. A., L. C. L. Favaro, C. R. Casela, and L. D. Paccola-Meirelles. 2003. Genetic recombination in *Colletotrichum sublineolum*. Journal of Phytopathology **151**:329-334.

Stein, N., D. Perovic, J. Kumlehn, B. Pellio, S. Stracke, S. Streng, F. Ordon, and A. Graner. 2005. The eukaryotic translation initiation factor 4E confers multiallelic recessive Bymovirus resistance in *Hordeum vulgare* (L.). Plant Journal **42**:912-922.

Stekel, D. J., Y. Git, and F. Falciani. 2000. The comparison of gene expression from multiple cDNA libraries. Genome Research **10**:2055-2061.

Sticher, L., B. Mauch-Mani, and J. P. Metraux. 1997. Systemic acquired resistance. Annual Review of Phytopathology **35**:235-270.

Stotz, H. U., J. G. Bishop, C. W. Bergmann, M. Koch, P. Albersheim, A. G. Darvill, and J. M. Labavitch. 2000. Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors. Physiological & Molecular Plant Pathology **56**:117-130.

Strange, R. N. 2003. Introduction to plant pathology. West Sussex, England:John Wiley & Sons Ltd.

Suman, A., S. Lal, A. K. Shasany, A. Gaur, and P. Singh. 2005. Molecular assessment of diversity among pathotypes of *Colletotrichum falcatum* prevalent in sub-tropical indian sugarcane. World Journal of Microbiology & Biotechnology **21**:1135-1140.

Taubes, G. 2008. The Bacteria Fight Back. Science **321**:356-361.

Tian, M. Y., B. Benedetti, and S. Kamoun. 2005. A second kazal-like protease inhibitor from *Phytophthora infestans* inhibits and interacts with the apoplastic pathogenesis-related protease P69B of tomato. Plant Physiology **138**:1785-1793.

Tian, M. Y., E. Huitema, L. da Cunha, T. Torto-Alalibo, and S. Kamoun. 2004. A Kazal-like extracellular serine protease inhibitor from *Phytophthora infestans* targets the tomato pathogenesis-related protease P69B. Journal of Biological Chemistry **279**:26370-26377.

Tian, M. Y., J. Win, J. Song, R. van der Hoorn, E. van der Knaap, and S. Kamoun. 2007. A *Phytophthora infestans* cystatin-like protein targets a novel tomato papain-like apoplastic protease. Plant Physiology **143**:364-377.

Tiffin, P. 2004. Comparative evolutionary histories of chitinase genes in the genus *Zea* and family Poaceae. Genetics **167**:1331-1340.

Valasek, L., A. A. Mathew, B.-S. Shin, K. H. Nielsen, B. Szamecz, and A. G. Hinnebusch. 2003. The yeast eIF3 subunits TIF32/a, NIP1/c, and eIF5 make critical connections with the 40S ribosome in vivo. Genes Dev. **17**:786-799.

Vallélian-Bindschedler, L., J.-P. Métraux, and P. Schweizer. 1998. Salicylic acid accumulation in barley is pathogen specific but not required for defense-gene activation. Molecular Plant-Microbe Interactions **11**:702-705.

Van der Hoorn, R. A. L., P. De Wit, and M. Joosten. 2002. Balancing selection favors guarding resistance proteins. Trends in Plant Science **7**:67-71.

van Loon, L. C., M. Rep, and C. M. J. Pieterse. 2006. Significance of inducible defense-related proteins in infected plants. Annual Review of Phytopathology **44**:135-162.

van Loon, L. C., and E. A. van Strien. 1999. The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. Physiological & Molecular Plant Pathology **55**:85-97.

Vettore, A. L., F. R. da Silva, E. L. Kemper et al. 2003. Analysis and Functional Annotation of an Expressed Sequence Tag Collection for Tropical Crop Sugarcane. Genome Res. **13**:2725-2735.

Wisser, R. J., Q. Sun, S. H. Hulbert, S. Kresovich, and R. J. Nelson. 2005. Identification and Characterization of Regions of the Rice Genome Associated with Broad-Spectrum, Quantitative Disease Resistance. Genetics **169**:2277-2293.

Wong, W. S. W., Z. H. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168**:1041-1051.

Yang, Z. H., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yeam, I., J. R. Cavatorta, D. R. Ripoll, B.-C. Kang, and M. M. Jahn. 2007. Functional Dissection of Naturally Occurring Amino Acid Substitutions in eIF4E That Confers Recessive Potyvirus Resistance in Plants. Plant Cell **19**:2913-2928.

Yu, J., S. N. Hu, J. Wang et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. *ssp indica*). Science **296**:79-92.

CHAPTER THREE

BALANCING SELECTION AND SELECTIVE SWEEPS IN *SORGHUM BICOLOR* UP-REGULATED AND DIVERGENT DISEASE RESPONSE GENES[3]

Abstract

Disease response genes (DRGs) are an essential component of a long term molecular arms race between hosts and pathogens. As expected, DRGs sometimes show high inter-specific divergence, and positive selection has been found at several of these genes. At the population level, DRGs can show either significantly lower or higher diversity levels, in relation to average genomic regions, corresponding respectively to selective sweeps or to the effect of balancing selection. Rapid evolution can occur through recurrent events of positive selection and therefore significantly reduced intra-specific polymorphism should be common in divergent DRGs. However, the most prevalent signal of selection found in disease response genes studied so far is that of balancing selection, and few examples of selective sweeps have been found. Additionally, few direct comparisons have been made between codon-based maximum likelihood analyses of adaptive evolution and intra-specific polymorphism of DRGs. Here we present a direct comparison of the macro-evolutionary history of *S. bicolor* disease response genes and their intra-specific polymorphism, and show that moderately rapidly evolving, positively selected genes have patterns of polymorphism consistent both with selective sweeps and balancing selection. The *Sorghum bicolor mlo* homolog and *SESPY*, a gene of unknown function, have divergent alleles consistent with balancing selection. On the contrary, a Class III peroxidase and a gene with an RNA Binding Domain

[3] Zamora-Meléndez, A., Aquadro, C.F. and Kresovich, S. Balancing selection and selective sweeps in *Sorghum bicolor* upregulated and divergent disease response genes. MBE, to be submitted.

(*RBD*), have significantly reduced diversity, consistent with the effect of recent cross-species selective sweeps. This evidence indicates that these disease response genes have evolved under positive selection for millions of years, along different cereal lineages, and continue to do so in current populations possibly due to the strong and constant selective pressure imposed by pathogens.

Introduction

Theoretical analyses of plant-pathogen interactions have resulted in several models describing their antagonistic co-evolution. Two main models have been proposed to explain the evolution of disease response genes (DRGs), namely, the co-evolutionaly arms-race model and the overdominance model. The first model theorizes that DRGs should be constantly changing through recurrent events of directional selection, without ever reaching an optimum state. This dynamic reflects the evolution of the pathogen's molecules, which are also always changing to avoid recognition and to find new ways to defeat the host. This model predicts that DRGs should show highly reduced levels of polymorphism as well as significantly increased inter-specific divergence, when compared with neutrally evolving genes or genomic regions. However, contrary to expectations, very few of the DRGs studied until now show this pattern of genetic variation, for instance, Rps4 in *A. thaliana* (Bergelson et al. 2001) and the rice blast R gene Pi-ta, in the wild rice *Oryza rufipogon* (Huang et al. 2008).

The second model is the overdominance/diversifying selection model, in which selective pressure would favor new alleles that confer higher fitness on heterozygous individuals. It was proposed long ago due to the predominance of balancing selection in the human Major Histocompatibility Complex, MHC, (Hughes, Ota, and Nei 1990) as well as in plant R genes, those having Nucleotide

Binding Sites (NBS) and Leucine Rich Repeat (LRR) domains. There are, however, few confirmed cases where a heterozygous individual has a selective advantage, such as β-globin in humans (Allison 1954) and MHC (Hughes, Ota, and Nei 1990). Moreover, Borghans, Beltman, and De Boer (2004) used computer simulations to show that overdominance does not explain the extremely high polymorphism observed at MHC loci, while negative frequency dependent selection does. Additionally, the presence of hemizygous loci and of loss-of-function alleles suggests that, for a few loci, overdominance is not correct. For these last two cases, the "trench warfare" model of diversifying selection seems more appropriate (Stahl et al. 1999; Ingvarsson 2005). In this model alleles at a DR locus are thought to have co-existed for millions of years due to advances and retreats in the frequency of resistance and susceptibility alleles. A special case of this model was previously described as the pleiotropy theory (Gillespie 1975; May and Anderson 1990; Parker 1990), which specifically states that "if genes conferring resistance also cause inferior fitness under disease-free conditions, the polymorphism can be maintained by a reversal in the relative fitness of different phenotypes in the presence and absence of pathogen attack" (Parker 1990). To date most plant disease response genes analyzed have shown the signals of balancing selection and not of directional selection (Stahl et al. 1999; Bergelson et al. 2001; Mauricio et al. 2003; Ingvarsson 2005; Bakker et al. 2006) or that of a more complex configuration referred to as allele cycling (Ingvarsson 2005). That the same general pattern has been found in humans (Ferrer-Admetlla et al. 2008) seems to indicate that balancing selection is the norm for DRGs and that the existing dynamic polymorphism seen at DRGs is maintained by selection at different functionally important alleles at the same locus (Bergelson et al. 2001; Tian et al. 2002; Mauricio et al. 2003). However, Bergelson et al. (2001) suggested that there is a bias in the sampled genes, since most of the

loci, R-genes in particular, had been commonly mapped by virtue of their polymorphism and that some DR loci should show the signals of selective sweeps. Unexpectedly, Bakker et al. (2008) found that most *Arabidopsis* genes involved in signal transduction pathways are conserved, with only a few showing minimal evidence of directional selection. This pattern contrasted strongly with that found for R genes using the same panel (Bakker et al. 2006) and suggests that the genes' function and position in gene networks may have an effect on the kind of diversity patterns they display.

Zamora et al. (2009) found that several highly (conditionally) expressed *Sorghum bicolor* Biotic Stress Upregulated (BSU) genes had above-average divergence with respect to rice, contrasting with the observation in some species that gene expression level is positively correlated with evolutionary conservation (Pal, Papp, and Hurst 2001; Krylov et al. 2003). Several of those divergent BSU genes showed signals of historical adaptive evolution in the form of a high dN/dS ratio, and some positively selected sites mapped to the active sites of enzymes, apparently due to these enzymes' interaction with pathogen-derived molecules (Zamora et al. 2009). These findings are in agreement with the arms-race hypothesis: genes important in disease resistance should be divergent because of the selective pressure to change and to adapt to a constantly evolving pathogen biota, leading to the observed divergence through a preponderance of non-synonymous substitutions. If the arms-race hypothesis is correct for these genes, we should be able to identify signals of selection in current populations of *Sorghum bicolor*, e.g. a significant excess of rare or high frequency derived alleles. It is possible, however, that the selective process took place long ago, and that the pathogens that caused it are no longer virulent or even present. Hence, it is important to determine whether there is evidence of recent selection in these DRGs, as this would indicate an on-going host-pathogen

interaction that could be biologically interesting and economically important to address.

Wild accessions of *Sorghum bicolor* from throughout its natural range in Africa show higher genetic diversity at neutral markers (Casa et al. 2005) and coding regions (Hamblin et al. 2005) than even the most divergent cultivated accessions. Wild sorghum accessions are likely to have lower genome wide linkage disequilibrium and therefore can be useful to increase the statistical power of the analyses used to identify signals of selection in the candidate genes, across the whole *S. bicolor* population or in subpopulations. Here we describe the polymorphism and divergence of several known and putative DRGs in the wild population of *Sorghum bicolor*, identified based on their expression profile and divergence to rice, including some of the rapidly evolving genes studied by Zamora et al. (2009), which include known and putative DRGs. We show that the polymorphism estimates at these loci map to the extremes of the distribution for genetic diversity for the species, having either very low or very high polymorphism, and that all show higher non-synonymous fixed differences, with respect to other grass species, than control genes. Two of the genes studied show a pattern of polymorphisms consistent with balancing selection, while two loci, a peroxidase and a gene containing an RNA Binding Domain (*RBD*), have evidence for a recent or strong, cross-species selective sweep.

Material and Methods

Plant material

We amplified and sequenced the candidate genes in a diverse panel of *Sorghum bicolor* accessions (Table 3.1), consisting of mostly wild plants from various geographic regions in Africa. We selected the wild species based on a microsatellite analysis (Casa et al. 2005) to include the most informative accessions while eliminating redundancies. We used mostly wild species because, first, diversity within landraces and elite varieties of *S. bicolor* is limited (Hamblin et al. 2004), and second, landraces have limited genetic structure (Casa et al. 2005), and we wanted to maximize the probability of finding evidence of recent selective sweeps at both the species and subpopulation levels. *Sorghum bicolor* accessions used (see Table 3.1) included individuals from natural populations (n=36), exotic cultivated lines (n=3), and inbred lines (n=1). Other *Sorghum* species (Price et al. 2005) were included in the analyses to compare intraspecific polymorphism with interspecific divergence in tests of selective neutrality. These *Sorghum* species were *S. angustum* S.T. Blake, *S. bulbosum* Lazarides, *S. ecarinatum* Lazarides, *S. extans* Lazarides, *S. laxiflorum* Bailey, *S. macrospermum* Garber, *S. matarankense* Garber & Snyder, *S. propinquum* (Kunth) Hitch, *S. purpureosericeum* (A. Rich) Aschers & Schweinf, *S. stipoideum* (Ewart & Jean White) C. Gardner and C. E. Hubb, *S. timorense* (Kunth) Buse and *S. versicolor* Anderss. Additionally, *Saccharum giganteum* (octoploid), a wild American sugarcane-related species, was also used to amplify and sequence the candidate loci. DNA was extracted from several leaves of one individual using the method of Doyle and Doyle (1987). Fresh tissue was ground in liquid nitrogen using mortar and pestle. The leaf powder was incubated in a CTAB buffer at 60°C for one hour and an equal volume of chloroform:isoamyl alcohol (24:1) was added. DNA was precipitated with 7.5M ammonium acetate and 90% ethanol chilled to -20°C.

After centrifugation the pellet was resuspended in water, RNAse was added, and a new extraction with chloroform was performed. The purified DNA was resuspended in TE and stored at -20° until used.

Identification of candidate disease response genes

In *S. bicolor,* the Biotic Stress subgroup of EST libraries is comprised of three EST libraries called the **P**athogen induced **I**ncompatible (PI1, resistant reaction, 9533 ESTs), the **P**athogen **I**nfected **C**ompatible (PIC1, susceptible reaction, 10209 ESTs) and the **S**alicylic **A**cid treated library (SA1, 5801 ESTs, see Pratt et al. (2005) for a detailed description of the 20 *Sorghum bicolor* EST tissue and treatment-specific libraries and analysis software). Salicylic Acid is an important signal molecule in plant disease resistance as it leads to systemic acquired resistance. We used the program MAGIC Gene Discovery (Laboratory for Genomics and Bioinformatics, The University of Georgia, Athens) to identify uniscripts constructed using ESTs that came primarily from these three libraries. A uniscript is defined here as a unique splicing form of a gene model made from ESTs that can come from only one (tissue- or treatment-specific genes) or many EST libraries (housekeeping genes). We selected uniscripts comprising more than 70% ESTs expressed in the biotic stress subgroup of libraries and with at least 7 ESTs composing the uniscript. A believability score reflecting the probability that a gene is upregulated in a particular library was generated by Pratt (2005) for each of the uniscripts, based on the method of Stekel, Git, and Falciani . Finally, we identified uniscripts of interest as those showing a divergence to rice of at least $E = 10^{-50}$ in BLASTP comparisons. In addition to the candidates identified using the described conditions, we included the candidates studied by Zamora et al. (2009).

Table 3.1. Panel of *Sorghum bicolor* accessions used in the sequencing of the moderately rapidly evolving disease response gene candidates to evaluate their polymorphism, inter-specific divergence.

Status	PI	Name	ICRISAT ID	US Source	Origin	Genus	Species	Subspecies	Race
I	BTx623	US Inbred line*				*Sorghum*	*bicolor*	*bicolor*	-
L	PI22913	Chinese Amber	IS 12711	Georgia PGRCU	China	*Sorghum*	*bicolor*	*bicolor*	*bicolor*
L	PI257595	NO. 1 Gambela	IS 12608	Georgia PGRCU	Ethiopia	*Sorghum*	*bicolor*	*bicolor*	*caudatum*
L	PI267380		IS 2694	Georgia PGRCU	Zimbabwe	*Sorghum*	*bicolor*	*bicolor*	*Kafir*
L	PI152705	Kokla	IS 12570	Georgia PGRCU	Sudan	*Sorghum*	*bicolor*	*bicolor*	*Durra*
L	NSL50876	65I 1634	IS 7173	Fort Collins	Tanzania	*Sorghum*	*bicolor*	*bicolor*	*Guinea*

Table 3.1 (Continued)

W	L-WA12	HD-769-1		Kansas	Sudan	*Sorghum*	*bicolor*	*drummondii*	
W	L-WA13	HD-552		Kansas	Sudan	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA15	HD-555		Kansas	Sudan	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA17		IS14215	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA18		IS14219	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA20		IS14233	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA22		IS14235	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA23		IS14237	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA25		IS14249	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA26		IS14250	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA27		IS14251	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA28		IS14252	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA29		IS14254	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA31		IS14259	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*

Table 3.1 (Continued)

W	L-WA38		IS14279	Kansas	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA41		IS14312	Kansas	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*verticilliflorum*
W	L-WA42		IS14313	Kansas	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA43		IS14329	Kansas	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA44		IS14330	Kansas	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*Verticilliflorum*
W	L-WA55	HD-629		Kansas	Benin	*Sorghum*	*bicolor*	*arundinaceum*	*arundinaceum*
W	L-WA58		IS14232	Kansas	Angola	*Sorghum*	*bicolor*	*arundinaceum*	*arundinaceum*
W	L-WA59		IS14300	Kansas	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*arundinaceum*
W	L-WA63		IS14359	Kansas	Malawi	*Sorghum*	*bicolor*	*arundinaceum*	*arundinaceum*

Table 3.1 (Continued)

W	L-WA67		IS14485	Kansas	Sudan	*Sorghum*	*bicolor*	*arundinaceum*	*aethiopicum*
W	L-WA71	27/80K		Kansas	Sudan	*Sorghum*	*bicolor*	*drummondii*	
W	L-WA72	PQ-728/80K		Kansas	Sudan	*Sorghum*	*bicolor*	*drummondii*	
W	L-WA81		IS14473	Kansas	Sudan	*Sorghum*	*bicolor*	*drummondii*	
W	L-WA88		IS18808	Kansas	Egypt	*Sorghum*	*bicolor*	*arundinaceum*	*virgatum*
W	L-WA89		IS18815	Kansas	Egypt	*Sorghum*	*bicolor*	*arundinaceum*	*virgatum*
W	PI199869		-	Georgia PGRCU	South Africa	*Sorghum*	*bicolor*	*drummondii*	
W	PI213900		IS 3106	Georgia PGRCU	Kenya	*Sorghum*	*bicolor*	*arundinaceum*	*verticilliflorum*
W	PI300118	294		Georgia PGRCU	South Africa	*Sorghum*	*bicolor*	*arundinaceum*	*verticilliflorum*
W	PI302105		IS 12687	Georgia PGRCU	Ethiopia	*Sorghum*	*bicolor*	*arundinaceum*	*aethiopicum*

Table 3.1 (Continued)

W	PI302233	A-7171	-	Georgia PGRCU	Egypt	*Sorghum*	*bicolor*	*arundinaceum*	*virgatum*
W	PI154831	66I 4913	IS3636	Georgia	Uganda	*Sorghum*	*bicolor*	*arundinaceum*	*verticilliflorum*
W	PI225905		IS 12693	Georgia PGRCU	Zambia	*Sorghum*	*bicolor*	*arundinaceum*	*arundinaceum*
W	PI247723	21221		Georgia PGRCU	Zaire	*Sorghum*	*bicolor*	*arundinaceum*	*verticilliflorum*

*Sorghum line used for genome sequencing (Paterson et al. 2009)

PCR amplification and sequencing of candidate genes in sorghum

Primer sets were designed using Primer3 (Rozen and Skaletsky 2000) and we tried to cover most of the contiguous region assembled from the uniscripts and other available genomic sequences. To amplify and sequence some of the wild accessions and other species it was necessary to design multiple primers to the predicted coding regions and to reduce the annealing temperature 5–10 degrees C. PCR products were treated with Exonuclease I and Shrimp Alkaline Phosphatase, sequenced using both forward and reverse primers with ABI (Columbia, MD) Big Dye, and analyzed with ABI 3700, at the Life Sciences Core Laboratories Center (Cornell University). All chromatograms were visually inspected in Sequencher (Gene Codes Corp., Ann Arbor, MI). Forward and reverse sequences of each accession were combined and aligned using Sequencher, and polymorphisms were accepted only where both forward and reverse sequences showed the SNP. Often more than one accession showed the SNP at a particular locus, which increased confidence in the call. Absolute singletons at the edges of the high quality sequence were disregarded to be conservative. The DNA sequences were deposited in the National Center for Biotechnology Information (NCBI) nucleotide database (*i.e.*, GenBank).

Molecular population genetics

Polymorphism and divergence parameters were calculated using DnaSP 4.00.4 (Rozas and Rozas 1999). The species-wide silent site nucleotide diversity, π (Nei, 1987) and θ_S (= θ_W (Watterson 1975)), were estimated. To test the strict theory of neutral equilibrium molecular evolution we used Tajima's D (1989) and Fu and Li's (1993) statistics, as implemented in DnaSP. We also used the McDonald-Kreitman test (McDonald and Kreitman 1991) to compare the polymorphism and divergence in the coding and non-coding regions, respectively. The genetic structure for each

locus for the whole African population of *Sorghum bicolor* was assessed using the G_{ST} statistic in DnaSP. To do this we sub-classified the individuals *a priori* into the broad geographic regions of North and South Africa.

Results

Using the combined information from the number of ESTs that make up a uniscript, the ratio of ESTs that are derived from BS libraries, and the BLASTP divergence value to the rice ortholog, we identified a set of 30 *S. bicolor* uniscripts having characteristics consistent with an arms race evolutionary model, namely, those highly divergent genes that are highly expressed in the Biotic Stress subgroup of libraries, from now on referred to as Highly Expressed and Divergent (HED) genes (Table 3.2). All these candidates have at least 7 ESTs, and at least 70% of their ESTs come from the biotic stress subgroup. These candidates also have an E value lower than 10^{-60} in the BLASTP vs. rice, showing moderate to high divergence, despite a relatively long match in amino acid residues (Table 3.2 shows the best 12 uniscripts).

The annotation for some of the HED orthologous rice genes is coherent with genes involved in disease resistance, such as an LRR gene -a gene with similarity to classical R genes and possibly involved in recognition of a pathogen ligand-, and also a thaumatin, a known Pathogenesis Related Protein (PRP-5). Additionally, there are genes that may be involved in signal transduction, such as the two-component response regulator, known to activate MAPKKKs that amplify signals through cascades of phosphorylation (Posas and Saito 1998), and genes coding for transcription factors such as the SIT4 phosphatase

.

Table 3.2 Highly expressed and divergent (HED) genes. These uniscripts have the highest number of ESTs, Biotic Stress ratio and show a moderate divergence between sorghum and rice, suggesting importance in defense response and rapid evolution through recurrent events of positive selection.

Uniscript	ESTs	BSR	Match length	E value	Rice best hit annotation
2_11508	19	0.789	72	2E-37	LOC_Os05g11910 GDSL-like Lipase /Acylhydrolase
2_13461	27	0.815	165	5E-37	LOC_Os11g46050.1 hypothetical protein
2_13771	20	0.9	173	9E-23	LOC_Os11g46000.1 protein expressed
2_8981	17	0.706	157	2E-49	LOC_Os12g43380.1protein Thaumatin
2_8605	15	0.733	135	4E-31	LOC_Os04g52940.1 SIT4 phosphatase
2_8705	14	0.714	110	3E-36	LOC_Os05g38040.3 expressed protein
2_8223	12	0.75	122	2E-46	LOC_Os09g36220.1 Two-component response regulator PRR95
2_6266	12	0.833	100	2E-36	LOC_Os11g03390.1 FHA domain containing protein
2_6582	8	0.75	78	3E-32	LOC_Os09g37540.1protein expressed
2_7407	7	0.714	144	3E-35	LOC_Os04g52960.1 RNA recognition motif family
2_7586	7	1	147	5E-49	LOC_Os02g11760.1 PDR5-like ABC transporter
2_7647	7	0.857	164	8E-48	LOC_Os05g40270.8 Leucine rich repeat (LRR) protein kinase

Table 3.3 Polymorphism survey of moderately divergent, positively selected candidate disease resistance genes.

