ramosa1 in the development and evolution of inflorescence architecture in grasses

Brandi Adell Sigmon
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Cell and Developmental Biology Commons, and the Genetics and Genomics Commons

Recommended Citation
Sigmon, Brandi Adell, "ramosa1 in the development and evolution of inflorescence architecture in grasses" (2010). Graduate Theses and Dissertations. 11804.
https://lib.dr.iastate.edu/etd/11804

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
ramosa1 in the development and evolution of inflorescence architecture in grasses

by

Brandi Adell Sigmon

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

PHILOSOPHY OF SCIENCE

Major: Genetics

Program of Study Committee:
Erik Vollbrecht, Major Professor
Philip Becraft
Volker Brendel
Lynn Clark
Jeanne Serb

Iowa State University
Ames, Iowa
2010
Copyright © Brandi Adell Sigmon, 2010. All rights reserved.
# TABLE OF CONTENTS

## ABSTRACT

### CHAPTER 1. GENERAL INTRODUCTION

- Dissertation Organization 1
- Literature Review 2
- Research Goals 14
- Literature Cited 16
- Tables and Figures 21

### CHAPTER 2. EVIDENCE OF SELECTION AT THE RAMOSAI LOCUS DURING MAIZE DOMESTICATION

- Abstract 23
- Introduction 24
- Materials and Methods 27
- Results 29
- Discussion 39
- Acknowledgements 44
- Literature Cited 45
- Tables and Figures 50

### CHAPTER 3. MOLECULAR EVOLUTION OF RAMOSAI IN THE ANDROPOGONEAE (POACEAE)

- Abstract 60
- Introduction 60
- Materials and Methods 63
- Results 65
- Discussion 74
- Acknowledgements 79
- Literature Cited 80
- Tables and Figures 83

### CHAPTER 4. GENOME-WIDE COMPARATIVE ANALYSIS REVEALS THE IMPORTANCE OF LINEAGE-SPECIFIC EXPANSION IN THE EVOLUTION OF THE EPF SUBFAMILY IN PLANTS

- Summary 94
- Background 95
- Results 98
- Discussion 107
- Conclusions 114
- Methods 114
- Acknowledgements 116
- Literature Cited 117
- Tables and Figures 121
CHAPTER 5. GENERAL CONCLUSIONS 131
Summary and Discussion 131
Literature Cited 134

APPENDIX A. FASCIATED PHENOTYPES IN THE MAIZE LANDRACE VERACRUZ 85 136
Abstract 136
Introduction 136
Methods and Materials 137
Results 139
Conclusions 141
Acknowledgements 142
Literature Cited 143
Tables and Figures 144

APPENDIX B. SUPPLEMENTARY DATA 152
Tables and Figures 152

ACKNOWLEDGEMENTS 157
ABSTRACT

Branching architecture of the flower-bearing structures in grasses, known as inflorescences, is a morphological trait that is subject to natural and artificial selection since it affects both reproduction and grain yield. Genes underlying this trait are sought to explain the molecular basis underlying the phenotypic diversity of these structures. The *ramosa1 (ra1)* gene encodes a transcription factor that controls branching architecture in maize inflorescences (tassel and ear). Reduced *ra1* activity in maize produces ears with crooked rows of kernels due to the generation of extra spikelets, a phenotype that may have been selected on during the derivation of modern maize. Patterns of nucleotide diversity coupled with statistical tests and phylogenetics suggest a regulatory element at the *ra1* locus was a target of artificial selection during maize’s domestication from its wild progenitor teosinte. We also narrowed the timeframe for the probable origin of *ra1* during the evolution of grasses and found sequence variations in some species correlate with their respective inflorescence architectures. These results suggest this gene was important in the evolution of inflorescence architecture in other grasses, most notably in sorghum where statistical tests show *ra1* may have been a target of artificial selection during its evolution, most likely to increase grain yield. Since this gene may have been important during the domestication and cultivation of two crops, maize and sorghum, this research may lead to future breeding projects to increase grain yield in these and other cereal grasses.
CHAPTER 1. GENERAL INTRODUCTION

DISSERTATION ORGANIZATION

This dissertation includes a general introduction (Chapter 1), three journal papers (Chapters 2-4), general conclusions (Chapter 5), and two appendices. The general introduction contains a literature review including a discussion of the importance of research into the evolution of plant development, a summary of grass inflorescence development, the history of the domestication of maize, and the identification of potential domestication loci. Chapter 2 comprises a paper published in Molecular Ecology, which presents evidence that the ra1 locus was a target of artificial selection during the domestication of maize. My contributions include the planning and execution of the experimental design and writing of the manuscript in conjunction with Dr. Vollbrecht. Chapter 3 is a manuscript planned for future publication in a collaborative project with Dr. Elizabeth Kellogg. It only consists of my contributions to the project, which includes the genomics, wet lab work, sequence analyses and phylogenetics for ramosa1 in the grass tribe Andropogoneae. The final paper will also include Dr. Kellogg’s contributions, primarily analyses for some grass species outside of the Andropogoneae. Chapter 4 is a paper to be submitted to BMC Genomics, which explores the evolution of the EPF subfamily of transcription factors, of which ra1 is a member of, in land plants. My contributions include the design of the experimental plan, the interpretation and implications of the results, and drafting the manuscript with Dr. Vollbrecht. Analyses including the data mining, gene annotations, construction of the alignments, and phylogenetics were done in collaboration with Rebecca Weeks. Chapter 5 is a summary of the general conclusions and a discussion of implications garnered from the totality of the three previous chapters. The first appendix presents results from a new project characterizing and mapping putative loci involved in the regulation of meristem size discovered in a maize landrace. Molecular work on this project was started toward the end of my Ph.D tenure and the uncompleted nature warrants its inclusion as an appendix instead of a full chapter. Finally, the second appendix contains supplementary figures and tables for the previous chapters. My contributions to a review article entitled “Amazing grass:
developmental genetics of maize domestication” published in *Biochemical Society Transactions* is not included in the dissertation.

LITERATURE REVIEW

The Evolution of Plant Development

The study of developmental evolution, known as evo-devo, a product of a recent synthesis of developmental and evolutionary biology, is still a young developing field. For example, in 1999 it was first given its own division in the Society for Integrative and Comparative Biology (Goodman and Coughlin 2000). Researchers in this field address questions concerning the evolution of phenotypic change and the genetic and developmental processes behind changes in shape and form of organisms. More specifically they address questions concerning developmental plasticity, modularity, and how genes can act as switches, such that spatial and/or temporal changes in gene expression and/or function can lead to phenotypic novelties. Most studies in this field consist of correlative evidence, as it is a daunting task to find causative evidence of how a genetic change results in a new phenotype. This is particularly true for plants. In fact only a few clear examples of plant evolutionary development exist that bridge the gap between genotype and phenotype at present, one of which is the maize domestication locus *teosinte branched1* (*tb1*) (Wright et al. 2005), which controls plant architecture in maize. After over fifteen years of work on *tb1*, its function in plant development and the role environment has on its function is well documented (Doebley and Stec 1993; Doebley, Stec, and Gustus 1995; Doebley, Gaut, and Smith 2006). Also, the identification of a *cis*-regulatory element as the genetic basis for the phenotypic change during the domestication of maize, bridges the gap between genotype and phenotype (Doebley, Stec, and Gustus 1995; Lukens and Doebley 2001; Clark et al. 2006). Another example comes from work on *Mimulus* (Bradshaw et al. 1998; Bradshaw and Schemske 2003; Ramsey, Bradshaw, and Schemske 2003). These researchers have shown there is a genetic difference in flower color and that the phenotype makes a difference in pollinator visitation. These two examples illustrate how information must be gathered on multiple levels in order to demonstrate that a genetic difference causes a phenotypic change that may, in turn, be adaptive.
As illustrated through the work on *tb1*, the study of the evolution of domestication loci offers a unique opportunity to contribute to the field of evo-devo. Darwin viewed artificial selection during domestication and breeding as a microcosm of natural selection. Therefore, the study of the genes behind phenotypic changes important during the domestication of crops or animals from their wild progenitors can offer insight into the role these genes may have in the adaptation of related species. The actual targets of selection are often resolved to regulatory genes, particularly transcription factors (Doebley and Lukens 1998; Zhao et al. 2008). In particular, changes in cis-regulatory sequences have been implicated as the genetic basis behind phenotypic changes, as illustrated by *tb1* (Clark et al. 2004; Clark et al. 2006). Multiple cis-regulatory sequences enable a gene to respond differentially to combinations of transcriptional regulators (Wray et al. 2003). This fine control allows for differential spatial and/or timing of expression in a developing organism. Therefore, it is intuitive that cis-regulatory elements are often the targets of selection, and changes in expression levels or timing of expression may account for many of the differences between domesticates and their wild progenitors. However, there are also examples of protein changes in domestication loci that can lead to selected, phenotypic changes. An example in maize includes the *teosinte glume architecture1* (*tga1*) gene, which is responsible for exposing the endosperm of maize kernels by the elimination of the hardened fruitcases that are present in the progenitor of maize; this change significantly increased the ease of harvest (Wang et al. 2005). It is hypothesized that a single amino change in the protein sequence is responsible for this change in phenotype. Domestication occurs over several hundred or thousand years at least, which is a blink in evolutionary time and changes in gene regulation or protein function can happen quickly, but may have huge morphological consequences. This situation provides a potential mechanism behind the formation of “Goldschmidt’s hopeful monsters” (Goldschmidt 1940) as the modern maize ear is truly a monstrosity compared to its progenitor (Figure 1). Goldschmidt hypothesized that small alterations, especially in genes that control timing of developmental changes, could lead to drastic differences in adult form. As such, maize domestication is an example of how small heterochronic or heterotopic changes can lead to great morphological differences that could be considered macroevolutionary (Vollbrecht and Sigmon 2005).
Grass Inflorescence Development as a Model System

To conduct evo-devo research, a model system with substantial resources is desirable due to the interdisciplinary nature of the research required to address questions concerning both development and evolution. These resources include, but are not limited to, the presence of phenotypic diversity in natural populations, developmental mutants and other genetic resources, a well-supported phylogeny, gene expression technology, and genomics resources such as transcriptome and genome sequence data. Grasses are fast becoming an ideal model system for plant evo-devo studies as all or most of these tools are available for several species. Rice, maize, and sorghum all have complete genome sequences (Goff et al. 2002; Paterson et al. 2009; Schnable et al. 2009), extensive EST data, means of mutagenesis in addition to classical mutants, and the ability to generate transgenic plants. Many of these tools will soon be available for other emerging grass models including *Brachypodium distachyon*, *Setaria italica* (Foxtail millet), *Pennisetum glaucum* (Pearl millet), and several subspecies of rice and other *Oryza* species (Buell 2009). In addition, the grasses have a well-supported phylogeny, which provides an evolutionary context (GPWG 2000; GPWG 2001). Grasses also possess a wealth of phenotypic diversity, which includes plant and inflorescence architecture, growth habits, flowering time, pathogen defense mechanisms, and responses to environmental stressors. Most of these traits have been researched in eudicot systems, but the conclusions from these studies are often not directly transferable to monocots due to the divergence between monocots and eudicots over the last 170 million years (Wikstrom, Savolainen, and Chase 2001).

In particular, grass inflorescences are a good model system due to their great phenotypic diversity in architecture. Being reproductive structures, they are visible to both natural and artificial selection (Doust 2007). It has been suggested that since inflorescence structure is so variable and is a homoplasious trait, it must have been relatively easy to modify in evolutionary time (Mathews et al. 2002). This suggests there may be many possible genetic mechanisms that can be modified or manipulated to facilitate change in inflorescence structure. Therefore, a comparative study of these multiple avenues to phenotypic change can lead to more general conclusions concerning the genetic causes behind adaptive phenotypic change in grasses, which is a main goal of evo-devo. This pursuit
is also aided by a wealth of mutants that affect inflorescence development in the grass model organisms rice, sorghum, and maize. Since these genes have a role in determining inflorescence structure and grain yield they may have been targets of artificial selection during domestication or improvement, as mentioned previously (Dillon et al. 2007; Doust 2007). Data gleaned from these model organisms can then be used to test hypotheses in nonmodel grasses, which is also facilitated by the whole genome sequences. In this way, moving from grass models to nonmodels, research can readily transition from a study of artificial to natural selection.

**Maize Inflorescence Development**

In order to develop hypotheses concerning molecular genetic mechanisms in the evolution of inflorescence development in grasses, it is necessary to have an understanding of inflorescence development and the roles of particular genes in that process. For example, consider inflorescence development in maize (Vollbrecht and Schmidt 2009). Inflorescence architecture is determined by the arrangement of long indeterminate branches and short determinate ones. Flowers, the last structures produced, are made after the branching architecture has been completed. Although the branching architectures of the male (tassel) and female (ear) maize inflorescences appear to be morphologically distinct, their basic organization and underlying development is similar until flowers are initiated. Understanding of their development has been elucidated through the characterization of inflorescence development mutants (Veit et al. 1993). Some mutants such as *liguleless2* (*lg2*), *barren stalk1* (*ba1*), *suppressor of sessile spikelets1* (*sos1*) reduce production of branches and/or the number of spikelets (Walsh and Freeling 1999; Gallavotti et al. 2004) whereas others like the *ramosa* genes, *branched silkless1* (*bd1*), and *indeterminate spikelet1* (*ids1*) increase branching and/or the number of spikelets (Chuck, Meeley, and Hake 1998; Chuck et al. 2002; Vollbrecht et al. 2005; Bortiri et al. 2006; Satoh-Nagasawa et al. 2006).

During maize ontogeny, the terminal shoot apical meristem (SAM) conditions indeterminate vegetative growth of the shoot until a signal is received to transition to a flowering stage of development. Once the signal is received the SAM is converted into an inflorescence meristem (IM), which will go on to produce the tassel. The ear arises from the
analagous conversion of an axillary meristem (AM) to a lateral IM, which occurs just after the conversion of the terminal SAM. Among the genes known to first act just after the transition of an SAM or AM to an IM are *fasciated ear2 (fea2)* and *thick tassel dwarf1 (td1)* (Taguchi-Shiobara et al. 2001; Bommert et al. 2005). Mutants of these two genes result in enlargement of the IM, which causes overproliferation of organs, primarily resulting in fasciation of the ear and the main tassel axis. During normal development, the IM then goes on to produce higher order meristems in a progressive manner (secondary, tertiary, etc.) in both the tassel and ear. In other words, the IM produces an indeterminate number of spikelet pair meristems (SPM), which in turn produces a pair of spikelet meristems (SM). The mutant *barren inflorescence2 (bif2)* affects the transition of the IM to SPM, which results in the production of spikelets and branches (McSteen and Hake 2001). Finally, in normal development, the SM produces a pair of floral meristems (FM), which will produce the floral organs. The mutant *branched silkless1 (bd1)* is a gene hypothesized to be required for FM identity as in mutants, the ear FMs are replaced by BMs, which then produce SMs (Chuck et al. 2002). One notable difference in ear and tassel development is that in the tassel IM first produces a few indeterminate BMs that make long branches before transitioning to the production of a short spikelet pair branches. However, the ear IM typically does not produce BMs and therefore remains unbranched. The *ramosa* genes are involved in this transition from BM to SPM production (Vollbrecht et al. 2005; Bortiri et al. 2006; Satoh-Nagasawa et al. 2006). Mutant inflorescences have more long branches, suggesting the role of these genes is suppression of BM production in maize inflorescences (Vollbrecht et al. 2005; Bortiri et al. 2006). With this basic introduction into inflorescence development, I will now focus more on the role of *ral* in the development of inflorescence architecture.