Gene[a]	bp	GO function	n	s	π	Taj D	Div
2_8981F1	1103	Thaumatin	7	15	9.17	0.967	3. 0
2_11684F1 [a]	403	Mlo	9	5	10.43	2.032*	7.5
2_11684F2 [b]	575	Mlo	8	2	3.77	0.242	0.287
2_7192F2	684	Peroxidase	14	3	1.8	0.576	8.2
2_7192bF3 [c]	402	Peroxidase 3' UTR	22	0	0	-	2.5
2_8705 [d, e]	810	Unknown (SESPY)	29	5	38.16	2.032*	6.7
2_7407	757	RNA Binding Domain (RBD)	23	5	2.6	-1.33	2.3
Wild random [f]	750		8.4	9.8	5.3	-0.024	12.7
Cultivated [g]	307.27	29,186bp total	24.7		2.25	-0.001	1.18
St. Dev.	73.69		2.4		4.00	0.755	2.01

Table 3.3 (Continued)

n: sample size (chromosomes); π: silent nucleotide diversity (Nei 1987) x 1000; D: Tajima's D (Tajima 1989); Div: net nucleotide divergence (Nei 1987) between *S. bicolor* and *S. propinquum* x 100, corrected using Jukes-Cantor method in DnaSP.

HED: highly expressed, divergent compared with rice; i.e., E value > 10^{-50}. Statistical significance: *, P < 0.05, based on coalescent simulations using DnaSP. Tajima's D 95% Confidence interval: -1.76 to 2.07.

a. Divergence to *Sorghum macrospermum*, instead of *S. propinquum*, which did not amplify.

b. All the variation is in the 5' intron, with no structure due to geography, but with a few high-frequency haplotypes. Both 2_11684F1 and F2 contain coding regions, but F2 is upstream of F1.

c. There are 8 fixed differences with respect to *S. propinquum* in this 3' UTR.

d. Divergence calculated to *S. laxiflorum*, the closest species after *S. propinquum*, which did not amplify.

e. The region used to calculate Tajima's D was the internal intron.

f. Summary statistics from random loci from across the *S. bicolor* genome for wild accessions (Hamblin et al. 2005).

g. Summary statistics from random loci for cultivated accessions (Hamblin et al. 2004).

Based on the recently available *Sorghum bicolor* genome sequence (Paterson et al. 2009), it appears that several of the best candidates (e.g. 2_11508, 2_8981, Table 3.2) are duplicated or belong to small gene family clusters. Interestingly, 2_13461 and 2_13771 are similar, adjacent genes, suggesting a recent duplication with posterior diversification. On the contrary, two of the best candidates, 2_8705 and 2_7407, which show high levels of expression under biotic stress, are single copy in the genome of sorghum and other grasses. One of these genes, 2_8705, named SESPY due to a conserved domain in the cereals (Zamora et al. 2009), is of particular interest since it is highly induced by the pathogen in both the compatible and incompatible reactions, has orthologs in other grass species, and so far its function is unknown. SESPY and other three genes analyzed here, the Class III peroxidase (2_7192), the barley *mlo* homolog (2_11684) and the thaumatin (2_8981), were previously shown to evolve rapidly and under positive selection at several amino acid sites (Zamora et al. 2009).

Known and candidate disease response genes show signals of selection in current populations

We designed primers for 20 candidate genes coming from both the 12 best HED genes and eight of the candidates with orthologs across the grasses previously described (Zamora et al. 2009). Of these, we were able to obtain high quality sequences in 7 to 29 wild accessions for five genes (Table 3.3). In three of these, 2_8981, 2_11684 and 2_7192, we were able to sequence two regions of the gene that provided more information on their polymorphism in coding regions, including introns or UTRs. In the case of 2_8981, these two regions overlapped so we could join and analyze them as one. Up to five primer sets had to be designed for some loci arranged in clusters of paralogs (e.g. thaumatin 2_8981, peroxidase 2_7192) in

order to get specific amplification and high quality sequences for the genes presented here. On the contrary, single copy genes SESPY and RBD, were easy to amplify and yielded high quality sequences across the diverse panel of accessions.

Several of the genes analyzed here appear to come from the distal portion of the two tails of the distribution of variation, showing either an excess or a dearth of polymorphism, when compared with the values obtained from a large random sample of sequences from across the sorghum genome from cultivated accessions (Hamblin et al. 2004; Hamblin et al. 2006) and from wild accessions (Hamblin et al. 2005). For instance, 2_7407 shows a reduced level of genetic diversity in our wild sorghum panel, similar to the average of a random collection of cultivated *S. bicolor* loci (Hamblin et al. 2004), while the peroxidase (2_7192) shows only minimal variation in the two regions we sequenced (Table 3.2). It should be noted that, although Hamblin et al. (2004) used mostly elite materials and landraces in their panel, they strived to include the most divergent accessions. Cultivated sorghum exhibits, however, substantially less variation than the wild sorghum population from across the whole African continent (Casa et al. 2005). Therefore it is not unexpected to find loci with greater variation in the current study than those previously reported (Hamblin et al. 2004), but the level of polymorphism found in this study for 2_8981 and 2_11684 is one standard deviation greater than in the reference study, and that of SESPY 2_8705 is over two standard deviations higher. Moreover, given the greater genetic diversity in the accessions used, both at microsatellites (Casa et al. 2005) and as described here by sequencing, it is surprising to find loci with few segregating sites, as is the case of 2_7407 (an un-annotated gene with two RNA Binding Domains –RBDs-) and 2_7192 (peroxidase), respectively. In 2_7192, the accessions used include 14 of the most divergent accessions in our panel, coming from differentiated clusters in the analysis by Casa

et al. (2005), which contain the overwhelming majority of the genetic variation. Such a reduced level of variation suggests the effect of a cross-species selective sweep in 2_7192 loci and is consistent with the observation described above of rapid divergence, increased expression due to pathogen attack, and a dN/dS ratio showing positive selection at several codons (Zamora et al. 2009). Additionally, the 2_7192 peroxidase enzyme is completely monomorphic across the African population of wild *Sorghum bicolor*, having a single synonymous substitution segregating at moderate frequency. This gene has only three segregating sites and an indel. Two of the SNPs and the indel are located in the second intron and are in complete linkage disequilibrium. One haplotype, LWA72 from Sudan, has an ancestral A instead of a T in position 158, and is the only one different, apparently due to recombination. The 6bp insertion is a derived character, missing in *S. laxiflorum*, and starts at position 720 of the genomic sequence of Sb02g042860.1 (Phytozome). It is due to a duplication of an AGTTCC region and produced two almost identical haplotypes, with 3 differences in a single block (indel and 2 SNPs). There is one silent SNP in the 3' UTR where three individual accessions show an additional T at the end of a polyT repeat. The genealogy of the peroxidase is star-shaped, indicating that all alleles are very similar and that these mutations may have originated after the selective event that involved the whole population of *Sorghum bicolor* (Figure 3.3). Furthermore, the number of replacement fixed differences is high when comparing this *S. bicolor* locus to *Zea mays*, suggesting a recent selective sweep (Table 3.4).

Table 3.4. Replacement and synonymous polymorphism in *S. bicolor* and fixed differences with *S. propinquum*, other wild *Sorghum* species and maize, used for McDonald-Kreitman (1991) tests of neutral evolution of the PS-DRGs.

	Polymorphism		*S. propinquum*		*Sorghum spp.*		*Zea mays*		Total fixed Differences	
	R	S	R	S	R	S	R	S	R	S
2_8981	2	8	0	1	-	-	53.17	43.83	53.17	44.83
2_11684F1	2	2	-	-	9[b]	12	14.5	14.5	23.5	26.5
2_7192F2	0 [a]	1	1	2	3	8	28	29	32	39
2_8705	1	4	-	-	23*[c]	9	4	2	27*	11
2_7407	1	3	3	3	-	-	-	-	-	-
Total	6	17	4	6	35*	29	99.67*	89.33	**135.67***	**121.33**
52 random loci[d]	37	37	47	78	-	-	-	-	47	78

Table 3.4 (Continued)

a. 2_7192 peroxidase enzyme is completely monomorphic across the African population of wild *Sorghum bicolor*, and the number of replacement fixed differences is high when comparing to *Zea mays*, suggesting a recent selective sweep.

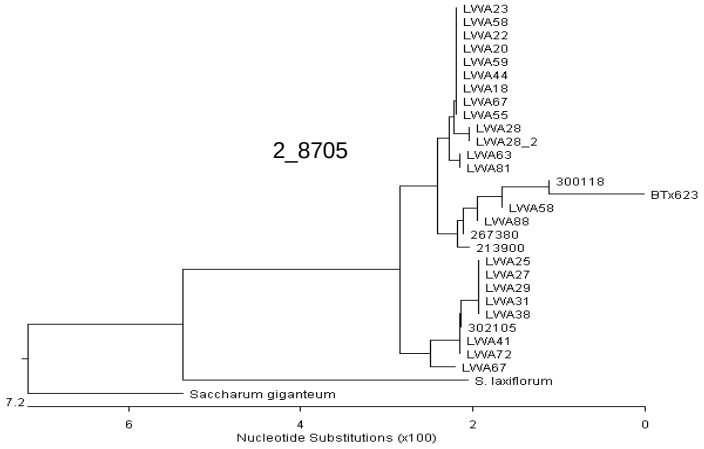
b. 2_11684 was compared to *Sorghum macrospermum* and *Saccharum giganteum*, as no *S. propinquum* sequences were obtained for this locus.

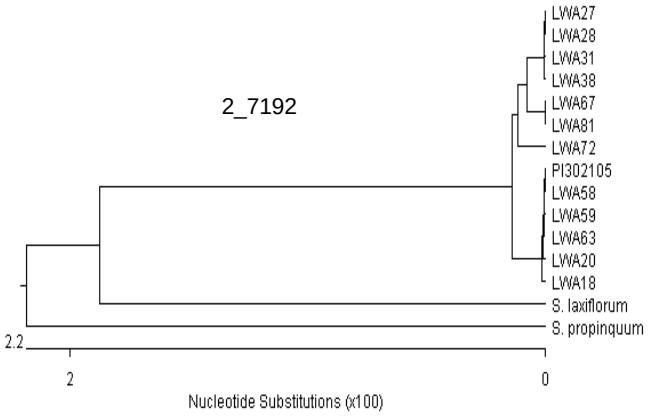
c. These are the sum of replacement and synonymous substitutions observed in several Australian wild *Sorghum* species and *S. giganteum*.

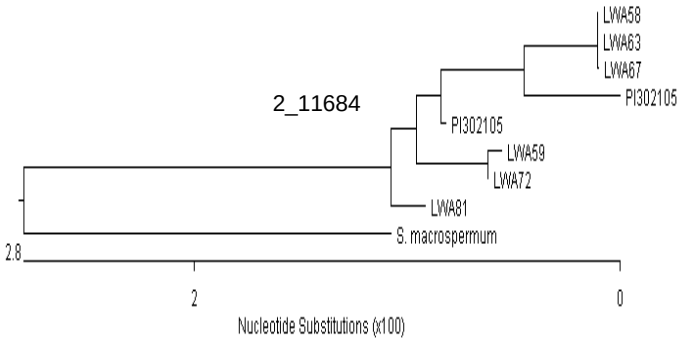
Significance: * ($p<0.05$), ** ($p<0.01$), *** ($p<0.001$)

d. Random loci in the genome of *Sorghum bicolor* studied by Hamblin et al. (2004).

Figure 3.1. Haplotype genealogies of DRGs. A. Thaumatin, B. Peroxidase, C. SESPY, D. RBD. Note that the most divergent haplotype of SESPY is that from an inbred cultivated line BTx623 (see discussion).


2_8705
LWA23
LWA58
LWA22
LWA20
LWA59
LWA44
LWA18
LWA67
LWA55
LWA28
LWA28_2
LWA63
LWA81
300118
BTx623
LWA58
LWA88
267380
213900
LWA25
LWA27
LWA29
LWA31
LWA38
302105
LWA41
LWA72
LWA67
S. laxiflorum
Saccharum giganteum
7.2
6
4
2
0
Nucleotide Substitutions (x100)


2_7192
LWA27
LWA28
LWA31
LWA38
LWA67
LWA81
LWA72
PI302105
LWA58
LWA59
LWA63
LWA20
LWA18
S. laxiflorum
S. propinquum
2.2
2
0
Nucleotide Substitutions (x100)


2_11684
LWA58
LWA63
LWA67
PI302105
PI302105
LWA59
LWA72
LWA81
S. macrospermum
2.8
2
0
Nucleotide Substitutions (x100)

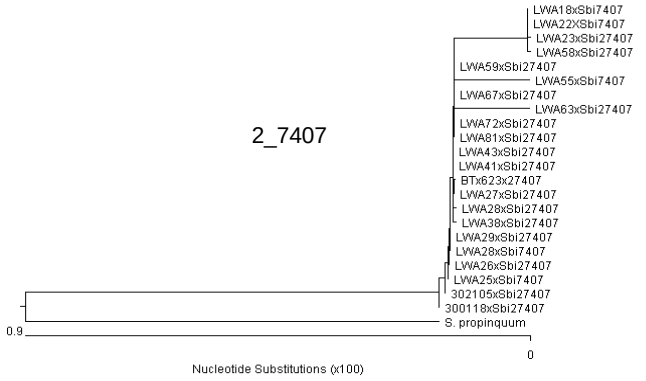
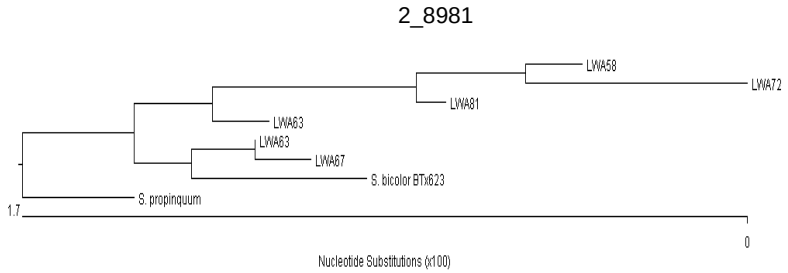
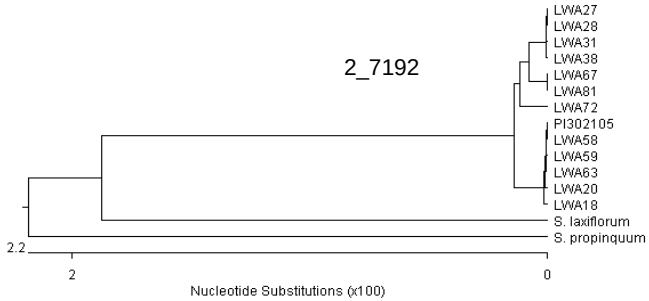

2_7407
LWA18xSbi7407
LWA22xSbi7407
LWA23xSbi27407
LWA58xSbi27407
LWA59xSbi27407
LWA55xSbi7407
LWA67xSbi27407
LWA63xSbi27407
LWA72xSbi27407
LWA81xSbi27407
LWA43xSbi27407
LWA41xSbi27407
BTx623x27407
LWA27xSbi27407
LWA28xSbi27407
LWA38xSbi27407
LWA29xSbi27407
LWA28xSbi7407
LWA26xSbi27407
LWA25xSbi7407
302105xSbi27407
300118xSbi27407
S. propinquum
0.9
0
Nucleotide Substitutions (x100)

Figure 3.1 (Continued)





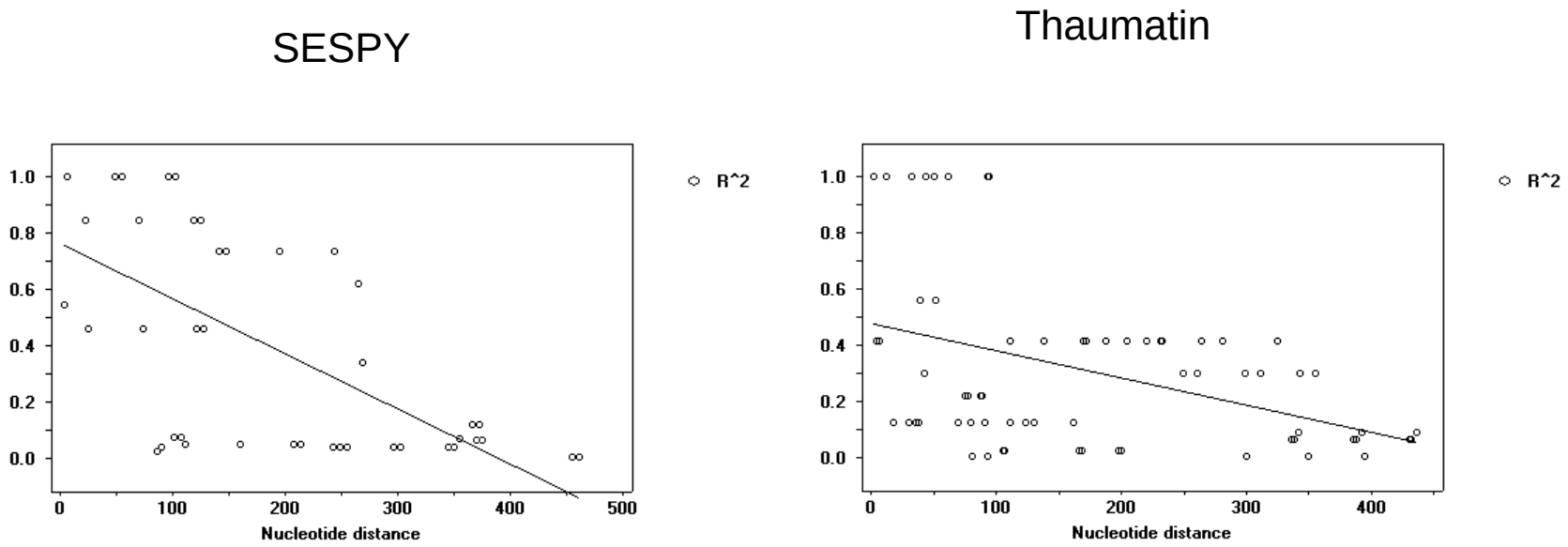


Figure 3.2. Linkage disequilibrium in two positively selected Disease Response Genes. A: SESPY, B: Thaumatin.

The RBD gene, 2_7407, also shows a very reduced polymorphism compared with that of random loci (Hamblin et al. 2004) and has a significantly negative Tajima's D, which indicates there is an excess of rare alleles in the sample, consistent with a directional selection event. This 665 residue protein is similar to *A. thaliana* RanGAP1, its closest match in that species, but the similarity occurs only in the C-terminal region from residues 388 to 664 (44% aa identity), where the two RNA binding domains (RBD) are located. The conserved WPP domain and the LRR, characteristic of RanGAP proteins, are missing from the N-terminus of 2_7407, and therefore this sorghum gene doesn't appear to have a full-length ortholog in *Arabidopsis* (or other dicots). Nevertheless, 2_7407 has very similar orthologs in sugarcane, maize, and rice, aligning well at both ends and including large highly conserved regions, suggesting an origin of this gene in the Poaceae, after the divergence of the dicots. Some smaller regions and 5% of amino acid sites within the conserved regions are hyper-variable, due to positive selection (codeml M7 vs M8, p = 3.8E-15, $\omega = 3.92$), determined as described elsewhere (Zamora et al. 2009). It seems therefore that this gene is unique to monocots, or that it has diverged so much it is not possible to find an ortholog in the dicots simply by using BLAST comparisons. Its similarity to RanGAP suggests, nevertheless, a role in nucleocytoplasmic transport or proteins and nucleic acids. RanGAP2 interacts with NBS-LRR canonical disease resistance (R) proteins both in *Arabidopsis* and *Nicotiana tabacum*, and is thought to be involved in recognition of pathogen ligands, as well as in the first steps of signal transduction leading to the earliest responses to pathogen attack (Sacco, Mansoor, and Moffett 2007).

The reduced diversity observed at both 2_7407 and 2_7192 might be explained by a recent bottleneck in the wild *S. bicolor* population, but the high polymorphism seen in neutral markers and other genes eliminate that possibility. Another explanation might

be the effect of purifying selection on these two genes, and eliminating this possibility requires the analysis of polymorphism at these loci at other species. In order to assess this, we used GRAMENE to study the sequence variation between the two rice sub-species. There are 24 SNPs between *indica* and *japonica* in the 4Kbp surrounding the NP_001060628.1 locus (orthologous to the *S. bicolor* peroxidase described here), including two synonymous and one non-synonymous substitutions, as well as one splice site variation. Fifteen of these variations occur in the 3' region of this gene. Other peroxidase genes in this cluster show similar levels of variation, although interestingly an unrelated neighboring locus, LOC_Os07g_48000, shows very few SNPs between these two genomes. Additionally, a BLASTN in PANZEA using the *Z. mays* cds NM_001112549 as query, indicated the position of the peroxidase ortholog in maize as Chr. 7 (BAC AC211735, contig 325), and a search of the reported polymorphism for this contig indicates that the locus has an average level of variation in maize. The evidence from rice and maize suggests that there are no intrinsic limits to the polymorphism in this locus and supports the idea that there was a selective sweep in the orthologous locus in sorghum. Furthermore, there is no correlation between polymorphism and divergence at these loci, as expected under neutrality (Tables 3.3 and 3.4). The low polymorphism and high divergence observed (compared with average values) suggest that these genes have evolved for long periods of time under positive selection, and are in agreement with the historical pattern of evolution in the cereals (Zamora et al. 2009).

On the other hand, two genes, 2_8705 and 2_11684 (barley *Mlo* homolog), showed both an excess of high-frequency polymorphisms and few divergent haplotypes (Figure 3.1), which are signals of some type of balancing selection, as evidenced by a significantly positive Tajima's D. The locus 2_8705 (SESPY) is a single copy gene with orthologs in the grasses and in *Arabidopsis*. This gene has several signals of

selection including a significantly positive Tajima's D (2.03, $p<0.05$), which is evidence for balancing selection. Tajima's D was calculated using the intron for this gene to maximize the polymorphism available from the samples and increase the power of the test. Twenty-nine diverse accessions were included in this test and the intron is 270 bp. Additionally, using all the sequence information for SESPY we found evidence of low linkage disequilibrium and at least one recombination event (Figure 3.2). Remarkably, the most divergent haplotype found at this locus belongs to BTx623, a highly inbred elite line used for hybrid cultivar development. Klein et al. (2008) showed that the whole genome of BTx623 is a mosaic of two widely divergent lines (a Kafir from South Africa, and a Zera Zera from Ethiopia), and the divergent haplotype for SESPY is the result of recombination between the alleles from those two lines. This suggests that the use of wide crosses in breeding is not only useful to introgress extant alleles, but that *de novo* generation of alleles occurs through recombination, potentially generating superior alleles for disease resistance. Additionally, a McDonald-Kreitman test indicated that the SESPY gene has a significant excess of replacement fixed differences, which is also a signal of recurrent selective sweep events (Table 3.4). Although not significant, the pattern of diversity for the ortholog of 2_8705 in the maize diversity panel also shows several frequent haplotypes, and in a comparison with *Tripsacum sp.*, the McDonald-Kreitman analysis of polymorphism and divergence was also significant due to an excess of fixed differences (G test, p-value = 0.04177), providing evidence of divergent evolution by recurrent selective sweeps.

The other locus showing the signal of balancing selection, 2_11684, is a homolog of the barley *Mlo* gene and has several differentiated haplotypes (Figure 3.1). In the 5' UTR (upstream part of 2_11684F2) there are several polymorphisms, but there are no mutations in the two exons surveyed. There are no differences with respect to *S.*

propinquum in the coding regions and only two fixed differences in non-coding, one in the 5' UTR and one in the first intron. It has a positive Tajima's D of 2.03, showing signs of balancing selection. There is no genetic structure observed in the diverse sorghum population at this gene.

Finally, while it is not significant for Tajima's D, 2_8981, a thaumatin homolog, also shows far greater variation than that seen at random loci, with several divergent haplotypes (Figure 3.1). It also shows evidence of at least one event of recombination, as well as low linkage disequilibrium (Figure 3.2).

HED candidate genes have an excess of non-synonymous fixed differences with respect to other grass species

Comparison of these candidate genes with *S. propinquum*, other *Sorghum* species and their orthologs in *Zea mays* (Table 3.4) showed that these genes generally exhibit an excess of replacement substitutions, consistent with a role of positive Darwinian selection in their rapid divergence. Interestingly, when comparing the candidate genes with *S. propinquum* only, there are no significant differences from the pattern expected under neutral variation. It appears that *S. propinquum*, as has been suggested before by Hamblin et al. (2004, 2006), is too closely related to *S. bicolor*, showing little divergence and therefore providing correspondingly low statistical power to detect selection. However, when using more divergent *Sorghum* species or *Zea mays*, the excess of non-synonymous fixed differences relative to non-synonymous polymorphism becomes statistically significant (Table 3.4), and seems to indicate that at longer time scales the effect of positive selection, taking to fixation the advantageous alleles, leaves a signature that can be detected using this method. It is also worth noting that, when comparing with maize, the number of synonymous and non-synonymous substitutions is quite similar, which would suggest a high mutation rate at these loci. There is, however, recent evidence that synonymous mutations may

not be silent and could potentially have an effect on the resistance phenotype (Komar, Lesnik, and Reiss 1999). Taken together, the information available for these genes indicates that they have evolved adaptively for over 200 million years, which is the sum of the lines in the phylogeny of the species studied (Zamora et al. 2009), and are still under selective pressure, as determined by their current non-neutral patterns of polymorphism.

Discussion

To identify the candidate genes presented in this study, we combined gene expression profiles, in terms of total level of expression and EST-library specificity, with inter-specific divergence between sorghum and rice. We did this under the hypotheses that, first, many disease response genes are up-regulated as a response to pathogen attack, since this is an energetically efficient way to deal with many different pathogenic taxa, some of which require unique responses (De Vos et al. 2005), and second, genes involved in disease resistance evolve rapidly due to the interaction of their protein products with pathogen elicitors, and the strong and constant selective pressure to survive and reproduce in the presence of pathogens. Using this approach we identified a set of genes showing high conditional expression during biotic stress, moderate to high divergence between sorghum and rice, and in some cases prior annotation suggesting a role in disease resistance. A few of the candidate genes identified in this way had been shown previously by us (Zamora et al. 2009) to be genes showing historical signals of adaptation.

To test whether these genes show signals of recent or current selection, we analyzed the pattern of polymorphism and divergence of five of these known and putative disease response genes in a highly diverse panel consisting of wild *Sorghum bicolor* accessions, with reference to that of many random sequences and found that most of

our candidate genes show a pattern of nucleotide diversity that is significantly different from that expected under neutrality, as observed for the set of random loci (Hamblin et al. 2004; Hamblin et al. 2006). There is strong evidence of diversifying selection in two of these genes, consistent with what has been found in disease response genes in several other species. Other two genes show minimal genetic diversity, significantly different from that expected under neutrality, particularly in a panel characterized by its maximum diversity within *S. bicolor*, suggesting the effect of a recent and strong species-wide selective sweep. The apparent selective sweep seen in the *Sorghum bicolor* peroxidase (2_7192) is in agreement with our previous finding that this gene has evolved, at least within the Poaceae, through recurrent events of positive selection, driven by the advantageous effect of non-synonymous substitutions (Zamora et al. 2009). Background selection (Charlesworth, Morgan, and Charlesworth 1993) may be a possible cause of reduced variation in a locus, by eliminating neutral variation surrounding a deleterious mutation. However, this usually happens in regions of low recombination, and although we don't have an estimate of the level of recombination around the loci found here to have limited variation, we expect that the region with the peroxidase locus should exhibit a moderate to high level of recombination (Akhunov et al. 2003). The position of the peroxidase gene cluster that includes 2_7192 in the *S. bicolor* genome, is close to the telomere (Sb02g042860, Chr. 2 at 76.6Mbp), is also typical of regions of high recombination frequency. This cluster shows variation in number with respect to rice and the presence of pseudogenes (Zamora et al. 2009), suggesting the effect of recombination in the birth and death of paralogs in this cluster, according to the model by Michelmore and Meyers (1998). Also, the neutralist explanation of high purifying selection acting at these loci requires that divergence to other species should be low (Kimura 1985), and that is not the case for either the peroxidase (2_7192) or the RBD

(2_7407) genes. In fact, the reduction in variation in both loci is quite substantial when compared with the average diversity for many random regions in the genome (Table 3.3), and its divergence to other species is very high (Table 3.4). This suggests that these genes have evolved through repeated events of directional selection.

Non-equilibrium demographic processes such as a recent population bottleneck or rapid population expansions may also lead to patterns of diversity similar to those caused by directional selection, although the former have an effect across the genome (Caicedo et al. 2007). However, since we used a highly diverse panel of wild accessions, we can rule these possibilities as highly unlikely. Wild *Sorghum bicolor* is an abundant and dominant species in many climax grassland ecosystems all across Africa. Furthermore, *S. bicolor* has several sub-species, adapted to various microhabitats (BOSTID 1996), and therefore there is no reason to believe it has gone through recent whole population crashes or expansions. Its neutral genetic diversity, measured by microsatellites (Casa et al. 2005), and several genes (Hamblin et al. 2005), including some described in this study, reflect its high genetic diversity. Therefore, reduced polymorphism is likely due to positive selection, which is consistent in both of these loci with their degree of divergence to other species. Indeed, a role in disease resistance has been amply demonstrated for class III peroxidases, due to their importance in basal defense response and its interaction with pathogen-generated molecules (Ye, Pan, and Kuc 1989; Johrde and Schweizer 2008).

The identification of signals of selection in wild accessions of Sorghum bicolor is possible given its greater diversity

Previous studies have used genome wide approaches to identify signals of positive selection in *Sorghum bicolor* (Hamblin et al. 2004; Hamblin et al. 2006). However, the accession panels used in these studies consisted mainly of cultivated accessions,

which have a reduced diversity compared to wild accessions (Casa et al. 2005) and only a fourth of that observed in maize (Wright et al. 2005; Hamblin et al. 2006). The limited genetic diversity, a common outcome of domestication and breeding, obscures the signals of selective events that undoubtedly have occurred in cultivated sorghum. Moreover, the random sequencing of DNA regions from the genome conducted in these studies covered only a small portion of the sorghum genome. In contrast, the approach used in this study took advantage of the greater polymorphism found in wild accessions of *Sorghum bicolor* and was directed towards disease response genes, which show signals of selection in many different taxa.

Although it would seem that genes involved in disease resistance have an almost certainty of showing signals of directional or diversifying selection, a large selection of *Arabidopsis thaliana* genes involved in signal transduction pathways show predominantly the effect of purifying selection (Bakker et al. 2008). Those genes may have been optimized early in evolution and may now be ruled by negative selection. It is tempting to speculate that the protein products of those conserved genes may be the guardees of R genes designed to detect the attack on signal transduction pathways (Caldwell and Michelmore 2009). Therefore, it is evident that there is a highly conserved component of disease resistance mechanisms, which we missed by using our present strategy. However, it is also clear that there are many proteins that interact directly with pathogen-derived molecules and whose pattern of evolution fits that predicted by one of the various models of host-pathogen antagonistic coevolution. Genes evolving adaptively due to an arms-race, pleiotropy or any other balancing selection processes are likely to show signals of selection in current populations, as well as historical signals of adaptive evolution. A few genes shown here as having patterns of genetic diversity consistent with directional or diversifying selection in current populations, were described previously by Zamora et al. (2009) as having

codons with a high dN/dS rate ratio when comparing orthologs of cereal species, a strong sign of recurrent events of positive selection. Although demographic events can generate patterns of polymorphism falsely suggesting selection, the combination of historical and current signals of selection (in that order) has several advantages. It facilitates the identification of candidate genes, reduces the number of genes to sequence, decreases the number of false positives, and when both sources of evidence agree, argues in favor of a selective hypothesis.

One additional reason we were able to identify loci showing reduced diversity is that our candidates came from expression profile studies and not from QTL analyses. As Bergelson et al. (2001) pointed out, loci derived from QTL studies are biased for polymorphism in the species in question. Loci that have been recently fixed throughout the whole species due to their fundamental importance and strong selective pressure would not appear as important genes in a QTL analysis (e.g. peroxidase). It is also possible that QTL studies generated by crossing highly inbred lines lead to the identification of recessive mutations that are not frequent in cultivated or wild populations. Genes under balancing selection should appear in both types of study, and in general, both kinds of analyses, QTL and differential expression profiling, are complementary.

Balancing selection results from multiple population-level processes in the host-pathogen interaction

A recent literary survey (Tiffin and Moeller 2006) and an empirical study (Caldwell and Michelmore 2009) show that the different components of molecular disease resistance differ with respect to the dominant form of selection displayed. However, most plant disease response genes for which patterns of polymorphism have been studied show the effect of some kind of balancing selection (Caicedo, Schaal, and

Kunkel 1999; Stahl et al. 1999; Tian et al. 2002; Mauricio et al. 2003; Shen, Francki, and Ohm 2006). Correspondingly, Ferrer-Atmella et al. (2008) found several cases of balancing selection in human innate defense response genes and argued it is the main evolutionary pattern.

Several processes can interfere with the development of a full selective sweep or otherwise generate a pattern consistent with balancing selection, including tight linkage between deleterious and adaptive mutations, pleiotropic effects, epistatic interactions, overdominance, spatial and temporal variations in the direction of selection, and frequency dependent selection. Tight linkage between deleterious and adaptive mutations, also referred to as the Hill-Robertson interference (McVean and Charlesworth 2000), is an interplay between the adaptive value of the allele and the level of recombination around it. Clearly, this effect would be important in regions of low recombination and particularly when the selective strength is weak. That doesn't appear to be the case in the genomic regions where the DRGs studied here are located, in particular the *mlo* locus (2_1684), located near the telomere in chromosome 1, in which recombination is expected to be high (Akhunov et al. 2003). All other genes studied here, except 2_7407, are located in sub-telomeric regions (data not shown) and others like 2_7192 and 2_8981 belong to clusters of genes, which likely evolve rapidly and adaptively through point mutations and recombination events (Parniske et al. 1997; Michelmore and Meyers 1998; Mondragón-Palomino and Gaut 2005). Clearly, the increase in fitness related to a new allele that confers a higher level of resistance at any of these loci is likely to be significant and the high recombination rate could indeed facilitate its rapid increase in frequency. Negative frequency dependent selection has been proposed as a viable mechanism that generates multiple alleles at a disease response locus (Borghans, Beltman, and De Boer 2004) and may be responsible for the high level of variation at the thaumatin and SESPY loci.

Conversely, pleiotropic effects of the allele under selection can make it difficult for its frequency to increase to fixation and will cause a quick reversal in frequency when the selective pressure decreases (Curtis, Cook, and Wood 1978). In the case of the barley *mlo* gene (orthologous to the sorghum 2_11684) it has been shown that having a knock-out mutation makes the defenses constitutively active (Büschges et al. 1997). This makes the plant more resistant when the biotrophic fungal pathogen *Blumeria graminis* is present, but in the absence of this pathogen, plants show reduced fitness due to the excess of energy wasted in the constitutive expression of defenses (Kjær et al. 1990). There is an additional trade-off since necrotrophic pathogens take advantage of the increased Hypersensitive Response (HR) and are able to infect *mlo* plants with greater success (Jarosch, Kogel, and Schaffrath 1999). Such opposing selective forces should lead to a pattern of diversity consistent with balancing selection at the *mlo* locus, and others with similar functions. For instance, in *Sorghum bicolor*, the Milo disease is caused by a host specific toxin produced by the necrotrophic pathogen *Periconia circinata.* This toxin targets an NBS-LRR gene, causing widespread necrosis due to out-of-control hypersensitive response (HR) (Ransom et al. 1992; Nagy et al. 2007). Susceptible and resistant forms alternate spontaneously in the sorghum population, suggesting the effect of recombination in the cluster of R genes where the target gene, *Pc*, is located (Nagy et al. 2007).

Therefore, there are many possible mechanisms that can lead to balancing selection and the genes showing such a pattern are likely to be functionally important. Tian et al. (2002) found that, in *A. thaliana*, the signal of increased allelic diversity and linkage disequilibrium extends approximately 10Kbp around the RPS5 locus, and suggested that, in this species, it would be possible to identify all loci that show old polymorphisms. A similar strategy could be used in sorghum where the linkage disequilibrium extends in average several kilobases but rarely extends more that 15 Kb

(Hamblin et al. 2005). The recent sequencing of the BTx623 *Sorghum bicolor* genome and the availability of highly diverse wild accessions make that strategy very feasible.

Selective sweeps are often partial and its signal is ephemeral

Evolutionary theory predicts a never-ending molecular race between plants and their pathogens and this process should generate many signals of selective sweeps. In fact, the high divergence values of some of the genes described here, as well as the high values of dN/dS ratios at some of the same genes (Zamora et al. 2009), provided evidence of recurrent events of positive selection, *i.e.*, multiple replacement substitutions occurring over millions of years. However, for one of several reasons, very few disease response genes show the signal of selective sweeps when the standing intraspecific polymorphism is studied. First, selective sweeps are discrete and ephemeral events. If a superior allele arises and the selective pressure is high enough, it will increase to fixation within a few generations. Later, mutations start to accumulate, increasing polymorphism at the flanks, and recombination erodes the linkage disequilibrium (Przeworski 2002). Both of these forces erase the signal of a selective sweep at a speed determined by the mutation rate, the effective population size (Ne), and the local recombination rate. The magnitude of the sweep or the extent of polymorphism reduction it caused, and the LD generated around the selected site also determine how fast the signal disappears. Second, sweeps can be soft when the selective pressure is not too high, not constant, or disappears at some point. Third, sweeps can be partial in time or in space, making it difficult or impossible to differentiate them from variation patterns produced under neutrality or non-equilibrium demographic processes. For instance, a selective process that started very recently may be occurring in the present time, with only a portion of the population

having the superior allele so far. Alternatively, the sweep may be restricted to a subpopulation that was not recognized as such in the analysis. In wild accessions of *Sorghum bicolor*, microsatellite-based analyses (Casa et al. 2005) and rapidly evolving genes expressed only in pollen show genetic structure between the Northern and Southern parts of Africa (our own unpublished data). However, there was no significant genetic structure in the genes presented here, which suggests that 2_11684 and 2_8705 present signals of balancing selection and that there are no partial sweeps.

Evidence for recent or on-going selection in S. bicolor disease response genes

The patterns of polymorphism and divergence found in the DRGs presented here suggest the effect of an active interaction between sorghum and one or multiple pathogens. Although these genes were identified using expression profiles generated by inoculating plants with *Colletotrichum sublineolum,* the fungal pathogen that causes anthracnose, the genes studied here could potentially be used in defense against other pathogens. Over two dozen fungal pathogens have been described for sorghum alone, and the number of bacterial and viral pathogens described is also increasing along with research in this species, all of which indicates that DRGs are needed constantly and possibly interact with many pathogen-derived effector molecules. For instance, Rosebrock et al. (2007) demonstrated that a bacterial E3 ubiquitin ligase physically interacts with a host protein kinase and disrupts an important signaling pathway. Additionally, it has been recently shown that tomato glucanases and *Phythophtora infestans* glucanase inhibitor proteins interact directly, that their tridimensional structures match, and that the antagonistic interaction between these organisms has lead to rapid, adaptive evolution of these genes and to non-conservative amino acid substitutions in the interacting positions of proteins from both pathogen and host (Damasceno et al. 2008). All this suggests that the DRGs described here, and

others that can be identified using the same strategy, may be interesting candidates for functional analyses geared to identifying the interacting molecules derived from currently important pathogens. Thaumatins, peroxidases and the barley *mlo* homolog have already been determined to be essential in defense against a broad range of fungal pathogens in several plant species. Conversely, the single copy RBD and SESPY genes have both historical and current evidence of positive selection, but their role in defense and their interaction with pathogen-derived ligands needs to be functionally validated.

It is also essential to test for differences in the disease resistance phenotype using wild accessions having non-synonymous variation at these loci, and to determine the fitness effect of different alleles for molecular breeding purposes. However, given the preponderance of balancing selection in nature, we propose that using crop varieties, e.g. hybrid multilines, that maximize the genetic variation at DRGs, while being homogeneous for other characters of interest, should reduce the loses due to pathogens and decrease the frequency and effect of epidemics.

Conclusions

Combining gene expression profiles and divergence patterns is a useful strategy to identify genes that have a role in disease response and that evolve rapidly through recurrent events of positive selection.

The use of highly diverse wild accessions of *Sorghum bicolor* increases the statistical power in the analyses and allows the detection of loci with patterns of polymorphism suggesting the effect of positive selection in the antagonistic co-evolution between host plant and fungal pathogens.

The information from historical and current signals of selection for the same genes provides strong evidence in favor of a selective hypothesis and facilitates the identification of candidates for functional validation studies and breeding.

In nature, a variety of selective mechanisms maintain multiple alleles at disease response loci, opposed to what we typically do in mainstream agriculture, where monocultures dominate, favoring the pathogens.

In *Sorghum bicolor,* the genomic information currently available, along with the wild germplasm, would permit a directed search for patterns of selection in genes involved in disease resistance, both with significantly high and low polymorphism.

Acknowledgements

The authors want to thank Martha Hamblin for her critical review of the manuscript. William Rooney kindly provided seed from some of the accessions. We are grateful to H.J. Price for providing seed for the various wild *Sorghum* species. AZM is supported by the Costa Rica-USA Foundation for Cooperation (CRUSA) through a special Fulbright scholarship in biotechnology and by the Institute of Genomic Diversity (Cornell University, Ithaca, NY). This research was funded by National Science Foundation grant No. 0115903 to SK.

REFERENCES

Akhunov, E. D., A. W. Goodyear, S. Geng et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Research **13**:753-763.

Allison, A. C. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. Brit. Med. J. **1**:290.

Bakker, E. G., C. Toomajian, M. Kreitman, and J. Bergelson. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. Plant Cell **18**:1803-1818.

Bakker, E. G., M. B. Traw, C. Toomajian, M. Kreitman, and J. Bergelson. 2008. Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana*. Genetics **178**:2031-2043.

Bergelson, J., M. Kreitman, E. A. Stahl, and D. Tian. 2001. Evolutionary dynamics of plant R-genes. Science **292**:2281-2285.

Borghans, J. A. M., J. B. Beltman, and R. J. De Boer. 2004. MHC polymorphism under host-pathogen coevolution. Immunogenetics **55**.

BOSTID. 1996. Lost Crops of Africa: Grains. Washington, D.C.:National Academy Press.

Büschges, R., K. Hollricher, R. Panstruga et al. 1997. The barley *mlo* gene: a novel control element of plant pathogen resistance. Cell **88**: 695-705.

Caicedo, A. L., B. A. Schaal, and B. N. Kunkel. 1999. Diversity and molecular evolution of the RPS2 resistance gene in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America **96**:302-306.

Caicedo, A. L., S. H. Williamson, R. D. Hernandez et al. 2007. Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genet **3**:e163.

Caldwell, K. S., and R. W. Michelmore. 2009. *Arabidopsis thaliana* Genes Encoding Defense Signaling and Recognition Proteins Exhibit Contrasting Evolutionary Dynamics. Genetics **181**:671-684.

Casa, A. M., S. E. Mitchell, M. T. Hamblin, H. Sun, J. E. Bowers, A. H. Paterson, C. F. Aquadro, and S. Kresovich. 2005. Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theoretical and Applied Genetics **111**:23-30.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289-1303.

Curtis, C. F., L. M. Cook, and R. J. Wood. 1978. Selection for and against insecticide resistance and possible methods of inhibiting the evolution of resistance in mosquitoes. Ecological Entomology **3**:273-287.

Damasceno, C. M. B., J. G. Bishop, D. R. Ripoll, J. Win, S. Kamoun, and J. K. C. Rose. 2008. Structure of the glucanase inhibitor protein (GIP) family from *Phytophthora* species suggests coevolution with plant endo-beta-1,3-glucanases. Molecular Plant-Microbe Interactions **21**:820-830.

De Vos, M., V. R. Van Oosten, R. M. P. Van Poecke et al. 2005. Signal Signature and Transcriptome Changes of *Arabidopsis* During Pathogen and Insect Attack. Molecular Plant-Microbe Interactions **18**:923-937.

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small amounts of leaf tissue. Phytochemical Bulletin **19**:11-15.

Ferrer-Admetlla, A., E. Bosch, M. Sikora et al. 2008. Balancing Selection Is the Main Force Shaping the Evolution of Innate Immunity Genes. J Immunol **181**:1315-1322.

Gillespie, J. H. 1975. Natural Selection for Resistance to Epidemics. Ecology **56**:493-495.

Hamblin, M. T., A. M. Casa, H. Sun, S. C. Murray, A. H. Paterson, C. F. Aquadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: Lessons from *Sorghum bicolor*. Genetics **173**:953-964.

Hamblin, M. T., M. G. S. Fernandez, A. M. Casa, S. E. Mitchell, A. H. Paterson, and S. Kresovich. 2005. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics **171**:1247-1256.

Hamblin, M. T., S. E. Mitchell, G. M. White, J. Gallego, R. Kukatla, R. A. Wing, A. H. Paterson, and S. Kresovich. 2004. Comparative Population Genetics of the Panicoid Grasses: Sequence Polymorphism, Linkage Disequilibrium and Selection in a Diverse Sample of *Sorghum bicolor*. Genetics **167**:471-483.

Huang, C.-L., S.-Y. Hwang, Y.-C. Chiang, and T.-P. Lin. 2008. Molecular Evolution of the Pi-ta Gene Resistant to Rice Blast in Wild Rice (*Oryza rufipogon*). Genetics **179**:1527-1538.

Hughes, A. L., T. Ota, and M. Nei. 1990. Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Molecular Biology & Evolution **7**:515-524.

Ingvarsson, P. K. 2005. Molecular population genetics of herbivore-induced protease inhibitor genes in European Aspen (*Populus tremula* L., Salicaceae). Molecular Biology and Evolution **22**:1802-1812.

Jarosch, B., K.-H. Kogel, and U. Schaffrath. 1999. The Ambivalence of the Barley *Mlo* Locus: Mutations Conferring Resistance Against Powdery Mildew (*Blumeria graminis f. sp. hordei*) Enhance Susceptibility to the Rice Blast Fungus *Magnaporthe grisea*. Molecular Plant-Microbe Interactions **12**:508-514.

Johrde, A., and P. Schweizer. 2008. A class III peroxidase specifically expressed in pathogen-attacked barley epidermis contributes to basal resistance. Molecular Plant Pathology **9**:687-696.

Kimura, M. 1985. The Neutral Theory of Molecular Evolution:Cambridge University Press.

Kjær, B., H. P. Jensen, J. Jensen, and J. H. Jørgensen. 1990. Associations between three ml-o powdery mildew resistance genes and agronomic traits in barley. Euphytica **46**:185-193.

Klein, R. R., J. E. Mullet, D. R. Jordan, F. R. Miller, W. L. Rooney, M. A. Menz, C. D. Franks, and P. E. Klein. 2008. The Effect of Tropical Sorghum Conversion and Inbred Development on Genome Diversity as Revealed by High-Resolution Genotyping. Crop sci **48**:S-12-26.

Komar, A. A., T. Lesnik, and C. Reiss. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Letters **462**:387-391.

Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution. Genome Research **13**:2229-2235.

Mauricio, R., E. A. Stahl, T. Korves, D. Tian, M. Kreitman, and J. Bergelson. 2003. Natural selection for polymorphism in the disease resistance gene Rps2 of *Arabidopsis thaliana*.

May, R. M., and R. M. Anderson. 1990. Parasite-host coevolution. Parasitology **100**:S89-S102.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351**:652-654.

McVean, G. A. T., and B. Charlesworth. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155**:929-944.

Michelmore, R. W., and B. C. Meyers. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research **8**:1113-1130.

Mondragón-Palomino, M., and B. S. Gaut. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. Molecular Biology and Evolution **22**:2444-2456.