*r al* is a classical mutant that was found in a framer’s field and first described in 1912. The morphology was so aberrant, it was first thought to be a separate species and was thus named *Zea ramosa* (Gernert 1912). The gene responsible was determined to be a putative EPF-type C2H2 zinc finger transcription factor that is expressed in the boundary domain near the meristem base and is responsible for imposing a more determinate short branch identity (Vollbrecht et al. 2005). Morphologically, strong mutant tassels have increased long branching that extends to the apex due to increased indeterminacy of the meristems.
(Vollbrecht et al. 2005). The ear also produces long branches leading to a highly disorganized ear. Weaker mutant alleles typically have ears with crooked rows and tassels with increased branching, though less severe than seen in strong mutants. Therefore, the function of *ra1* is to determine inflorescence architecture, which includes proper packing of kernels in straight rows on the ear, which may have been of significance during domestication and/or improvement. When *ra1* was examined in the maize inbred diversity lines, a signature of selection was detected, but with these data alone it could not be determined if selection occurred during domestication, improvement, or during the evolution of teosinte (Vollbrecht et al. 2005). This evidence of selection is not altogether surprising, as it is intuitive that inflorescence development genes may have been subjected to artificial selection at some point because of their role in organizing inflorescence architecture thereby putatively impacting grain yields.

**Identification and Characterization of Domestication Loci**

During the Agricultural Revolution, which began ~10,000 years ago, prehistoric farmers began to domesticate plants and animals through artificial selection on traits of importance including in the case of plants, growth habit, plant and inflorescence architecture, grain quality and yield, and ease of harvest among numerous other traits (Smith 1995). The domestication process typically spans hundreds to thousands of years in which traits may be selected on one at a time or in combination with other preferable traits. During domestication, the breeding population is reduced resulting in a bottleneck where there is an overall reduction in diversity for all genes. However, loci that underwent selection in the past, so-called domestication loci, will display an even greater reduction when compared to other non-selected neutral loci. Thus, this signature of selection may be detected through tests for altered nucleotide diversity. Typically, researchers will employ multiple tests to ensure existence of a signature of selection for a domestication locus. In this section I will describe in some detail the domestication process in maize and sorghum focusing on well-studied domestication loci in order to compare and contrast the effects domestication can have on different crops.
Maize Domestication

Archeological evidence indicates maize was domesticated ~5,000-10,000 years ago (Smith 1995) and molecular evidence supports a single domestication event (Matsuoka et al. 2002). Around 9000 years ago in the Balsas River Valley of Mexico, modern domestic maize was derived from *Zea mays* ssp. *parviglumis*, also known as a teosinte (Doebley 2004). The term “teosinte” is derived from the Nahuatl Indian word “teocintli” meaning “grain of the gods” and it refers to a group of annual (*Zea luxurians* and *Zea mays*) and perennial (*Zea perennis* and *Zea diploperennis*) species of the genus *Zea*, which is indigenous to Mexico and Central America (Doebley 2004; Fukunaga et al. 2005). Teosintes and maize were once placed into different genera due to their striking differences in plant architecture and ear morphology, although they do share similar growth forms (Doebley 2004). The differences in ear morphology present some of the most stunning differences. For example, the teosinte ear produces only 5-12 kernels, enclosed in stony fruitcases that disperse as the ear disarticulates, whereas the modern maize ear has several hundred kernels, each firmly attached to the cob without an enclosing fruitcase (Figure 1). Since maize lacks a dispersal mechanism and the kernels have no protective fruitcases, its survival is dependent on humans, which also illustrates how plants and humans have co-evolved during domestication (Purugganan and Fuller 2009).

During the twentieth century several hypotheses were put forward concerning the origins of domesticated maize. Most data support the teosinte hypothesis first proposed by George Beadle, which portrays teosinte as the single progenitor of maize (Beadle 1939). Beadle suggested ancient peoples cultivated teosinte as a source of food, and during that process mutations arose that improved its usefulness and these mutations were selected upon by humans. He also hypothesized that only five major mutations could convert teosinte into a primitive form of maize, and that humans had selected other major and minor mutations over time (Beadle 1939).

Later, a study looking for QTL involved in the morphological differences between maize and teosinte found five QTLs of strong effect (Doebly and Stec 1993), which was in line with Beadle’s teosinte hypothesis. Some of these QTL have been resolved to probable individual genes, which encode transcriptional regulators that have strong signatures of
selection. The first gene identified, was *teosinte branched1* (*tb1*), so named because the highly tillered plant architecture of *tb1* mutants resembles teosinte (Burnham 1959; Doebley, Stec, and Gustus 1995). *tb1* plays a major role in maize plant architecture by controlling the fate of axillary meristems through repressing organ growth where it is expressed (Doebley, Stec, and Hubbard 1997). Mechanistically, maize alleles express *tb1* at higher levels than teosinte alleles, which explains their differences in plant architecture. The difference in gene expression suggests a regulatory element may have been the target of selection. Indeed, patterns of nucleotide diversity at the *tb1* locus show a large reduction in diversity upstream of the transcribed region in maize (Clark et al. 2004). Recently the target was resolved to a region ~60 kb upstream from the coding region, which suggests a distant enhancer was the actual target of selection (Clark et al. 2006). Another gene that affects plant architecture, *barren stalk1* (*ba1*), is a putative domestication locus as it maps to another of the five QTL regions. *ba1* may interact with *tb1* to regulate vegetative shoot architecture as it is required for the initiation of axillary meristems (Gallavotti et al. 2004). A third candidate is *teosinte glume architecture* (*tga1*) gene, which confers soft glumes to maize and stony fruitcases to teosinte (Dorweiler and Doebley 1997) due to a single amino acid difference in the coding sequence (Wang et al. 2005). Both *tb1* and *ba1* have selected alleles in extant teosinte populations. However, *tga1* selected maize alleles seem to be mutant relative to teosinte *tga1* (Wang et al. 2005). Although the five QTL support the teosinte hypothesis, many of the differences between teosinte and maize may not be due to mutations but to novel combinations of teosinte’s incredible genotypic diversity (Fukunaga et al. 2005).

Domestication loci can affect a variety of traits beyond overall morphology including, for instance, biochemical traits such as genes involved in the starch pathway (Whitt et al. 2002) or carotenoid production. One example is the *yellow1* (*y1*) gene, which is responsible for the carotenoid-rich yellow endosperm typical of cultivated maize. This gene is suspected to have been subjected to strong artificial selection at some point as an asymmetrical selective sweep extends across a region of >0.5 Mb, the largest selective sweep identified in maize to date (Palaisa et al. 2004). Indeed, in the last few years many other candidate loci have been found, with the advent of sequence-based techniques for identifying signatures of selection through genome surveys. This approach is commonly called “bottom-up” because
genomic screens of diversity are used to identify candidate loci with no prior phenotypic knowledge. In a recent large-scale screen 2-4% of 774 genes examined have experienced selection implying as many as 1200 genes in the entire genome may have undergone selection (Wright et al. 2005). In an even larger screen of 1095 genes in maize inbreds, 35 potential targets of selection were identified which is a similar proportion (Yamasaki et al. 2005). Both of these studies show it is feasible to find candidate domestication loci in a high-throughput manner without any prior knowledge of gene function or phenotype. This approach to candidate loci identification is opposite to the other approach known as the “top-down” approach which uses *a priori* knowledge of phenotype to identify logical candidates followed by tests for evidence of selection. Both of these approaches have benefits and drawbacks, but both are important for elucidating the history of maize and its domestication from teosinte.