Nagy, E., T.-C. Lee, W. Ramakrishna et al. 2007. Fine mapping of the *Pc* locus of *Sorghum bicolor* , a gene controlling the reaction to a fungal pathogen and its host-selective toxin. TAG Theoretical and Applied Genetics **114**:961-970.

Pal, C., B. Papp, and L. D. Hurst. 2001. Highly Expressed Genes in Yeast Evolve Slowly. Genetics **158**:927-931.

Parker, M. A. 1990. The pleiotropy theory for polymorphism of disease resistance genes in plants. Evolution **44**:1872-1875.

Parniske, M., K. E. Hammond-Kosack, C. Golstein, C. M. Thomas, D. A. Jones, K. Harrison, B. B. H. Wulff, and J. D. G. Jones. 1997. Novel disease resistance

specificities result from sequence exchange between tandemly repeated genes at the cf-4/9 locus of tomato. Cell **91**:821-832.

Paterson, A. H., J. E. Bowers, R. Bruggmann et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**:551-556.

Posas, F., and H. Saito. 1998. Activation of the yeast SSK2 MAP kinase kinase kinase by the SSK1 two-component response regulator. EMBO J **17**:1385-1394.

Pratt, L. H., C. Liang, M. Shah et al. 2005. Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and skotomorphogenesis from a milestone set of 16,801 unique transcripts. Plant Physiol. **139**:869-884.

Price, H. J., S. L. Dillon, G. Hodnett, W. L. Rooney, L. Ross, and J. S. Johnston. 2005. Genome Evolution in the Genus *Sorghum* (Poaceae). Ann Bot **95**:219-227.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. Genetics **162**:2053-2053.

Ransom, R. F., J. Hipskind, B. Leite, R. L. Nicholson, and L. D. Dunkle. 1992. Effects of an Elicitor from *Colletotrichum graminicola* on the Response of *Sorghum* to *Periconia circinata* and Its Pathotoxin. Physiological and Molecular Plant Pathology **41**:75-84.

Rosebrock, T. R., L. Zeng, J. J. Brady, R. B. Abramovitch, F. Xiao, and G. B. Martin. 2007. A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. Nature **448**:370-374.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for General Users and for Biologist Programmers. Pp. 365-386 *in* K. S, and M. S, eds. Bioinformatics Methods and Protocols. Humana Press, Totowa, NJ.

Sacco, M. A., S. Mansoor, and P. Moffett. 2007. A RanGAP protein physically interacts with the NB-LRR protein Rx, and is required for Rx-mediated viral resistance. The Plant Journal **52**:82-93.

Shen, X. R., M. G. Francki, and H. W. Ohm. 2006. A resistance-like gene identified by EST mapping and its association with a QTL controlling *Fusarium* head blight infection on wheat chromosome 3BS. Genome **49**:631-635.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman, and J. Bergelson. 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. Nature **400**:667-671.

Stekel, D. J., Y. Git, and F. Falciani. 2000. The comparison of gene expression from multiple cDNA libraries. Genome Research **10**:2055-2061.

Tian, D., H. Araki, E. Stahl, J. Bergelson, and M. Kreitman. 2002. Signature of balancing selection in *Arabidopsis*. Proc Natl Acad Sci U S A **99**:11525-11530.

Tiffin, P., and D. A. Moeller. 2006. Molecular evolution of plant immune system genes. Trends in Genetics **22**:662-670.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7**:256-276.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, M. D. McMullen, and B. S. Gaut. 2005. The Effects of Artificial Selection on the Maize Genome. Science **308**:1310-1314.

Ye, X. S., S. Q. Pan, and J. Kuc. 1989. Association of pathogenesis-related proteins peroxidase, beta-1 3 glucanase, and chitinase with induced resistance of tobacco to *Peronospora tabacina* but not to systemic tobacco mosaic virus. Phytopathology **79**:1150.

Zamora, A., Q. Sun, M. T. Hamblin, C. F. Aquadro, and S. Kresovich. 2009. Positively selected disease response orthologous gene sets in the cereals identified using *Sorghum bicolor* L. Moench expression profiles and comparative genomics. Mol Biol Evol **26**:2015-2030.

CHAPTER FOUR

EFFECTS OF PATHOGEN SELECTION PRESSURE ON GENE AND GENOME STRUCTURE IN *SORGHUM BICOLOR*

Abstract

The eukaryotic genome is complex in structure, content and regulation, and we hypothesize that some of these characteristics may be the result of selection for greater adaptability. Plant pathogens collectively inflict a strong and constant selective pressure on their hosts. Several categories of disease response genes have been shown to have a high dN/dS rate ratio, and hence to evolve under positive selection as a consequence of antagonistic coevolution. However, other features of the gene structure, such as number of exons or size of the coding and non-coding regions, as well as the position in the genome have not been studied in detail to determine whether they show any sign of selection. Here we show that genes that are highly upregulated during biotic stress, and which show a moderate to high degree of divergence between sorghum and rice, have significantly more exons than a set of control house-keeping genes. Disease response gene candidates also have significantly more copies per cluster and genome than control genes. The disease response gene candidates also tend to be closer to the telomere than the control genes. Interestingly, genes with more copies are often single exon genes, and several of these have signs of positive selection and are located closer to the telomeric regions. This non-random structure may promote intragenic and intergenic recombination at single copy multi-exon and multicopy single exon genes respectively, thereby increasing the generation of functional diversity and efficiently eliminating damaged or targeted domains. It may also increase the speed of reaction to pathogen attack by having

multiple paralogs in subtelomeric regions. Overall, the differences in structure and chromosomal location of divergent disease response candidates seem to indicate that positive selection has had an effect on genome content, order and regulation.

Introduction

Pathogens are ubiquitous, numerous and diverse, and their collective selective pressure on plants is both unremitting and intense. This selective pressure has left a mark on the patterns of substitutions of genes involved in disease response (Bishop, Dean, and Mitchell-Olds 2000; Mondragón-Palomino et al. 2002; Nielsen et al. 2005; Zamora et al. 2009) as well as on their nucleotide polymorphism and diversity (Mauricio et al. 2003; Bakker et al. 2006; Bakker et al. 2008; Zamora et al. 2009)(see Chapter 3 this volume). Pathogens have a large array of effector molecules that interact with many different molecules of the plant's primary and secondary metabolism (Rosebrock et al. 2007; Yeam et al. 2007), driving their evolution. In the antagonistic coevolution between tomato and its oomycete pathogen *Phytophtora*, recurrent selective sweeps have left a signature of a coevolving molecular arms race in the interacting proteins from both species (Damasceno et al. 2008).

In addition to point mutations that result in non-synonymous adaptive substitutions, insertions and deletions, both disruptive and non-disruptive (*e.g.* 3n indels), have also been shown to be important in adaptation (Podlaha and Zhang 2003) and occur frequently in disease response genes in plants (Zamora et al. 2009). Furthermore, chimeric genes and copy number variation have been shown to occur at several loci of R genes within a single species (Kuang et al. 2008), and also between different species as in the case of peroxidases and thaumatins between sorghum and rice (Zamora et al. 2009).

The effects of transposable elements on adaptation through changes on copy number, gene structure and regulation have also been documented for both plants and pathogens (Gout et al. 2006). For instance, Gout et al. (2006) showed that in the rapeseed fungal pathogen *Leptosphaeria maculans* a retrotransposable element disrupted a putative avirulence factor resulting in the spread of virulence in a period of 3 years. All the factors described above contribute to variations in genome structure that may be positively selected so the plant can more efficiently defend itself from pathogens.

The genetic architecture of disease resistance has been addressed by Wisser et al. (2005), who found that R genes and other gene families are significantly co-localized with disease QTL (dQTL) in rice. However, due to the fact that QTL can generally cover several centimorgans, including several thousand genes, that particular study implicated close to 50% of the genome in disease response. Additionally, Boyko et al. (2002) showed that, in wheat, disease related genes are located in distal telomeric regions, while retrotransposon loci are pericentromeric. However, it is not clear from this study whether non-disease related genes are also located in telomeric regions.

In this study we attempt to determine whether the selective pressure imposed by a multitude of pathogens have had an effect on the structure of disease response genes and the genome itself. We show that a set of highly upregulated and divergent disease response gene candidates differs in its genomic position, as well as in their gene structure, from the non-disease response control set of genes.

Material and Methods

In order to test whether disease response genes have structural features or a genomic position that are significantly different than those of genes not involved in disease resistance, we first developed a set of non-disease related genes to use as controls. We

initially identified all *Sorghum bicolor* UniScripts (unique splice forms of a gene model) expressed as ESTs 30 or more times (n = 432, Figure 4.1) including all the sorghum libraries described in detail by Pratt et al. (2005). The libraries used are as identified in the columns of the heat map (Figure 4.1), where Rhizome consists of the RHIZ1 and RHIZ2 libraries together, Ovary is OV1 and OV2 combined. Drought is the combination of the Absicic Acid (ABA1), drought stress after flowering (DSAF1), drought stress before flowering (DSBF1), salt stress (SS1) and water stress (WS1). Biotic stress, the focus of our study is the combination of Pathogen Incompatible (PI1), Pathogen Induced compatible (PIC1), salicylic acid (SA1), and WOUND1. All data were fully normalized to compensate for different sequencing depths for each library or group of libraries, and to account for different strengths of expression. Hence, all values are from 0 (not found at all in the library) to 1 (all ESTs from that gene in just the one library). Of the 432 genes with 30 ESTs or more, many were not expressed at all (green color in heatmap) in the Biotic Stress sub-group and a few were expressed preferentially in this group (red). Those that are white in the heat map (Figure 4.1) are the ones expressed at the average value for all libraries combined. The latter are the ones of interest to us in this study as negative controls, *i.e.* house-keeping genes without any evidence of preferential expression. Therefore, to establish our set of control loci, we selected only those UniScripts that were expressed in Biotic Stress at a level between 0.04 and 0.06, where 0.05 would be expression at the same level as the overall average for all libraries together. We then selected further those that have a relatively low R statistic, or believability score (see Pratt et al. (2005) for details). The final list of 30 UniScripts includes those genes that are not expressed differentially under biotic stress and that are likely to be constitutive in their expression (Figure 4.2).

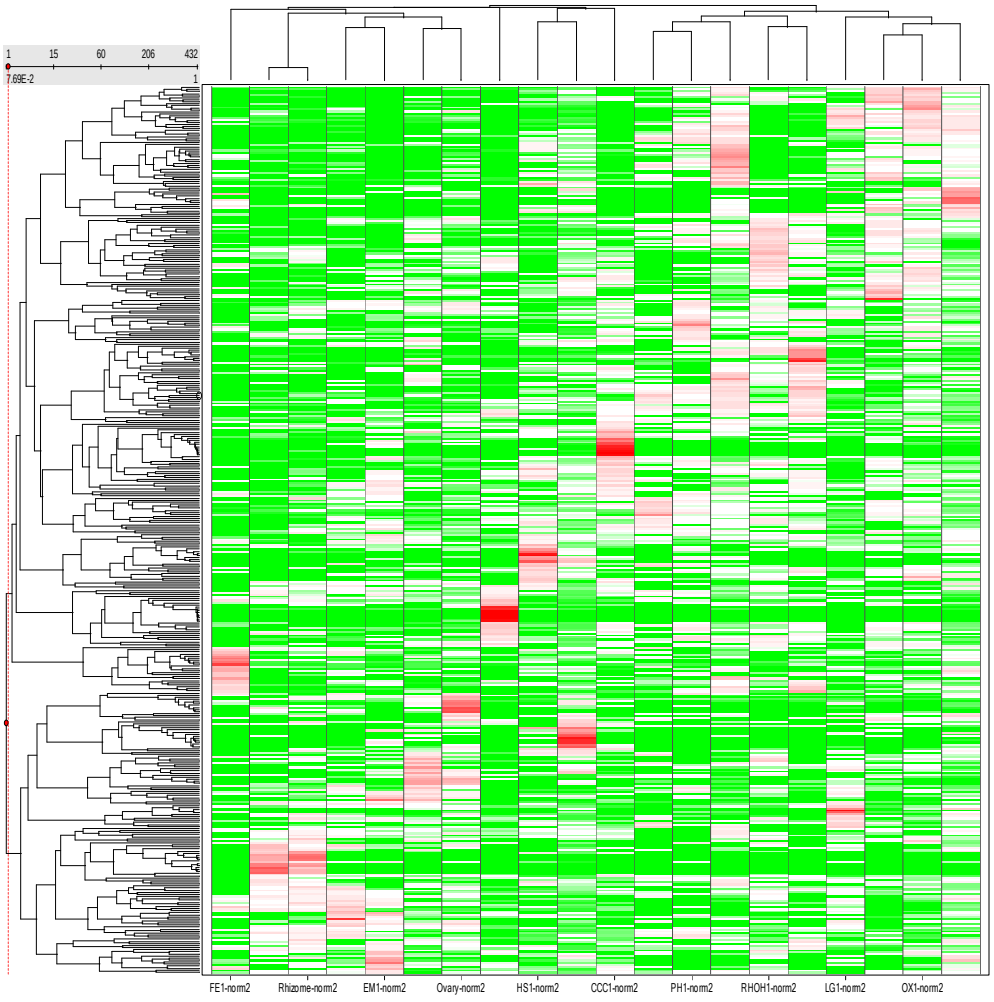


Figure 4.1. Heat map of sorghum 432 UniScripts expressed as ESTs 30 or more times. Uniscripts are clustered by identity (y-axis) and EST libraries are clustered by gene expression similarity. Rhizome consists of RHIZ1 and RHIZ2 libraries. Ovary is OV1 and OV2 combined. Drought is ABA1, DSAF1, DSBF1, SS1 and WS1. Biotic stress is PI1, PIC1, SA1 (salicylic acid), and WOUND1.

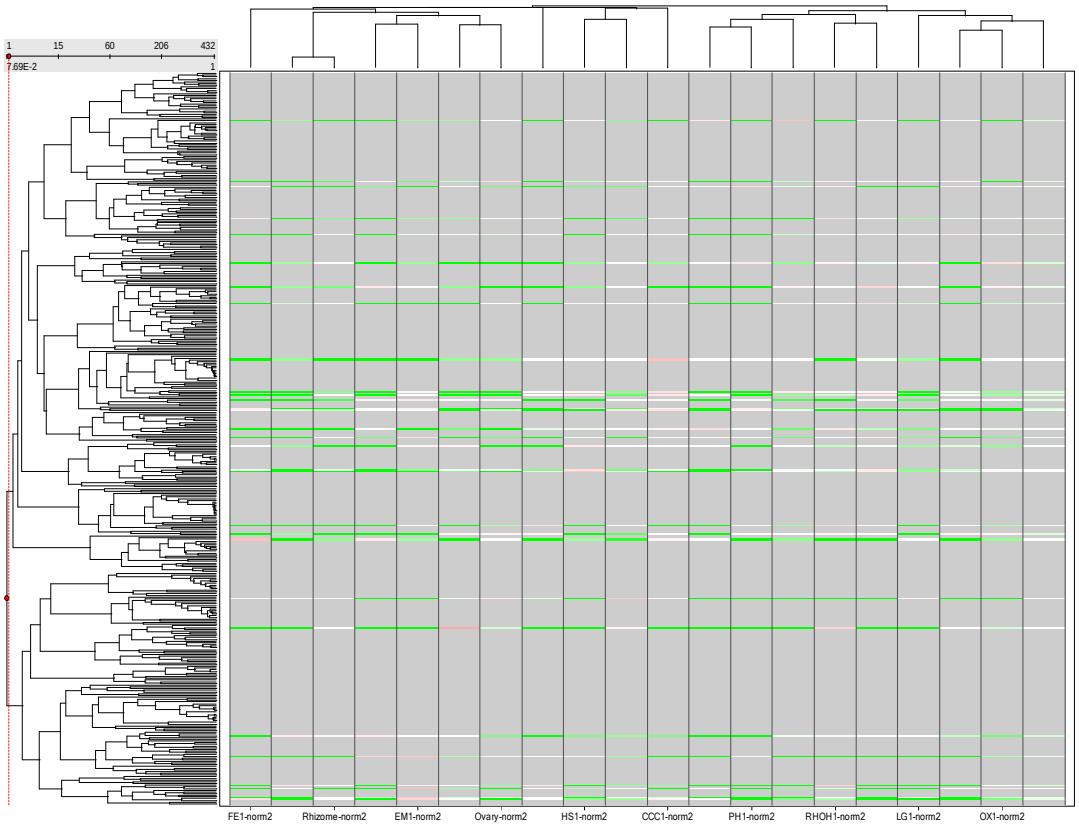


Figure 4.2. UniScripts that were expressed in Biotic Stress between 0.04 and 0.06. UniScripts that were expressed in Biotic Stress between 0.04 and 0.06 (0.05 would be expression at the same level as the overall average for all libraries together). We selected further those that have a relatively low R statistic, or believability score. The final list of 30 UniScripts are thus those genes that are not expressed differentially under biotic stress and that are likely to be constitutive in their expression.

Ideally, one would want a gene that was expressed at the same normalized value across all libraries - such a row would be a white line all the way across in the heat map. A file containing all the relevant data for these 30 UniScripts is included as supplementary material. This Excel file is ordered from top to bottom relative to the order of rows showing in Figure 4.2. Observed expression for every library or group of libraries is given in this Excel file, together with intermediate data used to get fully normalized values (-norm2 values). The Excel file also includes annotation information from UniprotSwissprot. Two of these control genes did not have a map position and other data in Phytozome and were not used.

We identified a set of 48 disease response gene candidates using the same libraries as described above, but having more than 70% of their EST coming from the Biotic Stress sub-group of libraries, at least 6 ESTs, and a divergence value in BLASTP comparisons to rice smaller than $E=10^{-60}$. Several of these candidates have other evidence of being involved in disease resistance responses (Zamora et al. 2009)(Chapters 2 and 3).

In order to increase the sample size, we additionally identified a set of 104 HED genes, including the genes described above, by relaxing the constraints to 4 ESTs per gene and 60% of the ESTs coming from the Biotic Stress libraries, and with a divergence in BLASTP smaller than $E=10^{-40}$. We also used 130 of the constitutively expressed control genes used in Zamora et al. (2009). Although this approach was done in order to increase statistical power and reduce the effect of a small sample in the results, it has the caveat that by relaxing the constraints for the pattern of expression it may lead to including false positives and false negatives in the test and control sets, respectively, *i.e.* non-disease response genes in the set of disease response genes and viceversa. It is known that certain disease resistance genes have a low level of constitutive expression and therefore it is possible that genes actually

involved in disease resistance but that were not upregulated during the induction tests used to generate the ESTs may be present in the control set.

We used all these uniscripts, both test and control sets, to BLAST agains the *S. bicolor* genome sequence in Phytozome and to obtain the full length sequence of the open reading frame (ORF) and the peptide sequence. We additionally obtained the chromosome map position of each gene, the number of exons, introns, total genomic span and total size of the coding sequence. We also obtained the number of copies in a cluster and the genome from the BLASTn values with a final cutoff of $E=10^{-80}$ and from the map positions of those hits.

In order to get the distance to the nearest telomeric region for each gene, we selected the smallest value of the gene position itself and that of the difference of the total chromosome size and the gene position, and we called this the shortest distance to the telomere.

We used a two sample t-test, assuming heterocedasticity, to compare the means and variances of these variables for the two groups of genes. For these tests we used the data for a single gene from those genes belonging to gene families. We mapped these genes in the genome of sorghum using GenomePixelizer (release October 2003), using the total chromosome sizes reported by Phytozome.

Results

We obtained the physical map position, number of paralogs in a cluster, paralogs in the genome, size of the gene in base pairs and number of exons for 104 and 130 genes from the control and disease response gene candidate sets, respectively and calculated their descriptive statistics (Table 4.1).

Table 4.1. Distance from the telomere, copies per cluster, gene size and number of exons measured for highly expressed and divergent disease response gene candidates.

	Distance to Telomere[a]	Copies per cluster[b]	Copies per Genome[c]	Genomic Span[d]	CDS size	CDS/ genomic span	Exons[e]	Introns
HED-DRG								
(n= 104)								
Mean	9054997.067	1.41	7.54	5260.29	1814.44	0.42	8.83	7.97
StDev	6524396.846	0.98	36.39	3913.28	1022.25	0.20	7.85	7.83
CONTROLS								
(n= 130)								
Mean	10709663.6	1.12	1.38	3758.70	1218.60	0.33	6.23	5.39
StDev	12569205.14	0.53	1.14	1804.15	544.19	0.16	4.55	4.57
p value	0.09	0.005**	0.004**	0.0002***	$1.7x10^{-7}$***	0.01*	0.001**	0.001**

a. Distance from the telomere to the gene.

b. Number of copies of paralogs per cluster.

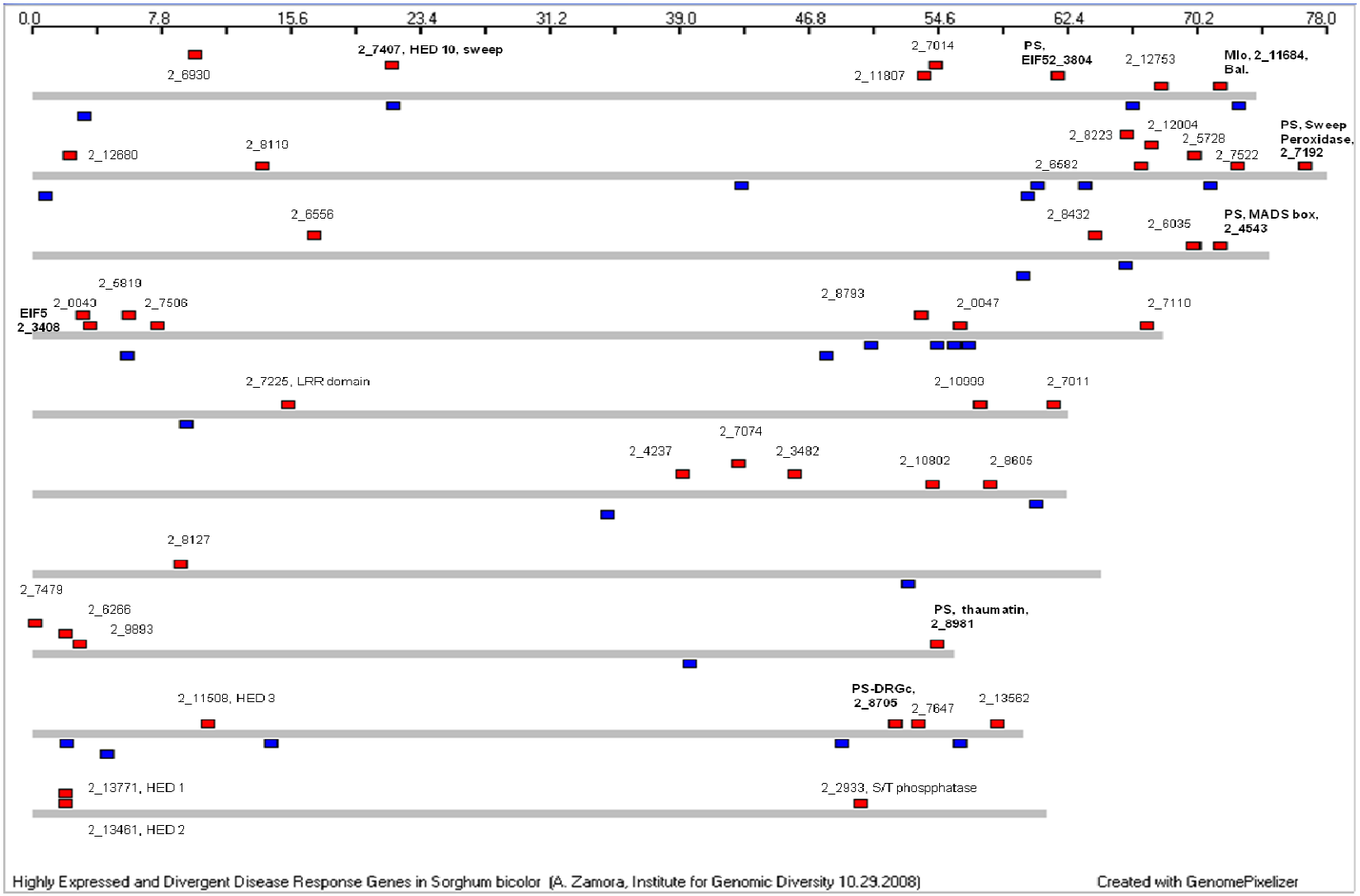
c. Copies of paralogs in the whole genome.

d. Total length of the gene including coding sequence, introns and UTRs (5' and 3').

e. Total number of exons in the gene.

Interestingly, the set of 104 highly expressed and divergent disease response gene candidates is, on average, over 1.5Mb closer to telomeres (9,053.218 bp) than the control set of genes (n=130; 10,709.663 bp), although this difference is only significant at an alpha of 0.10 (t-Test, p = 0.09, df = 202). However, the initial set of HED genes (n=48) is significantly closer to the telomeres than the control set (n=28; t-Test, p = 0.025, df=45; Figure 4.3). The distribution of all the mapped genes is aggregated and towards the telomeres instead of randomly or uniformly distributed (Figure 4.3). Noteworthy, several genes organized in clusters have the most distal positions of the genes mapped in several chromosomes (Figure 4.3), such as 2_7192 (Chr. 2; peroxidase, 4 copies), 2_12680 (Chr. 2, 3 copies); 2_7011 (Chr. 5; 5 copies), 2_8843 (Chr. 4, 3 copies), and 2_8981 (Chr. 8; thaumatin, 7 copies). Additionally, two genes, 2_13771 and 2_13461, located at 1,8Mb of the telomere in chromosome 10 (Figure 4.3), are similar to each other and appear to be a duplicated gene pair that has suffered some differentiation recently, possibly due to sub-functionalization. These two genes are the best disease response candidates based on their total level of expression, divergence to rice, and proportion of ESTs coming from the Biotic Stress sub-group of libraries, a measure of their validity and upregulation under pathogen attack. Another gene with two similar copies, but located in different regions of the genome is 2_3804 (a and b), located in Chr. 1 at 61.6Mb and in Chr. 4 at 3.3Mb. These genes (2_3804) are eukaryotic translation initiation factor-5, and were shown to evolve adaptively (Chapter 2), the same as 2_7192, 2_8981, and 2_11684, all of which are close to the telomeres in sorghum and have multiple paralogs (Figure 4.3).

Figure 4.3. Mapping of highly expressed and divergent *Sorghum bicolor* genes (red) and control set (blue). This figure shows only the 48 HED and 28 control genes from the first experiment. HED 1: number one candidate according to total expression, proportion of ESTs coming from the biotic stress subgroup of libraries and divergence with respect to rice. Bal.: evidence of balancing selection in *Sorghum bicolor;* Sweep: evidence of selective sweep in *S. bicolor* (Chapter 3). PS: positively selected, dN/dS rate ratio significantly >1, according to codon based maximum likelihood tests (Chapter 2)


0.0
7.8
15.6
23.4
31.2
39.0
46.8
54.6
62.4
70.2
78.0
2_7407, HED 10, sweep
2_6930
2_7014
2_11807
PS, EIF52_3804
2_12753
Mlo, 2_11684, Bal.
2_12680
2_8119
2_12004
2_8223
2_5728
2_7522
PS, Sweep Peroxidase, 2_7192
2_6582
2_6556
2_8432
2_6035
PS, MADS box, 2_4543
2_5819
2_0040
2_7506
EIF5 2_3408
2_8793
2_0047
2_7110
2_7225, LRR domain
2_10000
2_7011
2_7074
2_4237
2_3482
2_10802
2_8605
2_8127
2_7479
2_6266
2_8893
PS, thaumatin, 2_8981
2_11508, HED 3
PS-DRGc, 2_8705
2_7647
2_13562
2_13771, HED 1
2_2933, S/T phospphatase
2_13461, HED 2
Highly Expressed and Divergent Disease Response Genes in Sorghum bicolor (A. Zamora, Institute for Genomic Diversity 10.29.2008)
Created with GenomePixelizer

Higly expressed and divergent (HED) Disease response gene (DRG) candidates have a mean of 1.41 copies (SD=0.98) arranged in local clusters and have a significantly higher number of copies than control genes, both locally (t-Test, $p<0.005$, df= 149) as well as in the whole genome (t-Test, $p<0.04$, df= 102, Table 4.1). These candidate genes also have a significantly larger overall genomic span than the controls ($\bar{x}$ = 5260 ± SD 3913bp vs. 3758, SD=1804bp; t-Test, p=0.0002, df=136) as well as coding sequence size ($\bar{x}$ = 1814, SD=1022bp vs. 1218, SD=544; t-Test, $p=1.7x10^{-7}$, df=147).

Remarkably, the larger coding sequence of the DRG candidates is due to a significantly higher number of exons (t-Test, p=0.001, df = 154), where the disease response gene candidates have a mean of 8.83 exons/gene while genes from the control set have 6.23 exons/gene (Table 4.1, Figure 4.4).

Relationship between position in the genome and gene structure

Since DRGs are more distally located in chromosomes than controls, and are also made of more exons, we hypothesized that there should be a correlation between distance from the centromere and the number of exons. However, we found no significant correlation between these variables when using the large dataset (Figure 4.5). However, using the smaller but more stringent dataset, there is a slight suggestion that multicopy genes with few exons may be distally located, while multiexon genes with few copies may be preferentially located in a region between 7.5 and 10 Mbp (Figure 4.6). However, the polynomial regression used to fit the data explains only 6% of the variation and it is not clear whether this is an artifact due to a small sample size.

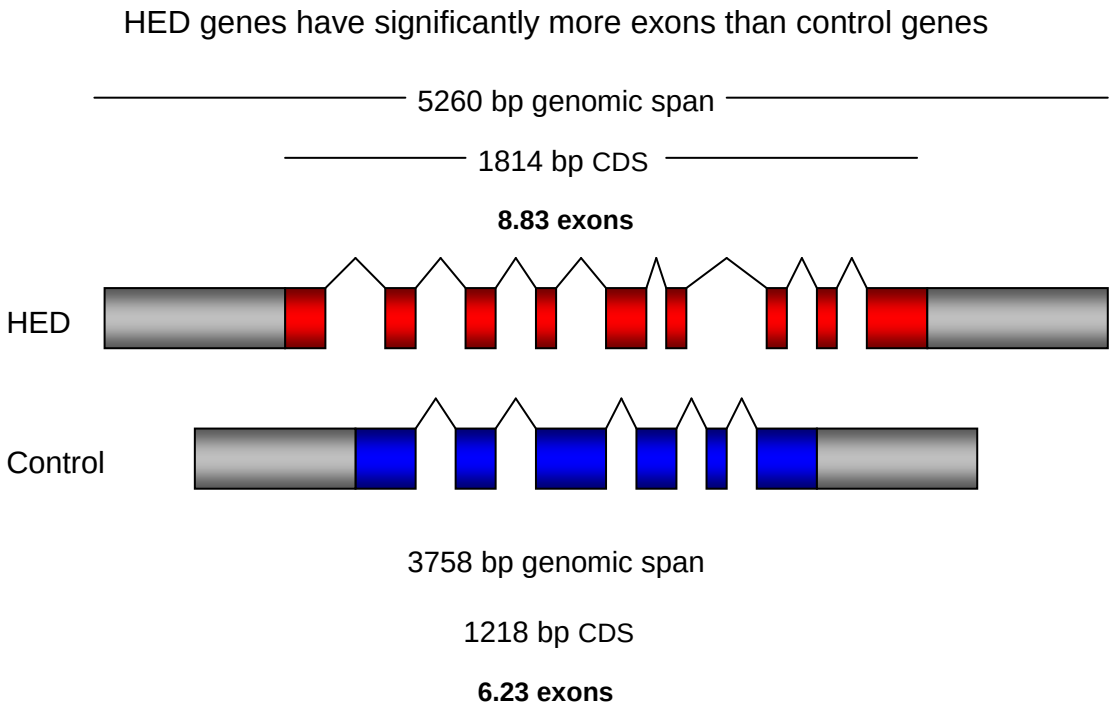


Figure 4.4. Average structural characteristics of the highly expressed and divergent set of disease response orthologs. Highly expressed and divergent genes (HED) have 8.83 exons per gene (red), while genes in the control set of constitutively expressed, house-keeping genes have a mean of 6.23 exons per gene (blue).

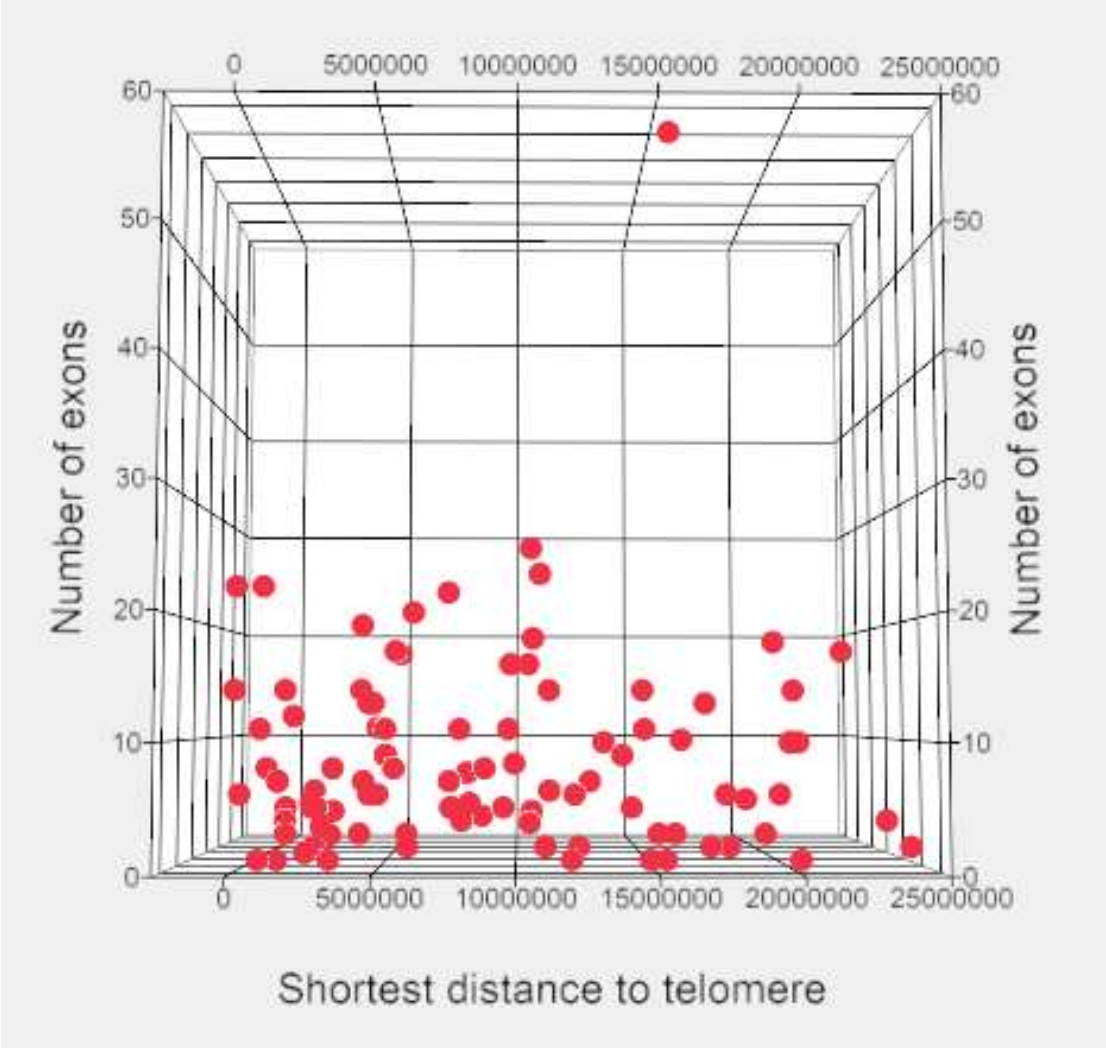


Figure 4.5. Distribution of HED-DRG candidates as a function of increasing distance from the telomere. Linear and polynomial regressions do not explain any of the variation observed for this data.

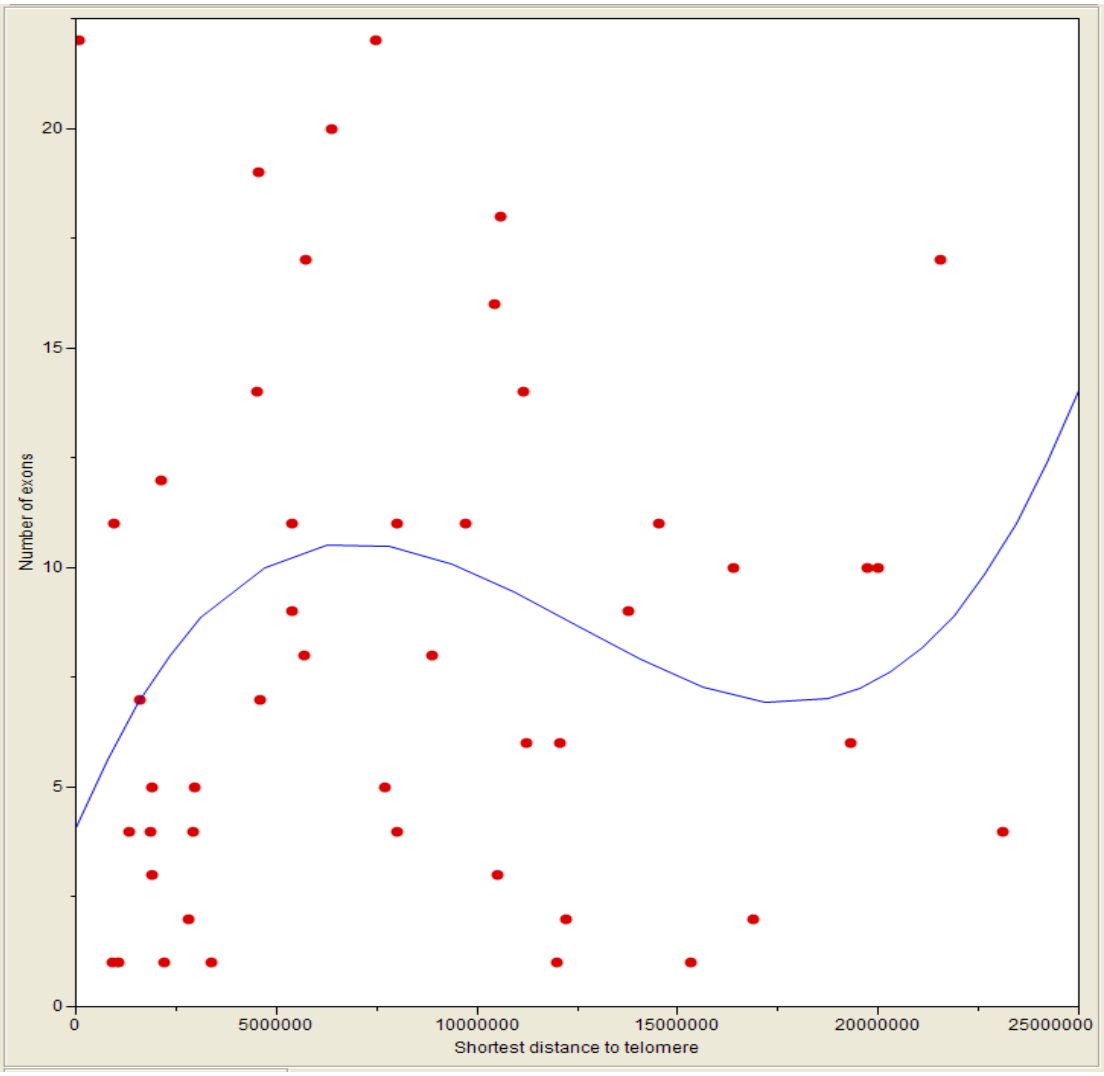


Figure 4.6. Distribution of HED-DRG candidates as a function of increasing distance from the telomere for the best 48 HED genes . Line: Polynomial regression degree 3, R square = 0.06.

We detected an inverse relation between number of paralogs in a cluster and the distance from the telomeres for HED genes (Figure 4.7). To test whether this inverse relation is significant we fit a linear and a quadratic regression of the number of copies on the number of exons per gene (Fig. 4.8), both of which indicate that 3 and 4% of the variation is explained by these models, respectively. Several points at the level of 1 for number of copies contain multiple genes, reducing the effect of genes with multiple copies and few exons. To further test this relationship, we divided the number of exons in two categories, <9 and ≥ 9 exons (since the average for HED genes is 8.83), and also divided the number of paralogs in two categories, ≤ 2 and >2 copies per cluster and used this data in a Fisher's Exact Test. This analysis also suggests that there is a significant inverse relationship between the number of exons and paralog copies per gene ($p= 0.04$). Similarly, it appears that genes with multiple paralogs tend to be closer to the telomere in the large HED dataset, although the large number of genes with a single copy obscure this relationship (Figure 4.9). Finally, a multivariate discriminant analysis (Figure 4.10) graphically shows that HED genes have more exons, and number of copies in the cluster than the set of control genes. An exception in the control set is a beta-expansin, which has 6 paralogs, although it is located ca. 24Mbp from the telomere.

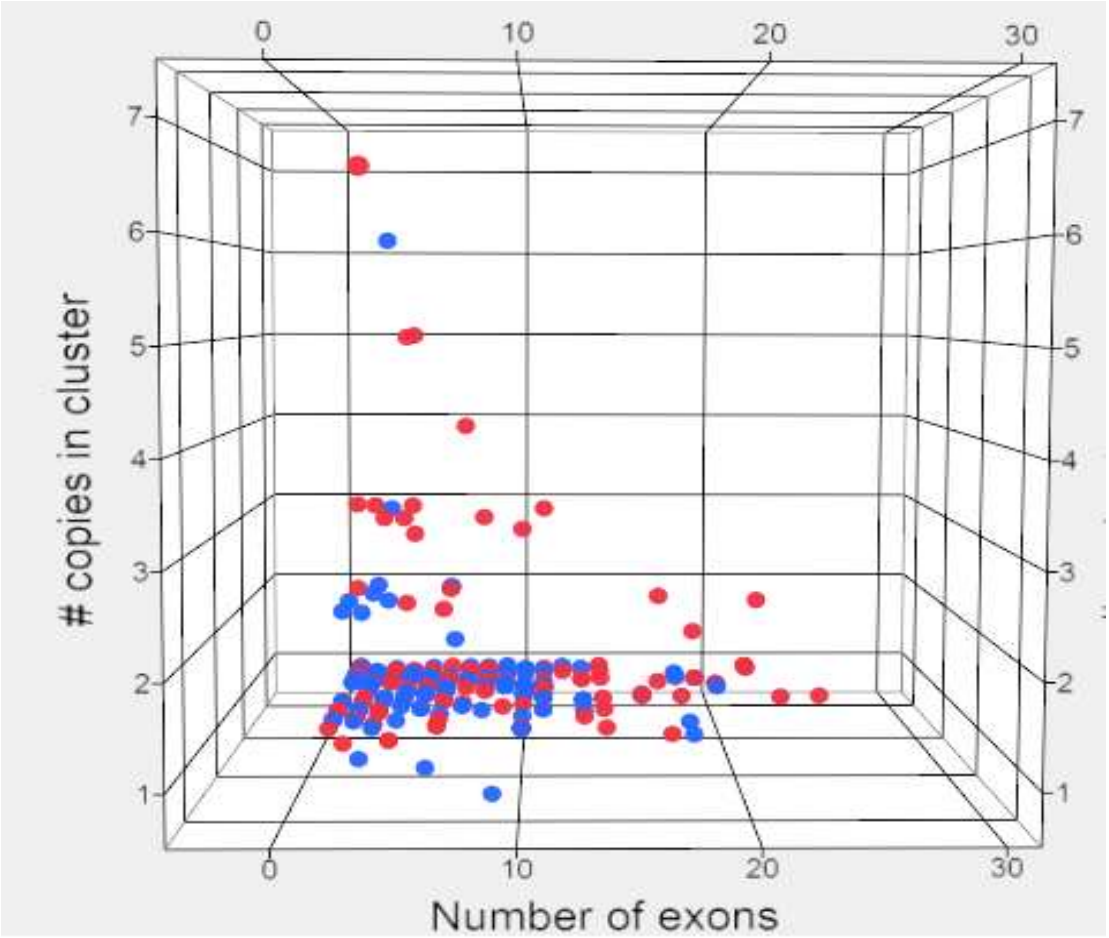


Figure 4.7. Inverse relation between the number of exons and the number of paralog copies in a cluster for highly expressed and divergent (HED) disease response gene candidates (DRGs) in the sequenced genome of *Sorghum bicolor*. HED genes (red) have more paralogs in clusters and more exons per gene (outlier with 58 exons, similar to a rapamycin target, not shown) than the set of evenly expressed control genes (blue). Z axis is the distance in Mbp from the telomere, blue dots in the front are further from the telomere (ca. 30Mbp).

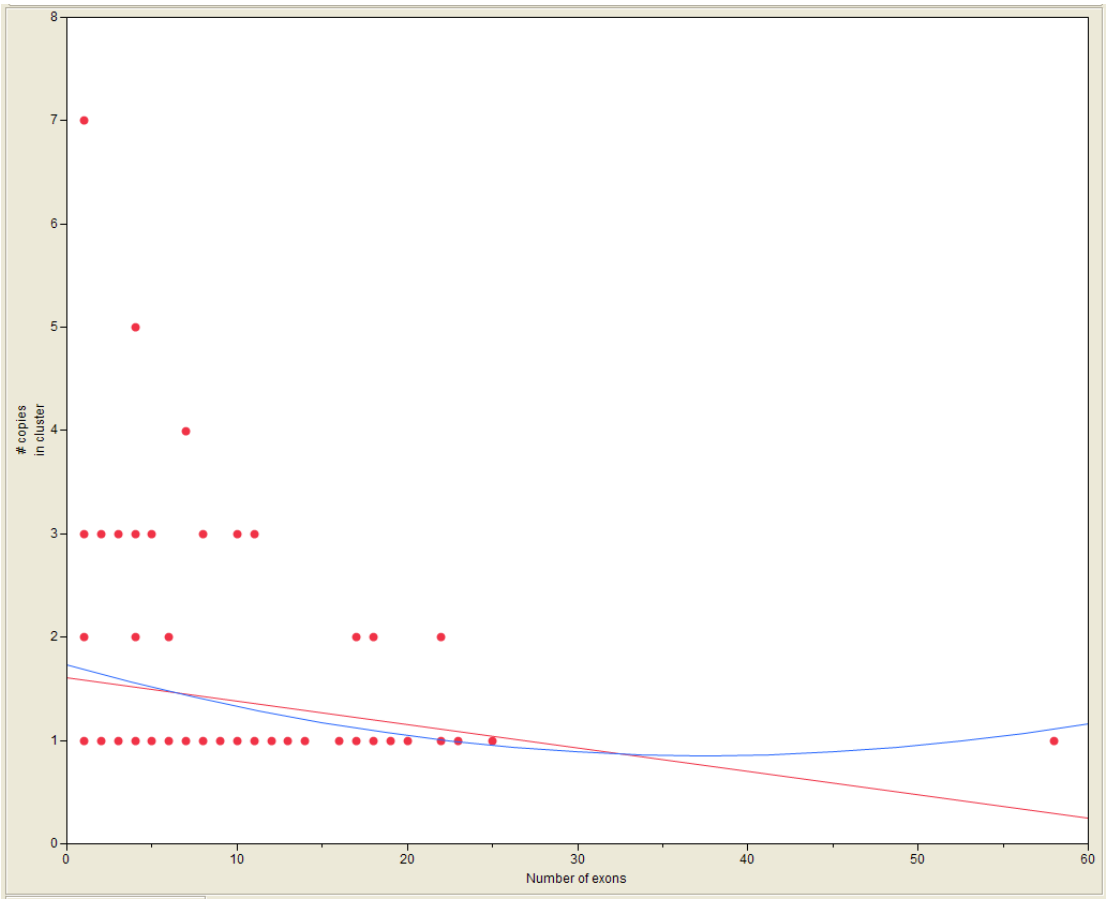


Figure 4.8. Linear (red) and quadratic (blue) regressions of the number of paralogs per cluster and the number of exons per gene for highly expressed and divergent (HED) disease response gene candidates (DRGs) in the sequenced genome of *Sorghum bicolor*. Linear and quadratic regressions explain 3 and 4% of the variation.