*Sorghum Domestication*

With the resolution of putative domestication QTL to candidate genes and identification through large-scale screens, much progress has been made in understanding maize’s domestication from its wild progenitor teosinte. The same cannot be said of sorghum because of inherent difficulties such as a self mating system, more extensive linkage disequilibrium compared to most other crops, and low nucleotide diversity across the genome (Hamblin et al. 2006). The study of sorghum’s domestication is important given it is the fifth most important grain crop in the world (FAO 2004). Sorghum last shared a common ancestor with maize 12-20 million years ago, and as such, is considered a close relative (for review, see Gaut et al. 2000). Unlike maize, sorghum was domesticated only 3000-6000 years ago. Domestication occurred in sub-Saharan Africa, where its nearest wild relatives are indigenous, and sorghum has been an important grain crop there ever since. Early in the domestication of sorghum, compact inflorescences were selected for, which necessitated an increase in inflorescence branching, a decrease in internode length and larger seeds (Dillon et al. 2007). Soon after domestication, it dispersed to the Indian subcontinent as there are ancient sorghum specimens dating back to the second millennium BC (Zohary and Hopf 2000). Although sorghum is believed to have been domesticated in Africa, the precise
location is not known in part because archeological exploration is still in its infancy (Zohary and Hopf 2000). This relatively recent domestication timeline in conjunction with a selfing mating system has important implications for the detection of signatures of selection. The hypothesized recent bottleneck means sorghum has diverged little since the bottleneck and therefore has a genome that is still highly homogenized. Standard tests for selection usually fail due to a lack of statistical power (Hamblin et al. 2006). In other words, despite the fact that sorghum has wild progenitors, primitive landraces, and inbreds like maize, there may not be enough DNA polymorphism between lineages to tell when selection has occurred for putative domestication loci. For now, tests for selection in sorghum must be limited to large data sets until more sensitive tests are developed.

Despite these challenges some progress has recently been made in detecting signatures of selection. Recently, in a simple sequence repeat (SSR)-based genome-wide diversity scan of *Sorghum bicolor*, several candidate loci with patterns of variation consistent with directional selection were detected in cultivated lines (Casa et al. 2005). Unfortunately, there was not enough data to determine if the signal was due to selection or to demographic effects. Hoping to utilize the population structure and the hypothetically high genetic differentiation, which is typical of self-pollinating species, a follow-up study analyzed sequence data for a panel of cultivated and wild sorghum accessions looking for clearer evidence of selection (Casa et al. 2006). Across 10 segments along 99kb on chromosome 1 they found low levels of variation and extensive haplotype structure (Casa et al. 2006). Furthermore, two of four tests of selection employed were significant supporting the hypothesis for a selective sweep although demography still cannot be ruled out completely. Despite this problem, the study does lend further evidence of a selective sweep and identified potential candidate loci that can perhaps be tested functionally in future studies.

**Evolution of Grass Inflorescence Architecture Traits**

Domestication loci, shown to be important in the evolution of crops, may also have been important in the evolution of wild grasses. In particular, genes involved in inflorescence architecture are important as branching varies widely in grasses. Both maize and sorghum are economically important crop plants, which has contributed to their intense study over the last
Maize and sorghum belong to the grass tribe Andropogoneae, which in turn belongs to the family Poaceae (GPWG 2000; GPWG 2001). Members of Poaceae are known as the true grasses and are thought to have evolved 55-70 million years ago (Kellogg 2001; Bremer 2002). Since the family’s origin, grasses have diversified and they now cover a significant portion of Earth’s landmass. The family includes around 600 genera, encompassing 10,000 different species, most of which can be placed into two major clades, the BEP clade (Bambusoideae, Ehrhartioideae, and Pooideae) and the PACCAD clade (Panicoideae, Aristidoideae, Arundinoideae, Chloridoideae, Centothecoideae and Danthonioideae) where each clade represents a radiation 40-50 million years ago (Bremer 2002). Most of the phylogenetic work to understand the evolutionary relationships of Poaceae genera has been done by the Grass Phylogeny Working Group (GPWG 2000; GPWG 2001), a group of grass systematists who decided to combine their data and collaborate in order to generate a comprehensive phylogeny of this immense family (Kellogg 2001). Work done by the GPWG has been insightful by recovering and confirming some larger clades, but there is still some uncertainty due to a lack of resolution and low support for parts of the phylogeny, which has lead to recent modifications or additions. For example, recently it has been suggested that the subfamily Micrairoideae be included into the PACCAD clade and that the Centothecoideae should be integrated with the Panicoideae, thus making the current acronym PACMAD (Sanchez-Ken et al. 2007; Sanchez-Ken and Clark 2010).

The Andropogoneae and the Paniceae are the two major tribes of the subfamily Panicoideae. The Andropogoneae comprises 85 genera with about 1000 different species,
which is roughly one tenth of all grass spp. (Clayton and Renvoize 1986). Moreover, all members use the C_4 metabolic pathway and most members have paired spikelets in the inflorescence. Members of Andropogoneae span the world with members on all continents in temperate and tropical regions. Despite the fact that the members of Andropogoneae are spread all over the world, the tribe is strongly supported as being monophyletic (GPWG 2000; GPWG 2001) with *Arundinella* being sister to the tribe. Notwithstanding the strong support for monophyly of the tribe, there is incongruence within the tribe due to a lack of resolution. Reconstruction of the phylogenetic relationships within the Andropogoneae is complicated by the presence of short internal branches, that are indicative of a rapid radiation or the diversification of the grasses with few molecular changes along the internal edges of the tree and most occur along the terminal edges (Bremer 2002). Previously, phylogenies have been constructed using highly conserved chloroplast genes, but the addition of nuclear genes adds further support to certain nodes of the phylogeny, such as the monophyly of the Andropogoneae and the existence of informal clade called the “core Andropogoneae” (Lukens and Doebley 2001; Mathews et al. 2002), but it still leaves most relationships unresolved or unsupported. Additions of molecular data to the current dataset are greatly needed to help resolve most of the generic relationships within the tribe.

**Using Comparative Genomics to Study Gene Families**

As more plant genomes are sequenced, comparative genomics is being used as a tool to study the evolution of genomes and gene families. Recently, the evolution of a number of gene families have been investigated using comparative genomics including the WRKY, MADS-box genes, ERFs, Aux/IAA, and TCP gene families (Vandenbussche et al. 2003; Wu et al. 2005; Nakano et al. 2006; Navaud et al. 2007; Mukherjee, Brocchieri, and Burglin 2009; Shan et al. 2009; Kagale, Links, and Rozwadowski 2010). Most of these papers compare similarities and differences between eudicots and monocots since Arabidopsis and rice were among the first genome sequences available (Initiative 2000; Goff et al. 2002). Since then, many additional “higher” plant genomes have been sequenced, although with a bias towards crop genomes. However, rapid sequencing of “lower” plant genomes at key nodes of the plant phylogeny is in progress, which will facilitate the investigation of broader
questions, such as comparing mechanisms of genome evolution along different lineages, discovering the origins of gene families and timing key events in their evolution to see if they correlate with developmental innovations. The hypothesis generating nature of this kind of research would provide fodder for more focused investigations. Currently, whole genome sequencing for many lower plants, such as the moss *Physcomitrella patens*, has been completed or nearly completed (Rensing et al. 2008). Also new models, such as *Aquilegia formosa*, are being developed rapidly in order to add another point of reference for the investigation of evo-devo questions. *Aquilegia* is a lower eudicot and a phylogenetic midpoint between Arabidopsis and rice, and therefore is an important taxon needed to address questions concerning the evolution of developmental differences between eudicots and monocots (Kramer 2009). Although, historically, comparative genomics has not been widely used as a tool for evo-devo research, as more plant genomes are sequenced it will prove to be an invaluable tool to understanding the evolution of developmental traits within a larger context.