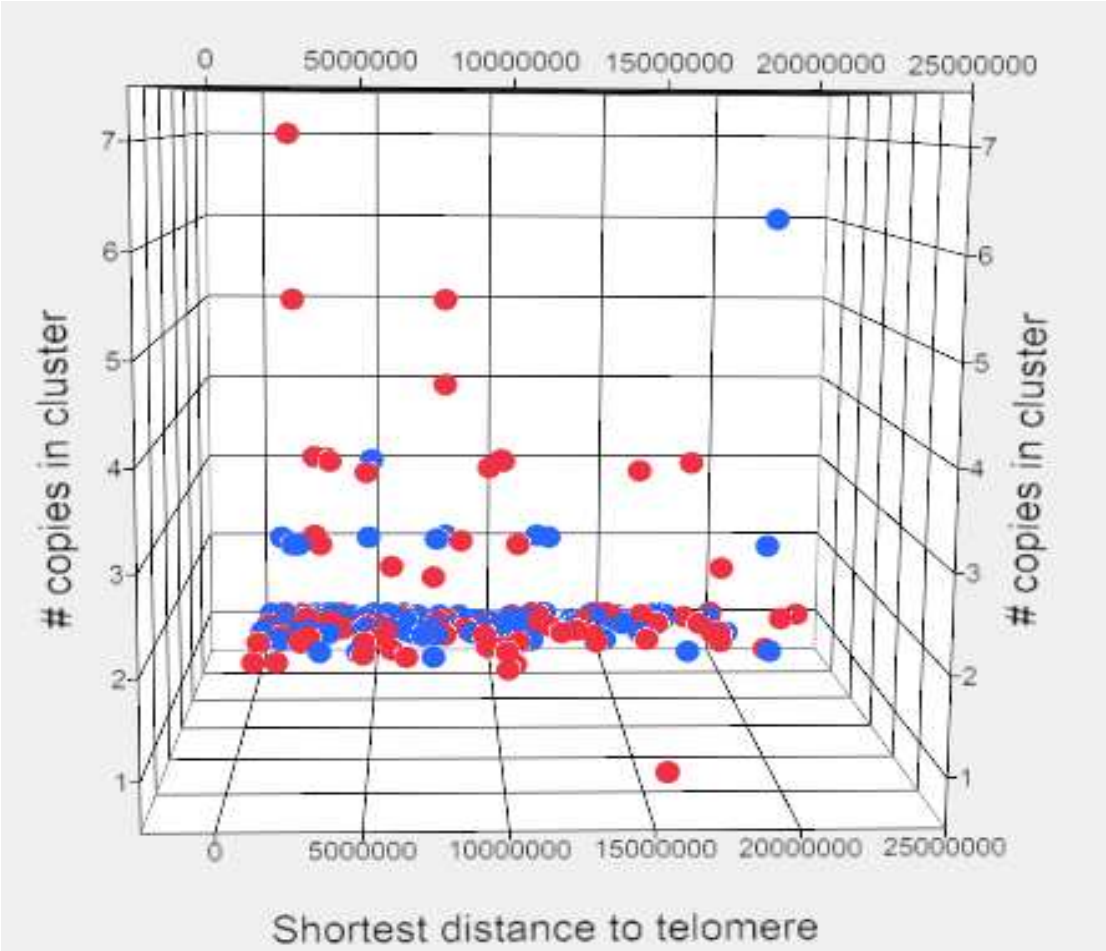


Figure 4.9. Inverse relation between the number of paralog copies in a cluster and the distance to the telomere for highly expressed and divergent (HED) disease response gene candidates (DRGs) in *Sorghum bicolor*. HED genes (red) have more paralogs in clusters than the set of evenly expressed control genes (blue). Z axis is the number of exons, a HED putative rapamycin target is an outlier (in the front) with 58 exons. A putative beta-expansin precursor is an outlier in the controls with 6 paralogs in a cluster, but with a position closer to the centromere.

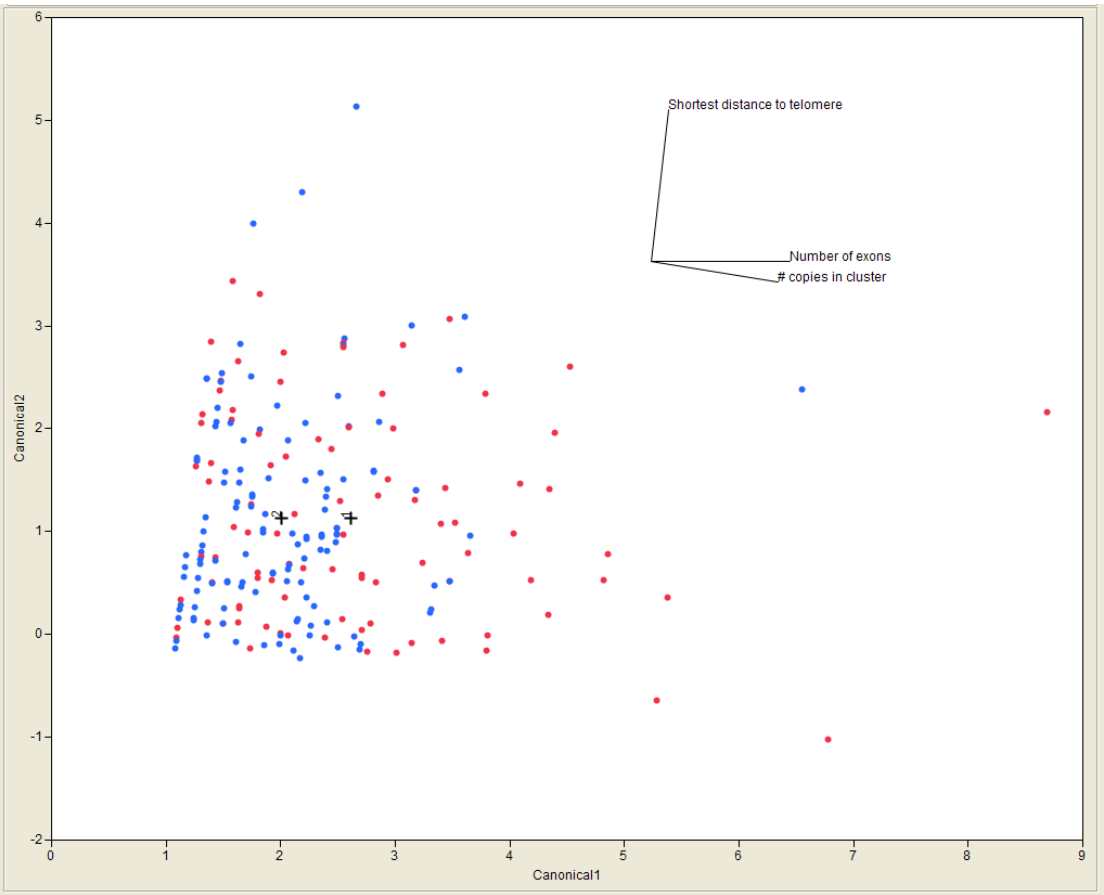


Figure 4.10. Canonical plot for multivariate discriminant analysis for HED (red) and constitutively expressed control (blue) sorghum genes. The biplot rays show the direction influenced by the three variables in the two-dimensional plot. HED genes have a greater influence than control genes in the direction of increasing Number of exons and number of copies in the cluster. Evenly expressed control genes have a greater influence in the direction of increasing distance from the telomere.

Discussion

Previous work on the evolution of disease response genes has focused mainly on the effect of point mutations on their structure and function, and there is now abundant evidence pointing out to the fact that, as a class, disease response genes show higher dN/dS rate ratios than, for example, house-keeping genes and several show signals of positive selection. This difference is both historical, through the comparative analysis of orthologs from different species, as well as current, inferred from the analysis of polymporphism in extant populations. That many plant disease response genes appear to have at least a few sites that seem to have been the target of positive selection is not surprising, given their proven or likely interaction with pathogen molecules (Bishop, Dean, and Mitchell-Olds 2000; Bittner-Eddy et al. 2000; Stahl and Bishop 2000; Bishop et al. 2005; Moeller and Tiffin 2005; Zamora et al. 2009). A similar scenario has been shown to occur in animals, where a significant portion of all genes shown to have signals of positive selection are involved in disease resistance mechanisms (Hughes, Ota, and Nei 1990; Borghans, Beltman, and De Boer 2004; Bryja et al. 2006), and of course in pathogens, where evidence of positive selection is clear and strong (Yang et al. 2000; Nielsen and Yang 2003).

It is evident that both parts implicated in this antagonistic coevolution are selective agents for each other, and that the evidence from this ancient struggle can be seen in the comparatively rapid evolution of the molecules that interact and are used as defense or attack mechanisms. This excludes, however, genes involved in the Salicylic Acid pathway that appear to be highly conserved (Bakker et al. 2008) and may be “guarded” by rapidly evolving R-genes (Van der Hoorn, De Wit, and Joosten 2002) or other components of the basal mechanism of disease resistance.

Other important component of disease response gene evolution, first described for R genes, but apparently valid for other kinds of genes arranged in clusters in the genomes of plants, is that of the effect of recombination and unequal-cross overs on the generation of new alleles, new paralogous copies, the homogenization through gene conversion of some of those copies and also the long term maintenance of some of those copies (Michelmore and Meyers 1998). However, this is the first time to our knowledge that disease response genes have been shown to have significantly more exons and paralogs per cluster, than a set of control genes not likely to be involved in disease resistance mechanisms. This suggests that not all gene categories are under the same selective pressures or have the same constraints, as there could also be a dosage effect acting to maintain some genes as single copy (Paterson et al. 2006). It is also interesting to see that HED genes tend to be closer to the telomeres than the controls. Although this difference was significant only in the more stringent dataset and only a trend in the larger dataset, it is consistent with that described by Boyko et al. (2002), who had previously shown that a set of 160 genes, including R genes and general disease response genes, mapped to distal telomeric regions in *Aegilops tauschii*. Additionally, Boyko et al. (2002) showed that these genes were arranged in clusters, and that there was a positive correlation between the physical density of genes in a region and the rate of recombination, measured as the ratio of the genetic length of a chromosome segment to the physical length of the same segment.

The significantly higher number of exons (and consequently introns) present in disease response genes, together with their arrangement in clusters, the higher recombination rate reported in these regions, the variation in copy number for these clusters, both within and between species, are all evidence of a strong selective pressure to generate variation in these genes. Additionally the larger number of exons may also be evidence of a set of genes assembled more recently from different

domains in evolutionary time, to accommodate the host defense needs. We found evidence of this in the 2_7407 gene (Chr. 1, at 21.5Mb, Chapter 3), which contains a highly conserved RNA Binding Domain (RBD), which matches only RanGAP1 in *Arabidopsis thaliana* but, which is not the sorghum RanGAP1 because the conserved WPP domain and the LRR, characteristic of RanGAP proteins, are missing from its N-terminus, therefore, this cereal protein doesn't have a full length ortholog in *A. thaliana*. Furthermore, the orthologs from cereal species align relatively well throughout the 665 amino acid residues of the sorghum protein, however, there are two regions that show extensive rearrangement that are very difficult to align and that show the evidence of recurrent (intraspecific) recombination.

It is therefore possible that these genes are younger and have not had enough time to become optimized for high expression by fusing exons and reducing the number of introns. An alternative explanation is that maybe there is selection against reduction in the number of exons and introns, since having more exons increases the modularity of these genes, fostering the generation of new combinations, allowing for the exonization of introns and permitting high rates of inter-allelic and inter-genic recombination that generate viable variation, without disrupting the open reading frame of functional and already tested exons.

One of the most surprising features discovered in this research was the increase in the number of paralogs per cluster as a function of greater distance from the centromeres. The biological meaning of this observation is not clear but two other features of the data studied here are related to it. First, several gene families arranged in clusters are located towards the telomeric regions, such as thaumatins and peroxidases, closer to the ends of the sequenced genome. These genes appear to evolve through the birth and death process (Michelmore and Meyers 1998) whereby

some genes (specifically those in the edges of the cluster, per our analyses) are rather conserved and maintained for long periods of time, showing a one to one orthology between sorghum and rice (Chapter 2), while other genes in the center of the cluster are more similar to each other, vary in copy number and may include pseudogenes, all of which points to the recurrent action of unequal crossing over, that homogenizes these central paralogs in the cluster while at the same time generating diversity. Interestingly, these genes, are single exon genes, probably because they need to be produced extremely rapidly to be used as weapons against the invading pathogen, as happens with several other pathogenesis related proteins. Therefore their arrangement in clusters, single exon structure and distal position in the genome may have been positively selected to increase diversity and speed of expression. On the contrary, in rice, the Pi20(t) R gene confers resistance to a broad spectrum of *Magnaporthe oryzae* isolates and it is located near the centromere (11.9Mb) in chromosome 12 (Li et al. 2008). This gene appears to be conserved, which is consistent with its position in the rice genome and with the fact that it confers broad spectrum resistance to rice blast, probably because it recognizes a conserved epitope in an essential molecule from *M. oryzae.*

Sample size vs. precision: the case of differentially expressed genes

In this study the two sample sets (small and stringent vs. larger and slightly relaxed in constraints) were considered different since increasing the sample size has the unavoidable consequence of increasing type I and type II errors in the analysis. All the comparisons made were significant using both data sets except two: the relationship between number of exons and distance from the telomere, which disappeared completely with the bigger sample; and the difference in position along the chromosome between HED and control genes, which is only a trend in the larger

sample. It remains to be seen whether the subtle increase in the number of exons around 7.5 to 10 Mbp from the telomere is an artifact caused by a small sample size, or a real phenomenon detected only when using a gene set with the precise characteristics as those used. To test this, one would require conducting many more differential expression analyses using different pathogens in order to elicit the expression of all the genes involved in disease response, thereby increasing the sample size without relaxing the candidate gene selection conditions and losing statistical power.

Nevertheless, we found significant differences between HED and control genes for several variables using the large dataset, which seem to indicate that selection has acted in these genes to increase the number of paralog copies per gene and the number of exons per gene. This suggests that there is a greater need for a structural flexibility that will allow faster and more efficient positive and negative selection to optimize the function of DRGs, and that rapid evolution in these genes may be a direct result of the structure and arrangement in clusters of these genes, in turn the result of intense pathogen selective pressure.

These genes may require an increased recombination rate to shuffle the variation in the different exons, to get rid of deleterious mutations, or domains that have become the target of pathogen molecules, without losing the ability to retain the functional or even improved alleles at other exons, due to Hill-Roberson interference. Additionally, an increased number of exons may also greater capability to generate functional variability through alternative splicing and through exonization of introns.

Finally, several non-disease response genes may face selective pressures as large as that imposed by pathogens and therefore this obscures the difference, but it also means that regardless of their function, genes under high selective pressure to change rapidly

will be closer to telomeres where higher recombination rates allows this to happen more efficiently.

In summary, we have provided evidence that the structure of highly expressed and divergent disease response gene candidates is significantly different from a control set of house-keeping genes, and this difference may have been the result of positive selection to increase diversity, speed of reaction and improved elimination of deleterious mutations.

Acknowledgments

We thank Lee Pratt, Charles Aquadro, Walter De Jong and Steve Kresovich for valuable comments on the manuscript. Steve Tanksley, Martha Hamblin and Sharon Mitchell also provided comments that improved the quality of this paper significantly. We also want to give special thanks to Peter Moffet, Julio Vega and Marianne Jaubert for their comments. AZM is supported by the Costa Rica-USA Foundation for Cooperation (CRUSA) through a special Fulbright scholarship in biotechnology and by the Institute of Genomic Diversity (Cornell University, Ithaca, NY). This research was funded by the National Science Foundation award number 0115903 to SK.

REFERENCES

Bakker, E. G., C. Toomajian, M. Kreitman, and J. Bergelson. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. Plant Cell **18**:1803-1818.

Bakker, E. G., M. B. Traw, C. Toomajian, M. Kreitman, and J. Bergelson. 2008. Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana*. Genetics **178**:2031-2043.

Bishop, J. G., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A **97**:5322-5327.

Bishop, J. G., D. R. Ripoll, S. Bashir, C. M. B. Damasceno, J. D. Seeds, and J. K. C. Rose. 2005. Selection on glycine beta-1,3-endoglucanase genes differentially inhibited by a *Phytophthora* glucanase inhibitor protein. Genetics **169**:1009-1019.

Bittner-Eddy, P. D., I. R. Crute, E. B. Holub, and J. L. Beynon. 2000. RPP13 is a simple locus in Arabidopsis thaliana for alleles that specify downy mildew resistance to different avirulence determinants in Peronospora parasitica. Plant Journal **21**:177-188.

Borghans, J. A. M., J. B. Beltman, and R. J. De Boer. 2004. MHC polymorphism under host-pathogen coevolution. Immunogenetics **55**.

Boyko, E., R. Kalendar, V. Korzun, J. Fellers, A. Korol, A. H. Schulman, and B. S. Gill. 2002. A high-density cytogenetic map of the *Aegilops tauschii* genome incorporating retrotransposons and defense-related genes: Insights into cereal chromosome structure and function. Plant Molecular Biology **48**:767-790.

Bryja, J., M. Galan, N. Charbonnel, and J. Cosson. 2006. Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism

of the DQA MHC class II gene in voles (Arvicolinae). Immunogenetics **58**:191-202.

Damasceno, C. M. B., J. G. Bishop, D. R. Ripoll, J. Win, S. Kamoun, and J. K. C. Rose. 2008. Structure of the glucanase inhibitor protein (GIP) family from *Phytophthora* species suggests coevolution with plant endo-beta-1,3-glucanases. Molecular Plant-Microbe Interactions **21**:820-830.

Gout, L., I. Fudal, M. L. Kuhn, F. Blaise, M. Eckert, L. Cattolico, M. H. Balesdent, and T. Rouxel. 2006. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Molecular Microbiology **60**:67-80.

Hughes, A. L., T. Ota, and M. Nei. 1990. Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Molecular Biology & Evolution **7**:515-524.

Kuang, H., K. S. Caldwell, B. C. Meyers, and R. W. Michelmore. 2008. Frequent sequence exchanges between homologs of RPP8 in *Arabidopsis* are not necessarily associated with genomic proximity. Plant Journal **54**:69-80.

Li, W., C. L. Lei, Z. J. Cheng et al. 2008. Identification of SSR markers for a broad-spectrum blast resistance gene Pi20(t) for marker-assisted breeding. Molecular Breeding **22**:141-149.

Mauricio, R., E. A. Stahl, T. Korves, D. Tian, M. Kreitman, and J. Bergelson. 2003. Natural selection for polymorphism in the disease resistance gene Rps2 of *Arabidopsis thaliana*.

Michelmore, R. W., and B. C. Meyers. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research **8**:1113-1130.

Moeller, D. A., and P. Tiffin. 2005. Genetic diversity and the evolutionary history of plant immunity genes in two species of Zea. Molecular Biology and Evolution **22**:2480-2490.

Mondragón-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Research **12**:1305-1315.

Nielsen, R., C. Bustamante, A. G. Clark et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. Plos Biology **3**:976-985.

Nielsen, R., and Z. H. Yang. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Molecular Biology and Evolution **20**:1231-1239.

Paterson, A. H., B. A. Chapman, J. C. Kissinger, J. E. Bowers, F. A. Feltus, and J. C. Estill. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis, Oryza, Saccharomyces* and *Tetraodon*. Trends Genet **22**:597-602.

Podlaha, O., and J. Zhang. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc Natl Acad Sci U S A **100**:12241-12246.

Pratt, L. H., C. Liang, M. Shah et al. 2005. Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and skotomorphogenesis from a milestone set of 16,801 unique transcripts. Plant Physiol. **139**:869-884.

Rosebrock, T. R., L. Zeng, J. J. Brady, R. B. Abramovitch, F. Xiao, and G. B. Martin. 2007. A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. Nature **448**:370-374.

Stahl, E. A., and J. G. Bishop. 2000. Plant-pathogen arms races at the molecular level. Curr Opin Plant Biol **3**.

Van der Hoorn, R. A. L., P. De Wit, and M. Joosten. 2002. Balancing selection favors guarding resistance proteins. Trends in Plant Science **7**:67-71.

Wisser, R. J., Q. Sun, S. H. Hulbert, S. Kresovich, and R. J. Nelson. 2005. Identification and Characterization of Regions of the Rice Genome Associated with Broad-Spectrum, Quantitative Disease Resistance. Genetics **169**:2277-2293.

Yang, Z. H., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yeam, I., J. R. Cavatorta, D. R. Ripoll, B.-C. Kang, and M. M. Jahn. 2007. Functional Dissection of Naturally Occurring Amino Acid Substitutions in eIF4E That Confers Recessive Potyvirus Resistance in Plants. Plant Cell **19**:2913-2928.

Zamora, A., Q. Sun, M. T. Hamblin, C. F. Aquadro, and S. Kresovich. 2009. Positively selected disease response orthologous gene sets in the cereals identified using *Sorghum bicolor* L. Moench expression profiles and comparative genomics. Mol Biol Evol **26**:2015-2030.

CHAPTER FIVE

DIVERSITY GENERATING MECHANISMS IN THE GENOMES OF PLANTS AND THEIR IMPORTANCE IN PLANT DEFENSE

> "Most of the phenotypic diversity that we perceive in the natural world is directly attributable to the peculiar structure of the eukaryotic gene, which harbors numerous embellishments relative to the situation in prokaryotes. The most profound changes include introns that must be spliced out of precursor mRNAs, transcribed but untranslated leader and trailer sequences (untranslated regions), modular regulatory elements that drive patterns of gene expression, and expansive intergenic regions that harbor additional diffuse control mechanisms. Explaining the origins of these features is difficult because they each impose an intrinsic disadvantage by increasing the genic mutation rate to defective alleles."
>
> Michael Lynch. 2006. The origins of eukaryotic gene structure. Mol. Biol. Evol. 23(2):450–468.

Plants have the amazing capacity to take energy directly from the sun and use it to produce all the organic compounds they need to survive and reproduce. As self-reliant organisms they are at the bottom of all land food chains and are attacked by a multitude of other species that strive to steal energy from them. Organisms from virtually all taxa, even including other plants, have become parasitic and can assault a broad range or a single species of plants. Bacteria, fungi, oomycetes, nematodes, insects, mollusks, vertebrates and parasitic plants are just a few of the taxa that play the role of thiefs on the stage of life. How do plants deal with such a vast diversity of parasitic and pathogenic organisms? How do they cope with their large numbers and

adaptability? Are there features of the genes involved in defense and the genome itself that have been modified by selection in order to give plants a chance in the fight against pathogens? These are the questions I try to address here.

In this review, I have used examples from pathosystems including many of the taxa mentioned before, although I focused more in bacteria and fungi, due to their preponderance as pathogenic agents. However, it is possible that all the species which attempt, or have attempted in the past, to steal resources from a plant species may have had an effect on that species' physiological and morphological characteristics, its genetic diversity, and as we argue here, their genome content, order, structure and regulation.

The main premise behind this review is that the myriad of pathogen's constant and strong selective pressure has been such that in order to avoid extinction plants have had to compensate for their smaller effective population size by taking advantage of several different genomic diversity-generating mechanisms, thus increasing their chances of producing functionally useful variation on which natural selection can act. In essence, my hypothesis is that the observed eukaryotic genome complexity in size, content, order and regulation compensates for a smaller N_e and is the reason that plants (and we) can survive despite the attack of myriads of pathogens. Comparative genomics and molecular population genetics have been used to observe the effects of these mechanisms, and can be used as well to identify them individually and quantify their contributions.

It has been extensively shown that defense mechanisms require and maintain high diversity (e.g. MHC, antibodies in animals and NBS-LRR genes, thaumatins, chitinases in plants) and therefore it is also for these kinds of genes that we can see the

strongest and clearest examples of the mechanisms that lead to genome diversification and rapid evolution.

The neutral theory of gene and genome structure evolution

Lynch (2006) proposed that the evolution of the gene and genome structure in eukaryotes can't be explained by selection due to the small effective population size of most eukaryotes, particularly multicellular ones. He argued that their genome structure is therefore the outcome of a neutral evolutionary process occurring under genetic drift. Although many plant species have what may appear to be large populations, their effective population size is rather small.

The effective population size (Ne) is the size of an ideal population showing a similar decay in heterozygosity across generations as a result of random sampling of alleles. It can also be defined as the size of an ideal, random mating population in which there is no mutation, migration or selection, that shows a value of heterozygosity similar to that observed in the actual population (Falconer and Mackay 1996; Halliburton 2004). The value of Ne is an important parameter since it determines the expected level of heterozygosity for a population, the decay in heterozygosity and the decay in additive genetic variance. If the population size is small the additive genetic variance, and hence, the heritability will decrease, reducing the response to selection, as described in the breeder's equation (Falconer and Mackay 1996):

$R = h^2S$

Moreover, the effective size of the population determines to a great extent the efficiency of selection, and how rapidly, if at all, a particular adaptive mutation will become fixed in the population. Therefore, Ne is a critical parameter in Lynch's

argument that the structures of the eukaryotic gene and genome have evolved neutrally since eukaryotes, and particularly land plants, have very low Ne compared to microbes, including plant pathogenic ones. Is there a mistake in the calculation of Ne for eukaryotes? Are plants destined to perish because they cannot adapt as rapidly as their microbial pathogens? Or are there factors that have not been considered which allow plants to adapt at a rate comparable to that of microbes so they can survive?

Lynchs' (2006) neutral model of gene evolution can be tested under different scenarios. One violation of such a neutral model involves the effect or evidence of selection and since the efficiency of selection in eukaryotes, or rather the lack there-of, was one of Lynch's (2006) main arguments, it is precisely this we have to verify. Selection, in essence, presumes a selective factor or agent that should be of enough intensity that upon a selective pressure, adaptive variants in the population should increase in frequency, as predicted theoretically (Halliburton 2004), as a function of the fitness conferred by the allele to the homozygous and heterozygous individuals. Moreover, selective efficiency implies that the combined effect of high intensity of selection and a large enough effective population size (Halliburton 2004) should produce a pattern of polymorphism across sites at a point in time (i.e. the site frequency spectrum), and at a single allele over generations, which would be distinguishable from random allele frequency fluctuations resulting from genetic drift in large populations. In these conditions, the probability of fixation of an adaptive mutation depends only on its effect on fitness (Pr = 2s; s = effect of the allele in fitness), otherwise its probability of fixation is similar to that of a neutral variant (1/2N). Clearly, a population should have enough genetic variation to respond to that pressure, allowing it to change physiologically or morphologically. Significant changes in allele frequencies at the loci that determine the selected traits are the

outcome of the positive selection process. Therefore it is important to determine first, whether the organisms we refer to as pathogens and pests constitute a selective factor of enough importance to override the effect of genetic drift in multicellular eukaryotic populations, and second, whether there is evidence of efficient selection in plants.

Pathogens are diverse, have a great reproductive capacity and large populations

There are billions of bacterial cells in every cubic centimeter of soil. There are more bacterial cells in our body than our own cells. The total mass of bacteria is greater than that of all eukaryotes combined. There are thousands of bacteria described, yet it has been estimated that we know only about 5% of the species in this group. In fact this taxon is much more diverse than any other group of organisms, and includes two major domains: the Eubacteria and the Archaea, each many times more diverse than all the Eukaryotes together. These trivia facts should convey a mental image of just how large and diverse bacterial populations are, and similar figures can be made for other kinds of potentially pathogenic organisms. Such huge and diverse populations, although haploid in the case of bacteria, provide the source for an equal amount of genomes exposed to mutation at any given time. Although mutation rates, particularly adaptive mutation rates, can be extremely low, -in the order of 1×10^{-8} to 1×10^{-10}-, the extraordinary numbers of bacterial chromosomes that exist provide the raw material for any adaptive mutation to appear very fast, and not in geological timeframe, but within a fraction of our lifespan. For instance, antibiotics, specifically penicillin, started to be used very successfully in 1942 during World War II. However, within only a few years the first penicillin-resistant strains of *Staphylococcus aureus* appeared, and the effectiveness of penicillin was completely eliminated in 10 years. Methicillin was introduced in 1959 and two years later

Methicillin-resistant *Staphylococcus aureus* (MRSA) was detected in Europe (Taubes 2008).

Plant pathogenic bacteria and fungi are not lagging behind in their speed of adaptation. For instance, in the rapeseed fungal pathogen *Leptosphaeria maculans,* a retrotransposable element disrupted a putative avirulence factor resulting in the spread of virulence in a period of 3 years (Gout et al. 2006), rendering many previously resistant cultivars susceptible.

Microorganisms are genetically diverse and can adapt quickly to new conditions through multiple genetic exchange mechanisms, and although bacterial species that can effectively attack and infect humans or plants have smaller effective population sizes than other bacteria (Lynch 2006), they still harbor and acquire the genetic variability necessary to become resistant to whatever chemicals we produce in a blink of an eye in evolutionary time. In addition to simple mutation events like base substitutions, bacteria possess genomic encoded mechanisms to exchange entire genes or multiple genes simultaneously trough plasmids, DNA loops and transposable elements. Furthermore, bacteriophage viruses can also transfer genetic material between bacteria. These varied forms of genetic exchange are not limited to a single species but can occur between very different bacterial taxa, as exemplified by the transfer of vancomycin resistance from *Enterococcus faecalis* to *Staphylococcus aureus,* producing vancomycin (and methicillin) resistant *S. aureus*. Plasmids carrying resistance genes to multiple antibiotics have been shown to spread very fast within and between bacterial species, producing plasmid epidemics. These resistance factors encode many different types of mechanisms, either attacking the antibiotic directly, in the case of penicillin, or by modifying the antibiotics target, in the case of methicillin. Besides the resistance mechanisms, pathogens produce dozens of effector molecules aimed at disrupting any and all of the host defense and homeostasis

mechanisms, including host selective toxins that attack directly the plant defense system (Nagy et al. 2007) or basic metabolic pathways.

Additionally, recent research has shown that bacteria have evolved to have genome diversity generating mechanisms of their own, which are set in motion as as result of DNA damage and other perceivable signals of stress (Blázquez, Oliver, and Gómez-Gómez 2002; Blázquez and Gómez-Gómez 2008). In this mechanism, an operon is regulated by a repressor that binds DNA fragments and upon release, diverse genes that contribute to hypermutation are expressed, generating variability and providing the basis for selection and adaptation.

To top things off, pathogens have diverse mechanisms to persist in the environment, even after long periods of time, storing genetic diversity and remaining always ready to reactivate and attack when conditions are favorable. For example, the parasitic angiosperm *Striga*, germinates only when the seed is exposed to organic molecules exudated by plant roots, such as sorgoleone and lactone. In the absence of these molecules the seeds will remain dormant for several years even when moisture in the soil is enough for germination (Rich, Grenier, and Ejeta 2004).

How strong is the pathogen selective pressure?

The selective pressure imposed by pathogens for cultivated plants has been estimated to be around 20-30%, but it can occasionally be much higher. Cultivated plants have numerous pathogens, each of which is putting pressure on the same species. Over 20 fungal pathogens have been described for *Sorghum bicolor* alone, and although the number of pathogenic bacteria and viruses are lower than that, the reason is likely to be lack of research and not lack of pathogens. Additionally, there are an unknown number of other organisms that try, successfully or not, to take energy from the plants. It is not known how much energy a plant invests in non-host

resistance but it is probably high and positively correlated with its longevity (trees produce many more toxic chemicals than annual plants).

The diversity of potentially pathogenic organisms, their amazing effective population size (Lynch 2006), and their absolutely perplexing capacity for rapid evolutionary change impose a constant and strong selective pressure on plants. So the question is: How are plants (and ourselves) still here? In fact, some authors have proposed that the extinction of the dinosaurs had more to do with microbes than with meteorites (May and Anderson 1983). Moreover, in today's health sciences community, there is a clear recognition that our survival is in peril if we do not devise new ways to mitigate the ever increasing number of bacterial species resistant to most antibiotics we now possess. Right now, at this very moment, we are at war with real microscopic terrorists that have killed several hundred million people, -in documented history only-, indirectly through malnutrition and famine, and directly through chronic and acute diseases. There is no other more important war happening.

However, it is also evident that this pressure has existed since the origin of the first eukaryotic organisms, and all along the evolution of the different eukaryotic lineages until today. Therefore, all extant multicellular organisms are the descendents of single individuals that somehow won the battle against bacteria and other potentially pathogenic organisms. The history of how they succeeded is written in their genomes.

Is selection efficient in plants?

Selection is the mechanism by which progress is achieved in breeding programs. As plant breeders have shown repeatedly, selection can be efficient even in relatively small populations, provided enough additive genetic variation is present in the parents and the right crosses are made (Falconer and Mackay 1996). Fisher (1930), in the

Fundamental Theorem of Natural Selection wrote: "The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time", which means that the response to selection is approximately equal to the additive genetic variance for fitness at that time (Halliburton 2004). Succinctly, if there are differences in survival and reproduction in a population there is selection. If the selection intensity, by the breeder, or selective pressure by the pathogen, is low, the efficiency of selection will be low or null, particularly in small populations where the effect of genetic drift will be more important. However, if the intensity of selection is high, for instance where only a fraction of the population survives and produces the next generation, it will have a greater effect than that of genetic drift, taking the alleles coding for the selected trait, in this case disease resistance, to higher frequencies. Of course, if the population becomes significantly smaller, the effect of the genetic drift will be greater, and therefore both processes interact, but this effect will be smaller on the allele frequencies of the loci conferring selection since they increased greatly in frequency or reached fixation. Therefore strong selection for one characteristic can have profound effects on other traits because of the genetic drift that will occur in the selected parental population and the inbreeding-like effects that will ensue in a small offspring population. In foxes, for example, successful artificial selection for tameness also resulted in increased variation in morphology, specifically fur color and body shape (Belyaev, Ruvinsky, and Trut 1981; Trut, Oskina, and Kharlamova 2009), both changes due to the expression of recessive characteristics typically hidden as heterozygotes and selected against in the natural populations. Such drastic reductions in population size due to artificial or natural selection can have an important effect on the heterozygosity and additive genetic variation in the population, and therefore a significative effect on further directional selection. Hence, moderate selection

intensities acting on populations with large Ne should have the greatest efficiency of selection in the long term (Falconer and Mackay 1996).

Examples of effective selection programs in plants include the famous corn oil and protein bidirectional selection long-term program (Goldman, Rocheford, and Dudley 1993); selection after domestication for white pericarp in cultivated rice (Sweeney and McCouch 2007); selection at the tb1 locus for reduced tillering in maize during domestication; several other domestication traits; and selection for increased disease resistance, involving single or multiple loci. Additionally, the effect of selection for disease resistance has been inferred from the polymorphism, divergence and pattern of substitutions of plant genes and has been described previously. Disease related genes often show signals of balancing selection (Caicedo, Schaal, and Kunkel 1999; Stahl et al. 1999; Bergelson et al. 2001; Tian et al. 2002; Mauricio et al. 2003) and (Zamora et al. in preparation, Chapter 3 this volume, *Sorghum* uniscripts 2_8705 and 2_11684) and in a few cases signals of selective sweeps in current populations have also been identified for pathogenesis related proteins (Zamora et al. in preparation; Sorghum peroxidase). Furthermore, signals of adaptive evolution over millions of years have also been found in several plant genes (Bishop, Dean, and Mitchell-Olds 2000; Mondragón-Palomino et al. 2002; Zamora et al. 2009) which indicate that for several genes there have been numerous events of selective sweeps, leading to rapid interspecific divergence and a high dN/dS rate ratio. In practical terms, this pattern of substitutions suggests that for these genes there is no optimal phenotype and that new variation at these plant loci has been used extensively by natural selection to counteract that occurring in pathogens. Multiple different forms of mutation generate variation in the many loci involved in disease resistance and in different individuals in the population, allowing for a continuous response to selection and a never ending antagonistic coevolution.

Introgression of adaptive characters from related species has been found to occur in wild species (Grant et al. 2004; Whitney, Randell, and Rieseberg 2006), often leading to transgressive segregation. This is similar to what plant breeders have been doing by crossing good parents from differentiated populations or from different, yet closely related species (Tanksley and McCouch 1997; Xiao et al. 1998; Xu, McCouch, and Zhang 2005). All of these historical or current evidences of adaptation related to the plant defense machinery suggest that selection is efficient in plants.

Therefore there is abundant evidence that selection can be efficient in plants and other multicellular eukaryotes and that the selection intensity imposed by pathogens can be great. Such a strong selective pressure was not included in the analysis by Lynch (2006) and we propose that just as selection by pathogens influences morphological and physiological features observed in plants, it may also explain, at least in part, some of the characteristics of the gene structure, order, content and regulation of plant genomes.

What are the features that suggest diversity-generating mechanisms integrated into the plant's genome?

The plant genome encodes thousands of genes involved in disease resistance. These genes code for multiple and complex structural and chemical defenses that make plants resistant to most non-specialist native microbes (non-host resistance), including enzyme-coding as well as development regulatory genes. Large genomes have thousands of genes regulated during resistance, many are common to several responses and a few are highly specific. Rice, for example codes for ca. 600 R-genes and close to 20% of its genes have been implicated in resistance, *i.e. ca.* 12,000 genes (Goff et al. 2002).

Plants have multiple signal transduction pathways, essential to take the information of attack through-out the cell, locally to surrounding cells and tissues, and systemically to the rest of the plant. The jasmonic acid, salicylic acid and ethylene pathways transport signals initiated by different molecules and from different pathogens. Interestingly, recent evidence shows that the effect of these signal molecules in the expression of particular genes can be synergistic or antagonistic (Salzman et al. 2005) and that plants deficient in either one of these signals have a greater bacterial diversity (Kniskern, Traw, and Bergelson 2007). Indeed, defense reactions are sometimes pathogen and even strain specific (Rahimi, Perry, and Wright 1993; De Vos et al. 2005).

Additionally, cross-talking between these pathways help plants detect not only pathogens but the effect of pathogens in the cell (Pathogen Activated Molecular Patterns, PAMPs), and very likely avoid the inhibitory effect of pathogen derived ligands designed to stop the signal. Within the cell, these webs and cascades of signaling molecules increase the signal of attack and curtail pathogen attempts at breaking the transmission (Popescu et al. 2009).

The great diversity of molecules involved in defense, their large numbers in relation to the whole genome and the substantial redundancy observed at some categories of genes, both in terms of multiple paralogs and alleles, suggest that there may be mechanisms in the genome that may be used to generate functional diversity and that these mechanisms themselves have been the object of positive selection for increased adaptability.

The perceived effects of the diversity generating mechanisms

Throughout the plant genome, there are multiple events of locus, segmental and whole genome duplications (Gaut and Doebley 1997; Vision, Brown, and Tanksley 2000; Meyers et al. 2003; Paterson, Bowers, and Chapman 2004). Many of the genes that have kept paralog copies after genome duplications are implicated in defense and the multiplication of genes into large gene families has often led to the subfunctionalization of members which then work in synergy to attack pest or pathogen structures (e.g. chitinases) (Bishop et al. 2005). Clearly, the probability of getting an advantageous mutation with two copies of a gene is greater that with one copy and the more copies there are in a cluster the faster they all evolve (Bergelson et al. 2001).

It is not necessary to invoke subfunctionalization to see the advantage of having multiple copies of a gene. Even having two identical copies of one gene can be better as it may be able to make more of a particular protein, and faster, as long as they are equally regulated. However, there are cases where having one or more copies of a defense gene is actually a disadvantage, as in the case of the RPM1 gene, which in the absence of the pathogen it decreases seed production (fitness) by 9% (Tian et al. 2002).

The duplication of a gene also leads to an increased probability of having a new duplication, therefore leading to acceleration in the rate of duplication as more copies exist (Graur and Li 2000), with the consequence that more copies can harbor more alleles and increase the additive genetic variance and the response to selection. In plant genomes these multiple paralogs can be arranged in several different ways, including clusters of closely related genes; clusters of genes with related and unrelated genes interspersed and gene family members both in clusters and in singletons.

Additionally, Leister et al. (1998) demonstrated the rapid reorganization of resistance gene homologues.

The analysis of multiple eukaryotic genomes has revealed that after segmental or whole genome duplications take place, selective deletion of genes at duplication resistant loci occurs (Paterson et al. 2006). And for some categories of disease response genes it is clear that selective retention of genes has been the norm. Finally, copy number variation or the differences in the number of genes in a cluster within individuals from a population (Leister et al. 1998) and from different species (Chapter 2) may also be important for the disease resistance phenotype.

The Genomic Diversity-Generating Mechanisms

Sexual reproduction, meiosis and recombination, have been recognized for a long time as a mechanism to generate diversity for rapid adaptation, and therefore essential for coping with rapidly evolving pathogens. Recombination has a primary role in the generation of diversity in the eukaryotic genome. Several processes occurring during meiosis can generate useful, functional diversity, including: 1. Inter-allelic recombination to generate more alleles; 2. Inter-genic recombination to create more haplotypes; 3. Unequal cross-over to generate variable number of gene copies; and 4. Unequal cross-over that generates alleles with variable number of codons or domains.

In addition to the large number of genotypes that can be produced by intergenic recombination, other aspects of this mechanism have been discovered. In plant R gene clusters, gene conversion has been detected which increases the similarity of paralogs within a cluster and at the same time lead to rapid interspecific divergence (Mondragón-Palomino and Gaut 2005). Remarkably, recombination hotspots have been detected in defense related genes (Büschges et al. 1997), while cold spots exist in other gene rich regions. Also, different locations in recombination hotspots have been

found for humans and chimps (Winckler et al. 2005), which indicates that selection has led to the development of mechanisms that can generate variation in particular regions of the genome, leaving other regions untouched presumably due to stronger conservation needs.

Furthermore, particular structures exist in the genome that may facilitate shifting, through recombination, between two morphological or physiological forms which are adaptive in different conditions. For instance, truncated non-functional repeats of the *mlo* gene interfere with the function of the wild type allele and lead to broad spectrum fungal disease resistance through the constitutive upregulation of multiple defenses. However, those repeats can be excised through recombination restoring of the function of the normal allele (Piffanelli et al. 2004) which confers an energetic advantage to the individual when the pathogen is not present (Büschges et al. 1997). Therefore, this gene shows balancing selection for different alleles in natural *Sorghum bicolor* populations (Zamora et al. 2009), while the loss-of-function allele *mlo* is a marker for plants grown in monoculture (Piffanelli et al. 2004). A strikingly similar pattern is found in cultivated sorghum, where a repeat structure associated to a domestication gene conferring dwarfism, is also meiotically instable and reverts to the wild type tall phenotype (Multani et al. 2003). These polymorphisms, involving repeats that obstruct the function of the wildtype allele but that can be eliminated restoring the normal phenotype, are heritable and can be selected for in nature and in cultivated plants, and suggest yet another genomic feature that confers rapid adaptability.

Even more remarkable is the increased level of recombination that results from pathogen attack in *A. thaliana* (Lucht et al. 2002), and that the homologous recombination rate, reported as the result of biotic and abiotic stress, increases and remains high for up to four generations, as this mechanism is regulated epigenetically (Molinier et al. 2006). This suggests that if the organism detects the onset of disease,

or if it passes through a threshold of disease, it triggers increased recombination as a mechanism to generate more variation and to increase the probability of generating offspring with different but better genotypes in the presence of the pathogen, hence leading to faster adaptation. Stokes, Kunkel and Richards (2002) had shown earlier that *Arabidopsis* R genes are also epigenetically regulated, and the bal variant as well as *Arabidopsis* transgenics overexpressing the At4g16890 gene are dwarfed and constitutively activate the salicylic acid (SA)-dependent defense response pathway.

Conservation is the rule in the genome and if conditions are stable excessive change can reduce the fitness of the parents, but occasionally change is needed and in the presence of a more virulent pathogen, the generation of a few resistant individuals is likely to increase the fitness. Therefore, recombination is regulated not only in terms of the position of hotspots and coldspots in the genome, but also in frequency.

Genome duplication can lead to new species with greater adaptation potential

Whole genome duplication, either as spontaneous autopolyploidization or and allopolyploidization after interspecific hybridization are great forces that lead to the generation of novel genotypes and species in relatively short periods of time (Dubcovsky and Dvorak 2007). Interestingly, after allopolyploidization a rapid and extensive silencing of genes and rearranging of the genome occurs, which suggests that plants, having suffered several rounds of polyploidization, have the internal mechanisms to select what should stay or not (see Adams (2005) for a review). Alternatively, this occurs at random and what we see is only the end result of a process of strong purifying (and likely some positive) selection lasting sometimes millions of years (Wendel 2000; Rieseberg 2001; Rieseberg et al. 2003a).

However, hybridization doesn't always lead to allopolyploidization. In fact that may be the least common result, although one with a huge effect. More commonly,

hybridization may lead to the introgression of useful alleles from other species through repeated backcrosses. This strategy used currently by plant breeders has been found to occur in nature (Rieseberg et al. 2003a; Rieseberg et al. 2003b; Grant et al. 2004).

Transposable Elements: a necessary evil

The effect of transposable elements in altering the expression pattern of genes has been studied for several organisms. Transposons aid in the generation of new genes through the movement of exons or whole genes around the genome, such as Pack-MULEs and helitrons, and their effect on gene regulation. It is not clear why there is increased expression of transposons under stress and if these elements are responsible for some of the differences in expression patterns observed within and between populations. However, there are intriguing reports of genes and promoters with similarities to transposable elements and TEs have indeed been implicated both in the disruption of regulatory elements and genes, but also as the donors of regulatory elements in both animals and plants. In addition to this, TEs themselves may have been coopted to become functional elements of the light response mechanisms in *Arabidopsis* (Lin et al. 2007). TEs have been also linked to the instability at disease resistance loci (Bennetzen et al. 1988; Bennetzen and Hulbert 1992).

The structure of the eukaryotic gene

The modular structure of the eukaryotic gene perplexed scientists since its discovery and almost immediately led Ohta (Ohta 1989) to propose that new genes should be formed by the joining of domains commonly used in other genes. Exon shuffling may be the origin of the RNA Binding Domain gene (Sorghum bicolor 2_7407) described in Chapter 3, which has a highly conserved RBD, found in several other types of genes, but it doesn't have a full length ortholog in *Arabidopsis*, suggesting that this

gene originated after the monocot-dicot divergence as the result of the fusion of the RBD with other domains. Additionally, as described in Chapter 4, highly expressed and divergent disease resistance gene candidates have more exons than controls and have a larger size, which is consistent with the idea that these multiple exon genes may be of recent origin, in evolutionary time.

This multiple exon structure may have originated by exon shuffling, but may also be maintained by selection to allow the production of multiple variants through alternative splicing. This structure may also lead to the expression of different versions of the gene aided by the presence of multiple start codons, and to the exonization of introns.

Endosymbiosis

Infrequent but hugely successful, the symbiotic relationships that led to the origin of mitochondria and chloroplasts, must have conferred a great advantage to the individuals involved, which is evident in the great diversity of species of algae and land plants that exist today. Being able to take energy directly from the sun must have led to a rapid population expansion and a rapid radiation, as well as a novel source of genes to produce chemicals to fight off enemies, and the energy needed to produce substantial amounts of these chemicals.