**RESEARCH GOALS**

Currently, there are few studies on the evolution of plant development in grass species. Given the interdisciplinary nature of these studies, they can potentially provide information important to the understanding of numerous topics in science. Of the few studies of grass developmental evolution, many have implicated domestication loci involved in the development of plant or inflorescence architecture (Doust 2007). The *ral* gene, as a member of the *ramosa* pathway, has been shown to be involved in the branching architecture of maize inflorescences (Vollbrecht et al. 2005). Previously, evidence of selection at the *ramosal* (*ral*) locus was found, but it was not known if the selection occurred during maize’s domestication from its wild progenitor or during the improvement process over the last several hundred years. Therefore, the first research goal was to determine when the selection occurred, as well as, to determine the developmental basis for the phenotype of weak mutant alleles of *ral* and make inferences about a possible adaptive, selected trait. The second objective, as an extension of the first, was to investigate the evolution of *ral* in grasses closely related to maize. Here we sought to determine if selection on *ral* has occurred and if
developmental modifications due to changes in gene regulation and/or protein structure are present that correlate with the natural variation in grass inflorescence architecture. Finally, we also characterized and studied the evolution of the EPF subfamily of C2H2 transcription factors in various plant genomes in order to correlate patterns of gene expansion, diversification, and lineage specific differences with the evolution of developmental life history traits in plants.
LITERATURE CITED


Goldschmidt, R. 1940. The material basis of evolution. Yale University Press.


FIGURE LEGEND

**Figure 1.** Morphological differences of inflorescences within *Zea* and its close outgroup Tripsacum. A) *Tripsacum dactyloides*, B) the teosinte *Zea mays* ssp. *mexicana*, C) a maize landrace, and D) a maize inbred. The figure illustrates the striking changes in ear morphology that occurred during the domestication of maize from its wild progenitor, teosinte.
Figure 1 Sigmon
CHAPTER 2. EVIDENCE OF SELECTION AT THE RAMOSAI LOCUS DURING MAIZE DOMESTICATION

A manuscript published in Molecular Ecology¹

Brandi Sigmon and Erik Vollbrecht²

ABSTRACT

Modern maize was domesticated from Zea mays parviglumis, a teosinte, about 9,000 years ago in Mexico. Genes thought to have been selected upon during the domestication of crops are commonly known as domestication loci. The ramosa1 (ra1) gene encodes a putative transcription factor that controls branching architecture in the maize tassel and ear. Previous work demonstrated reduced nucleotide diversity in a segment of the ra1 gene in a survey of modern maize inbreds, indicating that positive selection occurred at some point in time since maize diverged from its common ancestor with the sister species Tripsacum dactyloides and prompting the hypothesis that ra1 may be a domestication gene. To investigate this hypothesis, we examined ear phenotypes resulting from minor changes in ra1 activity and sampled nucleotide diversity of ra1 across the phylogenetic spectrum between tripsacum and maize, including a broad panel of teosintes and unimproved maize landraces. Weak mutant alleles of ra1 showed subtle effects in the ear, including crooked rows of kernels due to the occasional formation of extra spikelets, correlating a plausible, selected trait with subtle variations in gene activity. Nucleotide diversity was significantly reduced for maize landraces but not for teosintes, and statistical tests implied directional selection on ra1 consistent with the hypothesis that ra1 is a domestication locus. In maize landraces, a noncoding 3’ segment contained almost no genetic diversity and 5’ flanking diversity was greatly reduced, suggesting that a regulatory element may have been a target of selection.

¹ Reprinted with permission of Molecular Ecology
² To whom correspondence should be addressed: vollbrec@iastate.edu
INTRODUCTION

Agriculture began some 10,000 years ago, when prehistoric farmers in both the New and Old World began domesticating plants and animals through selection on desirable traits. In plants such characters included seed or grain quality, yield, and ease of harvest among numerous other traits (Smith 1995). The domestication process typically spans hundreds to thousands of years during which time preferred traits might be selected on singly or in combination with others. Due to continued selection after or in concert with initial domestication events, cultural and ecological adaptation results in locally adapted, diversified crops (Purugganan and Fuller 2009). For example, in many modern crops, accelerated selection of preferable traits has occurred within the last few hundred years, in a process known as crop improvement that is temporally well separated from domestication. Darwin recognized the artificial selection that occurred during the domestication of plants and animals as analogous to the process of natural selection that drives the evolution of species, and studies of domesticated species inform our understanding of the genetic basis of evolutionary diversification (Doebley, Gaut, and Smith 2006).

In the context of either natural or artificial forces, different modes of selection on phenotypes lead to different outcomes in allele frequencies. For instance, directional selection on a preferred phenotype as may occur in domestication and/or cultural adaptation leads to an increase in frequency of the corresponding allele(s), whereas balancing selection maintains multiple alleles. For genes that are associated with positively selected traits in domestication and improvement, the process may leave behind so-called signatures of selection in the form of distinctive patterns of reduced nucleotide diversity (Tanksley and McCouch 1997). When a wild progenitor and intermediate cultivars are extant and can be sampled, appropriate statistical tests (Sabeti et al. 2006) may identify and differentiate these signatures of selection in the genome, and pinpoint a selective event to a particular phase of crop evolution.

Molecular, archaeobotanical and paleoecological evidence suggest a single domestication event for maize, around 9000 years ago in Mexico (Matsuoka et al. 2002; Piperno et al. 2009). Maize was domesticated from its wild ancestor *Zea mays* ssp. *parviglumis*, (hereafter referred to as teosinte) (Doebly 2004). Maize and teosinte have
many similarities, but exhibit striking differences in plant architecture and ear morphology. Phenotype driven approaches such as mutant and QTL studies have shown that the molecular basis for some of these major phenotypic differences may be explained by selection on a few domestication loci of major effect. In particular, the *teosinte branched1* (*tb1*) gene explains most of the difference in plant architecture (Doebley, Stec, and Gustus 1995; Doebley, Stec, and Hubbard 1997; Clark et al. 2004; Clark et al. 2006). Other domestication and diversification loci may include genes involved in biochemical traits, such as those involved in the production of starch (Whitt *et al.* 2002) or those responsible for the carotenoid-rich yellow endosperm (Palaisa *et al.* 2004). These examples are among a few putative domestication loci in maize for which an accompanying, selected trait is clearly discernible.

Candidate domestication loci may also be identified by genotype driven approaches, such as strictly molecular, large-scale screens for genes with signatures of artificial selection (Casa *et al.* 2005; Chapman *et al.* 2008). Such studies in maize suggest that more than 1000 genes may have been selected on during the derivation of modern maize (Wright *et al.* 2005; Yamasaki *et al.* 2005). A disadvantage of this approach is that when candidate genes are defined by strictly molecular criteria, the corresponding gene functions are *a priori* unknown. Thus, while these approaches hold great promise for identifying genes that contribute to adaptive traits (Ross-Ibarra, Morrell, and Gaut 2007), relating such domestication loci to a selected trait is challenging. Notably, in maize many of these putative domestication loci identified by genomics are hypothesized to be expressed preferentially in the ear (Hufford *et al.* 2007), and to perhaps be related to selection on ear traits, because among the grasses the prolific maize ear is a unique and highly derived organ. Moreover, in the ear the basis of maize-teosinte phenotype differences is poorly understood. Thus, a modified, genotype driven approach in which a candidate gene(s) known to function in the morphological development of the maize ear is examined for evidence of selection may prove more fruitful.