Although the endosymbiosis that led to the origin of mitochondria occurred long ago, that of chloroplasts has occurred repeatedly in different lineages, indicating that these events are highly advantageous but difficult to achieve. Other kinds of symbiotic relationships exist that have an influence in disease resistance, for instance the symbiosis with arbuscular mycorrhizal fungi and nitrogen fixing bacteria. Additionally, it has been shown that endophytic fungi (Kelemu et al. 2001) and bacteria (Sturz, Christie, and Nowak 2000) produce secondary metabolites that confer

the plant some protection from potential herbivores, as well adaptation to diverse abiotic stress conditions (Rodriguez et al. 2008). These symbiotic relationships constitute the fourth dimension in the development of pathogenesis: the plant's genotype, the pathogen's genotype, the abiotic environment and the biotic environment. For many years we have known that bacteria in the external surface of our body, both our skin and our digestive system, help protect us from pathogenic organisms, but recent studies have shown the amazing diversity of bacteria (over 100 different taxa) that inhabit our different tissues and pits and the similarity of these bacteria with those found in mice, strongly suggesting a long term coevolution. Similar phenomena are being described for plants, where for instance, yeasts in the surface of grapes protect them from pathogens. In corn and other plant species the bacterium *Xanthomonas campestris* pv. zinniae rapidly degrades cercosporin, a toxin produced by the fungal pathogen *Cercospora spp.* (Taylor, Mitchell, and Daub 2006). One of the hot topics in sugarcane resistance against red rot is the identification of antagonistic microbes. Red rot is a disease caused by *Colletotrichum falcatum,* a close relative of *C. sublineolum* and *C. graminicola,* pathogens of sorghum and maize, respectively, and researchers in India[3] have been able to identify forty-nine endophytic bacterial strains that help sugarcane in the fight against red rot. These bacterial strains have been identified from several landraces, clones of *Saccharum spontaneum* and *Erianthus* species, again demonstrating the enormous value of genetic resources, not only for their own genes, but also for the beneficial organisms they harbor. Some of the beneficial bacteria identified include *Pseudomonas aeruginosa, P. fluorescens* and *P. putida.*

[3] Source: http://sugarcane-breeding.tn.nic.in/pathology.htm

Considering all the mechanisms and factors described above, it may be possible to make some predictions:

1. Wide crosses are more likely to generate useful diversity in which natural and artificial selection can be based to increment the values at traits of interest. It appears that this has been the case in rice (Xiao et al. 1998), where components for yield have increased although it is not clear whether in this case increased disease resistance occurred. In maize this has not occurred probably because being a mainly outcrossed species most of the diversity had been transferred naturally to *Zea mays* accessions already or because the variability within maize is already large and wide crosses do not improve the probability of generating useful variation substantially. Wide crosses generate this diversity both by the new combination of alleles and by the generation of new alleles by interallelic recombination and unequal crossing over (e.g. SESPY gene Chapter 3). Additionally, wide crosses could generate diversity in the population in the number of copies of genes at a locus.

2. Selection for increased number of copies in R gene clusters, or clusters of other functionally important defense related genes can lead to an increased response to selection. Evidently, more loci at disease related genes provides more opportunities for advantageous mutations to occur and can increase the efficiency of selection. Moreover, multiple copies of a gene decrease the effect of non-sense mutations on one copy: in a haploid or homozygous recessive it can be lethal, while in an organism with multiple copies it may have a minimal or no effect at all, depending on the gene (Dubcovsky and Dvorak 2007).

3. More loci with advantageous mutations can eventually lead to a cumulative effect of greater horizontal resistance. Stacking of multiple loci with advantageous mutations can amount to a very low probability that a pathogen can overcome all these

new barriers in a short to medium time frame, becoming similar to a non-host resistant phenotype.

4. Increasing the diversity at disease related loci while selecting for homogeneous harvest time, height, quality and yield can be a possible strategy to minimize the effect of pathogens while at the same time maximizing production.

5. Artificial polyploids, particularly allopolyploids should have, at least initially, an advantage due to their larger number of genes and different alleles. In autopolyploids however, despite the larger number of genes, the extreme reduction in diversity that occurs when they are generated as a new evolutionary lineage, pathogens may gain the upper-hand again quickly. Synthetic polyploids can be used to increase the diversity of polyploid crops (Zhang et al. 2005).

A proposed model: From balancing selection to gene families

Since the discovery that *A. thaliana* is a paleopolyploid (Vision, Brown, and Tanksley 2000) and the realization that most or all plants are of polyploidy origin, it has become evident that whole genome duplication is one of the most significant forces behind the generation of gene families. One truly intriguing question is that of identifying the factors that determine what turns into a multigene family and what doesn't. Paterson et al. (2006) identified several genes that have not duplicated, even though they had ample opportunity to do so after several whole genome duplications and in several widely distinct evolutionary lineages. These duplication-resistant loci appear to have a particular function that requires a single copy per genome. On the contrary, many categories of genes have developed into small, medium, and large gene families, and have acquired slightly or radically different functions. Different sources of evidence have shown that there is a correlation between the number of copies and the rate of evolutionary divergence. Bergelson et al. (2001) showed that, for plant R

genes, the Ka/Ks ratio was related to the number of paralogs in a cluster. Single copy R gene loci orthologs, *Rpm1*, *Rps2*, and *Rps5*, compared between *A. thaliana* and *A. lyrata*, showed lower rates of adaptive evolution than those exhibited by R genes belonging to clusters (Bergelson et al. 2001). Indeed, recombination within and between paralogs appears to have generated greater diversity in terms of copy number and allele diversity.

Unequal crossovers and the birth and death model of alleles (Michelmore and Meyers 1998) have been used to explain how new copies of a gene would be generated. In this scenario the two genes are generated by recombination and are conceived to be initially identical. These paralogs would then start changing, mainly through the fixation of beneficial non-synonymous substitutions (but see Komar, Lesnik, and Reiss (1999)), who show that synonymous substitutions can also have an effect)—particularly because purifying selection is likely to be smaller in members of multigene families—and then one or both would subfunctionalize or neofunctionalize.

Alternatively, when diverse selective forces acting on a locus lead to balancing selection, it is possible that through unequal cross over a heterozygote individual could generate a gamete having the two different alleles in tandem, essentially leading to a fixed heterozygote individual. A pattern fitting this model has been shown to occur in potatoes, where two canonical R genes, Rx1 and Gpa2, are located in the same cluster, are very similar to each other (92% identity) and yet recognize the coat protein of Potato Virus X and a ligand from *Globodera pallida*, a nematode, respectively (van der Vossen et al. 2000; Bakker et al. 2004). Also, in tomatoes, multiple R genes conferring resistance on *Cladosporium fulvum* are located in a cluster. *Cf-2* and *Cf-5* are closely linked, and *Cf-4* and *Cf-9* are allelic or very closely linked (Jones et al. 1993; Hammond-Kosack and Jones 1994). Complex disease resistance loci made of multiple R genes, each with severall alleles is common in crop plant species (Jones et

al. 1993). Another example is the duplication of the X-linked opsins in Old World Monkeys, which allowed both genders to achieve trichromacy by having the red and green opsin genes (96% identity). This duplication occurred after their divergence from New World Monkeys, ca. 35 MYA, in which high-frequency alleles allow heterozygous females to have trichromatic vision (Graur and Li 2000). Presumably, the selective advantage conferred by trichromatic vision could have maintained a similar polymorphism in Old World monkeys and may have led to the fixation of the duplication by unequal crossover. Recent studies in the MHC genes of voles (Arvicolinae) have also revealed the presence of extensive haplotype variation, recombination and gene duplication, all due to the strong positive selection imposed by parasites (Bryja et al. 2006). It appears then that, in these examples, subfunctionalization may have predated duplication and it is likely that this process may have lead to positive selection of individuals due to their advantageous genotypic condition, which is a special case of copy number variation.

Peroxidases, glucanases, chitinases, thaumatins, polygalacturonase inhibitors, and other families of genes involved in disease resistance have multiple paralogous copies arranged in clusters located close to the telomeres in cereals, and have been shown by us (Zamora et al. 2009) and others (Bishop, Dean, and Mitchell-Olds 2000; Bishop et al. 2005) to diverge through recurrent events of positive selection. As we have shown here, some of these genes show either multiple alleles or patterns of variation suggesting recent and strong directional, and it is possible that these gene families could have evolved as previously described. Clearly, other cases like null alleles (needed when the functional copy is targeted by the pathogen) and multiple alleles at a strictly single copy gene are exceptions to this model.

Concluding remarks

The interaction with pathogens has driven the evolution, content and organization of the plant's genome and very possibly that of other eukaryotes as well, due to at least two factors, first pathogens attack virtually any structural or molecular feature in the organism and therefore any gene can be a susceptibility factor; second, in order to cope with pathogens the plant genome (and our's) has had to develop multiple diversity generating mechanisms which include sexual reproduction, recombination, the use and regulation of transposable elements, capacitors and perhaps even polyploidization and the subsequent genome restructuring and regulation of genome functionality that normally follows such a major event.

There are two major counteracting forces in biology: parasitism and cooperation. Both of these have affected the structure and function of the Eukaryotic genome to a large extent. The signal transduction pathways that lead from recognition to action (gene expression, RNA and protein production and finally physiological and morphological changes) are varied and complex. They are also specific, in that those reactions will be of a particular form and intensity depending on the interacting organism in question. Cooperation has resulted in the extreme symbiotic relationships that eventually lead to the mitochondria and chloroplast, two organelles now essential to the functioning of the plant's cell and particularly in acquiring and the efficient use of energy. Parasitism on the other hand, has resulted over hundreds of millions of years in life forms with fantastic specialization for stealing that energy from plants and other organisms.

In addition to the diversity generating mechanisms described above, other factors may also help explain why multicellular organisms are able to cope with microbes. The cooperation that occurs among billions of cells in an organism helps in gathering

energy and resources that can be used in defense in a modulated way, as well as in growth and reproduction. External barriers are produced and reinforced as needed (e.g. callose deposition and lignification). Reservoirs of toxic compounds are stored in specialized compartments, like peroxisomes and urticating trichome vesicles. Also, numerous channels of communication to transmit the signal of attack throughout the cell and to other cells, tissues and organs, are also an essential part of how multicellular organisms fight off potential pathogens. In animals, multiple specialized cells patrol the entire organism, locating and destroying intruders.

Additionally, the nature of the game that pathogens play is also a factor that may help eukaryotes survive. If a pathogen is too successful in invading and killing the host, its success also spells its own doom, since the rarer the host becomes the smaller the pathogen population size becomes, and it could also become extinct (Nowak 2006). Theory predicts that parasites will increase in virulence when the population of hosts is increasing, and that after some time, less virulent forms will appear, mainly due to the strong selective pressure against the most virulent forms. Eventually, the parasite will remain, largely as a less virulent form, although this really depends in the host genotype and its ability to defend itself. Therefore, as it has been shown by Nowak (2006), pathogen strains in a population usually show a normal distribution for virulence, as a result of cyclical expansions and contractions of their host population, which leads to increases and decreases in the fitness of different strains, making moderately virulent strains the most successful.

Lynch (2006; 2007) has argued that the evolution of the gene and genome architecture has occurred mainly by non-adaptive processes, namely genetic drift and mutation. However, there is abundant evidence that some of these small and large

mutations have had a positive selective effect in relation to defense against pathogens, therefore the evolution of the genome may not have been completely neutral. And just as we see sexual reproduction as a normal and often advantageous diversity generating mechanism in the genome, other mechanisms may exist that have been selected for as well. The neutral theory is correct from the level of the gene to that of the genome: most mutations are deleterious, a few are neutral, and a very few are advantageous, and those latter are the ones that make the difference in fitness. We find their history written in the genomes of extant organisms.

Maybe bacteria, with all their diversity and huge population sizes, can only dream of being able to generate as much diversity as plants can, using all the mechanisms described above, to produce billions and billions of gametes with great numbers of genotypic combinations, epigenetic patterns and symbiotic associations. Considering how much information we have now in terms of gene and genome sequences and all the programs to analyze them and visualize them, it is amazing to realize how little we still now. Most genes in the genomes are still annotated as: weakly similar to putative uncharacterized protein. There is still a lot of work to do before we understand how genomes react to attack.

REFERENCES

Adams, K. L., and J. F. Wendel. 2005. Polyploidy and genome evolution in plants. Current Opinion in Plant Biology **8**:135-141.

Bakker, E., J. Van Vliet, H. Overmars, G. Smant, H. Sandbrink, E. Van der Vossen, J. Bakker, and A. Goverse. 2004. R gene homologues in potato confer resistance to distinct pathogens: a virus and a nematode. Proceeding of the Fourth International Congress of Nematology **2**:359-365.

Belyaev, D. K., A. O. Ruvinsky, and L. N. Trut. 1981. Inherited activation-inactivation of the star gene in foxes - its bearing on the problem of domestication. Journal of Heredity **72**:267-274.

Bennetzen, J. L., and S. H. Hulbert. 1992. Extramarital sex amongst the beets: Organization, instability and evolution of plant disease resistance genes. Plant Molecular Biology **20**:575-580.

Bennetzen, J. L., M. M. Qin, S. Ingels, and A. H. Ellingboe. 1988. Allele-specific and mutator-associated instability at the *rp1* disease-resistance locus of maize. Nature **332**:369-370.

Bergelson, J., M. Kreitman, E. A. Stahl, and D. Tian. 2001. Evolutionary dynamics of plant R-genes. Science **292**:2281-2285.

Bishop, J. G., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A **97**:5322-5327.

Bishop, J. G., D. R. Ripoll, S. Bashir, C. M. B. Damasceno, J. D. Seeds, and J. K. C. Rose. 2005. Selection on glycine beta-1,3-endoglucanase genes differentially inhibited by a *Phytophthora* glucanase inhibitor protein. Genetics **169**:1009-1019.

Blázquez, J., and J. M. Gómez-Gómez. 2008. Evolution of antibiotic resistance by hypermutation. Evolutionary Biology of Bacterial and Fungal Pathogens:319.

Blázquez, J., A. Oliver, and J. M. Gómez-Gómez. 2002. Mutation and evolution of antibiotic resistance: Antibiotics as promoters of antibiotic resistance? Current Drug Targets **3**:345-349.

Bryja, J., M. Galan, N. Charbonnel, and J. Cosson. 2006. Duplication, balancing selection and trans-species evolution explain the high levels of polymorphism of the DQA MHC class II gene in voles (Arvicolinae). Immunogenetics **58**:191-202.

Büschges, R., K. Hollricher, R. Panstruga et al. 1997. The barley *mlo* gene: a novel control element of plant pathogen resistance. Cell **88**: 695-705.

Caicedo, A. L., B. A. Schaal, and B. N. Kunkel. 1999. Diversity and molecular evolution of the RPS2 resistance gene in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America **96**:302-306.

De Vos, M., V. R. Van Oosten, R. M. P. Van Poecke et al. 2005. Signal Signature and Transcriptome Changes of *Arabidopsis* During Pathogen and Insect Attack. Molecular Plant-Microbe Interactions **18**:923-937.

Dubcovsky, J., and J. Dvorak. 2007. Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Science **316**:1862-1866.

Falconer, D., and T. Mackay. 1996. Introduction to Quantitative Genetics (4th Edition):Prentice Hall.

Fisher, R. 1930. The genetical theory of natural selection. Oxford:Clarendon Press.

Gaut, B. S., and J. F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences of the United States of America **94**:6809-6814.

Goldman, I. L., T. R. Rocheford, and J. W. Dudley. 1993. Quantitative trait loci influencing protein and starch concentration in the Illinois Long Term Selection maize strains. TAG Theoretical and Applied Genetics **87**:217-224.

Gout, L., I. Fudal, M. L. Kuhn, F. Blaise, M. Eckert, L. Cattolico, M. H. Balesdent, and T. Rouxel. 2006. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Molecular Microbiology **60**:67-80.

Grant, P. R., B. R. Grant, J. A. Markert, L. F. Keller, and K. Petren. 2004. Convergent evolution of Darwin's finches caused by introgressive hybridization and selection. Evolution **58**:1588-1599.

Graur, D., and W.-H. Li. 2000. Fundamentals of Molecular Evolution. Sunderland, Massachusetts:Sinauer Associates, Inc.

Halliburton, R. 2004. Introduction to population genetics. Upper Saddle River, New Jersey:Pearson Prentice Hall.

Hammond-Kosack, K. E., and J. D. G. Jones. 1994. Incomplete dominance of tomato Cf genes for resistance to *Cladosporium fulvum*. Molecular Plant Pathogen Interactions **7**:58-70.

Jones, D. A., M. J. Dickinson, P. J. Balint-Kurti, M. S. Dixon, and J. D. G. Jones. 1993. Two Complex Resistance Loci Revealed in Tomato by Classical and RFLP Mapping of the Cf-2, Cf-4, Cf-5, and Cf-9 Genes for Resistance to *Cladosporium fulvum*. Molecular Plant Pathogen Interactions **6**:348-357.

Kelemu, S., J. F. W. Jr, F. Munoz, and Y. Takayama. 2001. An endophyte of the tropical forage grass *Brachiaria brizantha*: Isolating, identifying, and characterizing the fungus, and determining its antimycotic properties. Canadian Journal of Microbiology, . **47**:55-62.

Kniskern, J. M., M. B. Traw, and J. Bergelson. 2007. Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on *Arabidopsis thaliana*. Molecular Plant-Microbe Interactions **20**:1512-1522.

Komar, A. A., T. Lesnik, and C. Reiss. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Letters **462**:387-391.

Leister, D., J. Kurth, D. A. Laurie, M. Yano, T. Sasaki, K. Devos, A. Graner, and P. Schulze-Lefert. 1998. Rapid reorganization of resistance gene homologues in cereal genomes. Proc Natl Acad Sci U S A **95**:370-375.

Lin, R. C., L. Ding, C. Casola, D. R. Ripoll, C. Feschotte, and H. Y. Wang. 2007. Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. Science **318**:1302-1305.

Lucht, J. M., B. Mauch-Mani, H.-Y. Steiner, J.-P. Metraux, J. Ryals, and B. Hohn. 2002. Pathogen stress increases somatic recombination frequency in *Arabidopsis*. Nature Genetics **30**:311 - 314.

Lynch, M. 2007. Colloquium Papers: The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences **104**:8597-8604.

Lynch, M. 2006. The Origins of Eukaryotic Gene Structure. Mol Biol Evol **23**:450-468.

Mauricio, R., E. A. Stahl, T. Korves, D. Tian, M. Kreitman, and J. Bergelson. 2003. Natural selection for polymorphism in the disease resistance gene Rps2 of *Arabidopsis thaliana*.

May, R. M., and R. M. Anderson. 1983. Epidemiology and genetics in the coevolution of parasites and hosts. Proc R Soc Lond B Biol Sci **219**.

Meyers, B. C., A. Kozik, A. Griego, H. Kuang, and R. W. Michelmore. 2003. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*.

Michelmore, R. W., and B. C. Meyers. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research **8**:1113-1130.

Molinier, J., G. Ries, C. Zipfel, and B. Hohn. 2006. Transgeneration memory of stress in plants. Nature **442**:1046-1049.

Mondragón-Palomino, M., and B. S. Gaut. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. Molecular Biology and Evolution **22**:2444-2456.

Mondragón-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Research **12**:1305-1315.

Multani, D. S., S. P. Briggs, M. A. Chamberlin, J. J. Blakeslee, A. S. Murphy, and G. S. Johal. 2003. Loss of an MDR Transporter in Compact Stalks of Maize br2 and Sorghum dw3 Mutants. Science **302**:81-84.

Nagy, E., T.-C. Lee, W. Ramakrishna et al. 2007. Fine mapping of the *Pc* locus of *Sorghum bicolor* , a gene controlling the reaction to a fungal pathogen and its host-selective toxin. TAG Theoretical and Applied Genetics **114**:961-970.

Nowak, M. 2006. Evolutionary Dynamics: Exploring the Equations of Life:Belknap Press.

Ohta, T. 1989. Role of gene duplication in evolution. Genome **31**:304-310.

Paterson, A. H., J. E. Bowers, and B. A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America **101**:9903-9908.

Paterson, A. H., B. A. Chapman, J. C. Kissinger, J. E. Bowers, F. A. Feltus, and J. C. Estill. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis, Oryza, Saccharomyces* and *Tetraodon*. Trends Genet **22**:597-602.

Piffanelli, P., L. Ramsay, R. Waugh, A. Benabdelmouna, A. D'Hont, K. Hollricher, J. H. Jorgensen, P. Schulze-Lefert, and R. Panstruga. 2004. A barley cultivation-associated polymorphism conveys resistance to powdery mildew. Nature **430**:887-891.

Popescu, S. C., G. V. Popescu, S. Bachan, Z. Zhang, M. Gerstein, M. Snyder, and S. P. Dinesh-Kumar. 2009. MAPK target networks in *Arabidopsis thaliana* revealed using functional protein microarrays. Genes & Development **23**:80-92.

Rahimi, S., R. N. Perry, and D. J. Wright. 1993. Induction and detection of pathogenesis-related proteins in leaves and roots of potato plants infected with pathotypes of *Globodera pallida*. Fundamental & Applied Nematology **16**:549-556.

Rich, P. J., C. Grenier, and G. Ejeta. 2004. Striga Resistance in the Wild Relatives of Sorghum. Crop sci **44**:2221-2229.

Rieseberg, L. H. 2001. Chromosomal rearrangements and speciation. Trends in Ecology & Evolution **16**:351-358.

Rieseberg, L. H., O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone, T. Nakazato, J. L. Durphy, A. E. Schwarzbach, L. A. Donovan, and C. Lexer. 2003a. Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. Science **301**:1211-1216.

Rieseberg, L. H., A. Widmer, A. M. Arntz, and J. M. Burke. 2003b. The genetic architecture necessary for transgressive segregation is common in both natural

and domesticated populations. Philosophical Transactions of the Royal Society of London B Biological Sciences **358**:1141-1147.

Rodriguez, R. J., J. Henson, E. Van Volkenburgh, M. Hoy, L. Wright, F. Beckwith, Y. O. Kim, and R. S. Redman. 2008. Stress tolerance in plants via habitat-adapted symbiosis. The ISME journal **2**:404-416.

Salzman, R. A., J. A. Brady, S. A. Finlayson et al. 2005. Transcriptional Profiling of Sorghum Induced by Methyl Jasmonate, Salicylic Acid, and Aminocyclopropane Carboxylic Acid Reveals Cooperative Regulation and Novel Gene Responses. Plant Physiol. **138**:352-368.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman, and J. Bergelson. 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. Nature **400**:667-671.

Stokes, T. L., B. N. Kunkel, and E. J. Richards. 2002. Epigenetic variation in *Arabidopsis* disease resistance. Genes Dev. **16**:171-182.

Sturz, A. V., B. R. Christie, and J. Nowak. 2000. Bacterial Endophytes: Potential Role in Developing Sustainable Systems of Crop Production. Critical Reviews in Plant Sciences **19**:1 - 30.

Sweeney, M., and S. McCouch. 2007. The Complex History of the Domestication of Rice. Ann Bot **100**:951-957.

Tanksley, S. D., and S. R. McCouch. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. Science (Washington D C) **277**:1063-1066.

Taubes, G. 2008. The Bacteria Fight Back. Science **321**:356-361.

Taylor, T. V., T. K. Mitchell, and M. E. Daub. 2006. An Oxidoreductase Is Involved in Cercosporin Degradation by the Bacterium Xanthomonas campestris pv. zinniae. Appl. Environ. Microbiol. **72**:6070-6078.

Tian, D., H. Araki, E. Stahl, J. Bergelson, and M. Kreitman. 2002. Signature of balancing selection in *Arabidopsis*. Proc Natl Acad Sci U S A **99**:11525-11530.

Trut, L., I. Oskina, and A. Kharlamova. 2009. Animal evolution during domestication: the domesticated fox as a model. Bioessays **31**:349-360.

van der Vossen, E. A. G., J. van der Voort, K. Kanyuka, A. Bendahmane, H. Sandbrink, D. C. Baulcombe, J. Bakker, W. J. Stiekema, and R. M. Klein-Lankhorst. 2000. Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. Plant Journal **23**:567-576.

Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. Science **290**:2114-2117.

Wendel, J. F. 2000. Genome evolution in polyploids. Plant Molecular Biology **42**:225-249.

Whitney, K. D., R. A. Randell, and L. H. Rieseberg. 2006. Adaptive introgression of herbivore resistance traits in the weedy sunflower Helianthus annuus. American Naturalist **167**:794-807.

Winckler, W., S. R. Myers, D. J. Richter et al. 2005. Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. Science **308**:107-111.

Xiao, J. H., J. M. Li, S. Grandillo, S. N. Ahn, L. P. Yuan, S. D. Tanksley, and S. R. McCouch. 1998. Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. Genetics **150**:899-909.

Xu, Y. B., S. R. McCouch, and Q. F. Zhang. 2005. How can we use genomics to improve cereals with rice as a reference genome? Plant Molecular Biology **59**:7-26.

Zamora, A., Q. Sun, M. T. Hamblin, C. F. Aquadro, and S. Kresovich. 2009. Positively selected disease response orthologous gene sets in the cereals identified using *Sorghum bicolor* L. Moench expression profiles and comparative genomics. Mol Biol Evol **26**:2015-2030.

Zhang, P., S. Dreisigacker, A. E. Melchinger, J. C. Reif, A. M. Kazi, M. Van Ginkel, D. Hoisington, and M. L. Warburton. 2005. Quantifying novel sequence variation and selective advantage in synthetic hexaploid wheats and their backcross-derived lines using SSR markers. Molecular Breeding **15**:1-10.