*ramosal* is a mutant of maize that has been studied by geneticists for many years (Gernert 1912) but only recently come to be understood at the molecular level. The *ral* gene encodes a plant specific C2H2, EPF-subclass zinc finger transcription factor. It is expressed in a boundary domain near the base of particular meristems in the inflorescences (the ear and tassel), thereby regulating fate of the adjacent meristem (Vollbrecht *et al.* 2005). *ral* is a
component of a genetic pathway, termed the *ramosa* pathway, that imposes a spikelet pair or short branch identity as branch meristems are initiated during tassel and ear development (Vollbrecht et al. 2005; Bortiri et al. 2006; Satoh-Nagasawa et al. 2006). Morphologically, normal *ral* gene function results in the unbranched appearance of the ear and the upper part of the tassel, by causing both structures to produce short, determinate spikelet pair branches (Figure 1A, 1D). In plants that contain strong mutant *ral* alleles and therefore lack *ral* gene function, the ears and tassels have additional long branches, leading to a highly disorganized ear (Figure 1C, 1F). In contrast, in weak *ral* mutants the tassels contain just a few extra branches (Figure 1B, arrowhead) and ears have crooked rows but are otherwise relatively normal (Figure 1E). Crooked rows likely affect grain yield in maize and are common to other mutants, including *ramosa2* mutants where they form as a consequence of spikelet triplets being produced in place of spikelet pairs (Bortiri et al. 2006). Thus, it is reasonable to hypothesize that the phenotype of weak *ral* mutants has a similar developmental basis, but this has not been examined.

Existing molecular data concerning *ral* function beg the question of whether or not *ral* may have been a target of selection given the gene’s role in ear development. Previously, nucleotide diversity in a ~700 base pair segment of the *ral* gene in maize was examined in a panel of diverse, modern inbred lines and a signature of selection was detected relative to the neighboring genus *Tripsacum* (Vollbrecht et al. 2005). These data suggested an event of positive selection somewhere between the present and the point in time when maize and *tripsacum* diverged from their common ancestor several million years ago, with caveats. First, studies have shown that sampling a small region of a locus may lead to erroneous conclusions about positive selection (Yamasaki et al. 2008). Second, previous work could not distinguish during which time period the putative selection occurred: the evolution of teosinte, maize domestication, and/or modern maize improvement. To find evidence of selection at *ral* during the domestication process would not be altogether surprising, as it is intuitive that inflorescence development genes may have been subjected to artificial selection because their role in organizing inflorescence architecture may impact grain yields. We hypothesize that if slight changes in *ral* function alter the packing of kernels into straight rows on the ear then particular *ral* alleles may have been selected during maize
domestication and/or improvement for increasing grain yields. To address this question, we first determined the developmental basis of the crooked row morphology in the ears of weak \textit{ra1} mutants to better understand this putative, selected trait. We also examined the nucleotide diversity around the \textit{ra1} locus in a diverse panel of unimproved maize landraces and throughout the teosintes. Our statistical and phylogenetic analyses provide compelling evidence that positive selection on \textit{ra1} occurred during the domestication of maize from teosinte, and we discuss the implications of \textit{ra1} as a candidate domestication locus.

**MATERIALS AND METHODS**

**Sampling Strategy and Plant Materials**

To examine the developmental basis of crooked rows in weak \textit{ra1} mutants, the recessive \textit{ra1-63.3359} weak mutant allele (Vollbrecht et al. 2005) was obtained from the Maize Genetics Cooperation Stock Center and crossed six times to the inbreds Mo17 and B73. Similarly, the recessive \textit{ra1-RS} weak mutant allele, obtained from Robert Schmidt, UC-San Diego, was crossed four to six times to the inbreds. Field-grown mutant (crossed six times to inbreds) and normal plants from our summer nursery in Ames, Iowa were sampled at various developmental stages. Scanning electron microscopy (SEM) was used as previously described (Vollbrecht et al. 2005) to obtain micrographs of younger ears from greenhouse-grown weak \textit{ra1-RS} mutants (crossed four times to B73).

For nucleotide diversity analysis, accessions of maize and teosinte were chosen to optimize geographic distribution and limit bias toward North American varieties. Accessions of more distant teosintes including \textit{Zea mays} ssp. \textit{mexicana}, \textit{Zea mays} ssp. \textit{huehuetenangensis}, \textit{Zea perennis}, \textit{Zea diploperennis}, \textit{Zea luxurians} and \textit{Tripsacum dactyloides}, which were sampled for either phylogenetic or outgroup purposes, were provided by the USDA Agricultural Research Station, Iowa State University, Ames, IA.

**PCR and Sequencing**

Regions of \textit{ra1} were amplified from genomic DNA by PCR using Ex Taq polymerase (TaKaRa). Primers were designed to amplify a region of approximately 2400 base pairs centered on the approximately 560 base pair \textit{ra1} coding region. Primers were designed from
known maize sequences at GenBank and the PCR was used over 35 cycles, with appropriate annealing temperatures and standard PCR conditions. Amplicons were then purified using QIAquick PCR Purification Kit (Qiagen, Valencia, CA) and cloned into pCR 2.1-TOPO (Invitrogen) to separate haplotypes in heterozygous samples. Eight clones from each PCR reaction were then sequenced to control for PCR errors. Sequence data were collected from both strands to minimize ambiguities. Sequences were assembled and edited using the PREGAP and GAP4 software from the Staden package (Staden 1996).

Sequence Analysis

Sequences were initially aligned using ClustalW (Larkin et al. 2007) as part of the MacVector Version 8.1 software (http://www.macvector.com) and further adjusted by hand. All diversity statistics were estimated using DnaSP, Version 3.51 (Rozas and Rozas 1999). To test for selection, the Hudson-Kreitman-Aguade (HKA) test (Hudson, Kreitman, and Aguade 1987) was employed using the three neutral loci *adh1*, *bz2* and *glb1* (Eyre-Walker et al. 1998; Hilton and Gaut 1998; Tenaillon et al. 2001). *P* values from multiple tests were combined as previously described (Whitt et al. 2002). Maximum likelihood HKA (MLHKA) tests (Wright and Charlesworth 2004) were conducted using *Zea luxurians* as the outgroup with the previously mentioned three neutral loci. Three separate runs starting from different random seeds were performed, each using a Markov chain length of $1 \times 10^6$ simulations. The tests were conducted to compare the fit of *ra1* in a population to a neutral versus a selected model. A hypothesis of selection at domestication is considered supported if the landrace likelihoods differ significantly, but the teosinte likelihoods do not. Significance was assessed using the likelihood ratio test.

Tajima’s *D* (Tajima 1989), Fu and Li’s *D* (Fu and Li 1993) and the minimum number of recombination events were estimated using DnaSP, Version 3.51 (Rozas and Rozas 1999). The test statistics Tajima’s *D* and Fu and Li’s *D* were conducted on total sites and significance was assessed using critical values. The minimum number of recombination events ($R_m$) was obtained using the four-gamete test (Hudson and Kaplan 1985). Fay and Wu’s *H* statistic (Fay and Wu 2000) was calculated and significance assessed using a publicly available web interface (http://www.genetics.wustl.edu/jflab/htest.html). The *H*
statistic determines the level of high frequency variants (ancestral polymorphisms) to detect hitchhiking. Recombination was examined with the pairwise module in LDhat version 2.1 (http://www.stats.ox.ac.uk/~mcvean/LDhat/).

Phylogenetic Analysis

Phylogenies were reconstructed using PAUP* Version 4.0b10 (Swofford 2003) and MrBayes Version 3.1.2 (Ronquist and Huelsenbeck 2003). Bootstrap support was assessed utilizing a full heuristic search with 1000 bootstrap replicates. Phylogenies made using MrBayes were reconstructed using a Markov chain Monte Carlo (MCMC) algorithm under a general time reversal (GTR) model of evolution. Parameters and priors were set to account for nucleotide frequencies, substitution rates as well as transition/transversion ratios. The branch length prior was set to be uniform and unconstrained. All other parameters and priors were set at default. Two independent runs of four chains each (one cold and three heated) were run for 1,000,000 generations and the first 25,000 trees were removed. All phylogenies were viewed using the software TreeView X Version 0.4.1 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

RESULTS

Ear Morphology in Weak ra1 Mutants

Maize ears and tassels (inflorescences) develop similarly, by producing a series of different meristem types on the main inflorescence axis (Vollbrecht and Schmidt 2009). The primary inflorescence meristem, located at the growing tip of the inflorescence, produces the axis and initiates second order meristems on its flanks (Figure 2A, black box). A few second order meristems at the base of the tassel produce long branches; the rest of the second order meristems in the tassel and all of the second order meristems in the ear, are determinate in that they quickly cease growing. Thus, most second order meristems produce short, spikelet pair branches that in turn bear third and fourth order meristems. The strong ra1 mutant phenotype is due to indeterminacy of spikelet pair meristems, resulting in an outgrowth of long branches that leads to a highly disorganized ear (Figure 1F) (Vollbrecht et al. 2005).
Ears of weak mutants, such as plants homozygous for the *ra1-63.3359* or *ra1-RS* allele, typically have crooked rows (Figure 1E), and occasionally produce long branches at the base of the ear in some genetic backgrounds. The developmental basis of this phenotype has not been described. The *ra1-63.3359* allele, which arose spontaneously, has a four base pair insertion into the stop codon that presumably results in the addition of 17 amino acids to the carboxyl terminus of the RA1 protein (Vollbrecht et al. 2005). On the other hand, the *ra1-RS* lesion is predicted to eliminate nine amino acids from the amino terminus of the RA1 protein and kernel rows are crooked on ears from *ra1-RS* mutants (Vollbrecht et al. 2005), slightly more so than from *ra1-63.3359* plants. The *ra1-63.3359* and *ra1-RS* mutant phenotypes vary in different genetic backgrounds, but are relatively weak in the Mo17 inbred background (R. Weeks and E. Vollbrecht, unpublished). After introgressing these alleles into inbred backgrounds we examined ears at various stages of development (Figure 2). At very early stages (length = 0.5 cm, Figure 2A-2B), *ra1-RS* ears showed normal behavior of the primary inflorescence meristem. Initiation of second order meristems was perhaps slightly delayed but otherwise normal. However, second order meristems became abnormally elongated due to a delay in the initiation and differentiation of third order meristems (Figure 2D) when compared to wild type (Figure 2C). This delay in the initiation of third order meristems resulted in the beginnings of row disorganization in the ear. At this developmental stage spikelets still appeared strictly as pairs, as in normal ear development (Vollbrecht and Schmidt 2009), but these positions eventually produced spikelets in groups of three (*i.e.*, triplets) in the mutants (Figure 2E). For example, of developing top ears from twelve *ra1-63.3359* mutants in the weaker Mo17 background (length = 0.8 cm to 2.1 cm), five ears had in addition to many paired spikelets an average of ten spikelet triplets per ear (Figure 2E-2G), and seven ears had no spikelet triplets (Table 1). When the *ra1-63.3359* mutant was examined in the B73 background, spikelet triplets were present in all twelve ears, and with greater frequency (Table 1). For *ra1-RS*, the same pattern was observed where more spikelet triplets were present in the B73 versus the Mo17 genetic background. In contrast, twelve ears of each control (normal inbred) contained no spikelet triplets. In each inbred background *ra1-RS* mutants consistently produced more spikelet triplets, and had relatively more disturbed kernel rows at maturity, than *ra1-63.3359*, indicating that *ra1-RS* is a stronger mutant allele.
Thus, the delayed initiation of third order meristems was sometimes accompanied by initiation of an extra (third) spikelet. The presence of spikelet triplets in weak *ra1* mutants and their absence in wild type ears strongly suggests these extra spikelets create disorganized rows. These data imply weak loss of function alleles of *ra1* condition crooked rows, supporting the hypothesis that natural variation in the *ramosa* pathway could also manifest as altered kernel rowing, and serve as the basis for a visibly selectable trait during domestication.

**Nucleotide Diversity of the *ra1* Locus**

To characterize nucleotide diversity at *ra1* within the genus *Zea*, we isolated and sequenced alleles of the *ra1* gene from 43 different plant accessions including 22 maize landraces, 11 *Z. m. parviglumis*, five *Z. m. mexicana*, and one each of *Z. m. huehuetenangensis*, *Z. perennis*, *Z. diploperennis*, *Z. luxurians*, and *T. dactyloides* (Table 2). Accessions were chosen across the geographic range of maize and teosinte focusing on Central and South America in order for the sampling to be representative of both teosinte and maize landrace diversity. Sequence analysis of the 2.4 kilobase amplicon, which includes almost 1 kilobase each of immediate 5’ and 3’ noncoding sequence in addition to the *ra1* coding sequence, identified 18 distinct haplotypes within the landrace sampling and 18 within the *Z. m. parviglumis* (teosinte) sampling. For landraces, two haplotypes were represented multiple times within the sampling comprising 24% and 12% of the alleles, whereas all other alleles were unique. For the teosinte population, all haplotypes were unique. There were no shared haplotypes between maize landraces and teosintes. Nucleotide polymorphism (θ) (Watterson 1975) and nucleotide diversity (π) (Nei 1987) were estimated (Table 3). To visualize variation in polymorphism throughout the 2.4 kilobase *ra1* region, a sliding window analysis of π was performed (Figure 3).

In the maize landraces, nucleotide diversity is unusually low across the entire region, within the range of $0 \leq \pi < 0.005$ as expected for domestication loci (Hufford et al. 2007), and differs significantly from the expected values of $\pi_{\text{average}} = 0.0087$ for neutral genes (t-test, $P << 0.01$). Within the known functional motifs including the zinc finger DNA-binding domain, a post-zinc finger, putative EAR repression motif (B. Sigmon and E. Vollbrecht,
unpublished), and the previously identified terminal EAR motif, the landraces have virtually no nucleotide diversity. When the nucleotide diversity analysis is partitioned into 5’, coding, and 3’ components separately, the immediate 3’ sequence has the lowest nucleotide diversity. The $\pi$ value for this region is 0.00065, approximately four-fold lower than values for both the coding and immediate 5’ regions of the gene. This $\pi$ value is among some of the lowest values found to date in surveys of maize landraces (Yamasaki et al. 2005; Hufford et al. 2007; Yamasaki et al. 2008).

In contrast, $ra1$ nucleotide diversity in the teosintes is 0.013 across the entire region and does not significantly vary among 5’, coding, and 3’ regions of the gene. These levels of diversity do not differ significantly from that expected for the average teosinte gene of $\pi = 0.012$ (Hufford et al. 2007), (t-test, $P > 0.5$). There is higher diversity in known functional regions of the gene (zinc finger and two EAR motifs) compared to the landraces, but the changes are all synonymous. In summary, these results of lower than average nucleotide diversity in maize landraces and average diversity in teosinte indicate that genetic diversity at $ra1$ was significantly reduced when the landraces were derived from teosinte during the initial domestication process.

Similarly, patterns of retention of nucleotide diversity at $ra1$ also match expectations for a domestication locus. Low gene diversity in maize landraces is consistent with selection, but diversity can also vary due to differential functional constraints on sequence evolution. The ratio of landrace nucleotide diversity to teosinte diversity ($\pi_{lr}/\pi_{teo}$) has been used to measure retention of genetic diversity correcting for functional constraints (Clark et al. 2004). For neutral genes a range of 60-80% retention is expected whereas a lower ratio is indicative of selection (Zhang et al. 2002; Hufford et al. 2007). For $ra1$, the low diversity 3’ noncoding region in the landraces retained only 5% of the diversity found in the teosinte panel, but the 5’ noncoding and coding region retained approximately 20% and 17% respectively. These low levels of genetic retention reinforce the notion that much of the reduction in diversity, especially for the 3’ region, occurred during the domestication of maize from its progenitor teosinte.
Tests of Selection

Although a number of selection tests are commonly used, the HKA test (Hudson, Kreitman, and Aguade 1987) is among the most widely used for maize candidate domestication loci due to its high statistical power to detect positive selection by combining information from comparative and population genetic data (Zhai, Nielsen, and Slatkin 2009). The HKA selection test is a stringent test for departure from neutrality that examines whether selection has significantly altered diversity at a locus relative to changes in diversity of neutral, control loci. We applied this test to our $ra1$ data using $adh1$, $bz2$, and $glb1$ as the neutral genes (Tenaillon et al. 2001; Tiffin and Gaut 2001) and both Tripsacum dactyloides and Zea luxurians as outgroups (Table 4). Tripsacum is preferable as an outgroup for tests of selection because it is in the sister genus to Zea and does not naturally interbreed with maize, but if tripsacum sequence can not be recovered for a portion of a gene then a distant outgroup within genus Zea may be utilized (Hanson et al. 1996; Hilton and Gaut 1998; Clark et al. 2004). Thus, for some tests we used Zea luxurians as an outgroup, with controls as described below. Zea luxurians is indigenous to Southern Guatemala and therefore geographically isolated from most other teosintes, and there has been only minimal, historical gene flow between Zea luxurians and maize (Fukunaga et al. 2005; Ross-Ibarra, Tenaillon, and Gaut 2009). Initial HKA tests using tripsacum as the outgroup (Table 4) and a ~1.3 kilobase fragment that contains the $ra1$ coding sequence plus some 5’ and flanking sequence and a small portion of the 3’-UTR, showed a departure from neutrality for maize landraces. The departure was due to reduced diversity at $ramosa1$. By contrast, HKA tests performed on the same fragment for teosintes were not significant. These results strongly support the hypothesis that the $ra1$ region experienced positive directional selection specifically during domestication.

In order to investigate whether or not a particular region of $ra1$ may have been the target of selection, pairwise HKA tests were performed individually on the 5’ noncoding, coding, and 3’ noncoding sequence. Because we did not recover much 3’ noncoding sequence from tripsacum despite attempts using a variety of approaches, and to enable analyses of more extensive flanking sequence in the tests, Zea luxurians was used as the outgroup. The full 2.4 kilobase region (see Figure 3) was easily recovered from Zea
As a control we first used Zea luxurians as the outgroup in HKA tests for the same 1.3 kilobase region previously analyzed relative to tripsacum and compared the results with the two different outgroups. As for tests relative to tripsacum, control tests relative to Zea luxurians were similarly significant for the landraces but not for teosinte. Thus, results did not differ according to outgroup, validating the use of Zea luxurians to analyze the more extensive region. For the larger, 2.4 kilobase region, HKA tests for maize landraces were significant when the whole region was tested, or when either 5’ or 3’ noncoding sequence were tested, but not when coding sequence was analyzed alone (Table 4). All of the analogous HKA tests for teosinte did not differ significantly from neutrality (Table 4), indicating no evidence for selection on ra1 in Z. m. parviglumis since it diverged from its most recent shared ancestor with Z. luxurians. When two more divergent teosinte haplotypes, evident from the phylogenetic analysis (Figure 3), were removed from the HKA analysis the tests were still not significant (data not shown). Therefore, even the remaining teosinte haplotypes, though more closely related to landrace haplotypes, are divergent enough to reveal selection in the landraces. These results imply that nucleotide diversity decreased in derivation of the landraces, i.e., during domestication. Moreover, the extremely low nucleotide diversity of the 3’ region and lack of a detectable signature in the coding region, together suggest a target of selection may be located in a regulatory region, perhaps in the 3’ direction of the ra1 coding sequence.

In addition to pairwise HKA tests we used a maximum likelihood HKA (MLHKA) test that also compares polymorphism within species and divergence between species, but allows for an explicit test of selection at a locus using multilocus data. Departure from neutrality is assessed by the likelihood ratio test of a locus compared to neutral reference genes (Wright and Charlesworth 2004). MLHKA tests also provided strong evidence of selection in the landraces for both 5’ and 3’ noncoding regions of ra1, but not the coding sequence, and no evidence of selection on the ra1 locus in teosinte (Table 4).

The maximum likelihood estimate of the selection parameter ($k$) measures the degree of reduction caused by selection (Wright and Charlesworth 2004). Neutral genes are expected to have $k$ values equal to one whereas those under strong selection are expected to have values of $k > 2$ or $k < 0.5$ (Moeller and Tiffin 2005). For ra1, the landraces had $k$ values
less than 0.5, but the 5’ and 3’ noncoding regions had significantly lower values ($k = 0.1$ and $0.06$) than the coding region ($k = 0.42$), indicating an extreme loss of diversity in these regions (Table 4). In contrast, under the model that assumes ra1 is under selection for all regions in teosinte, no regions have values of $k < 0.5$ (Table 4), indicating the region does not have a significant reduction in diversity in teosinte. Compared to other putatively selected maize genes, $k$ values for the 5’ and 3’ noncoding regions of ra1 are comparable to those for other regions in the maize genome that exhibit evidence of a selective sweep (Moeller and Tiffin 2005; Wright et al. 2005; Camus-Kulandaivelu et al. 2008; Tian, Stevens, and Buckler 2009). These results suggest the nucleotide diversity, especially in noncoding regions of ra1, has been greatly reduced by selection.

In addition to performing HKA tests, we assessed non-neutral evolution by calculating Tajima’s $D$, Fu and Li’s $D$, and Fay and Wu’s $H$ test statistics. A significantly negative Tajima’s $D$ test statistic indicates an excess of low frequency of polymorphism, which is consistent with directional selection or population expansion (Tajima 1989). Similarly, a significantly negative Fu and Li’s $D$ also indicates directional selection although this test statistic is based on the number of singletons in a sample (Fu and Li 1993). For ra1, Tajima’s $D$ was negative for both landraces and teosinte, however, only the value for the 3’ region was significant. Fu and Li’s $D$ was also negative for both populations but only significant for the 3’ region in maize landraces. Fay and Wu’s $H$ statistic (Fay and Wu 2000) was calculated for the whole ra1 2.4 kb sequence and for partitions of the gene, in both the teosinte and maize landrace populations. The $H$ statistic detects the prevalence of high frequency variants in order to detect hitchhiking. For teosintes all of the calculated $H$ statistics were similar in value, and none were significant (Table 4). Similarly, in the landraces none of the $H$ statistic values were significant enough to suggest evidence of hitchhiking, although the value for the 3’ region in the landraces was the lowest (Table 4). Since this region only has four low frequency variants (singletons) and no high frequency variants (ancestral polymorphisms), there may not be enough sequence polymorphism in this partition of the gene to detect hitchhiking. Alternatively, a significant Tajima’s $D$, which detects low frequency variants, and a non-significant value for the $H$ statistic may indicate a
region recovering from a recent bottleneck where all the ancestral polymorphisms were removed from the population (Fay and Wu 2000).

These test statistics have low power to detect selection especially if few segregating sites are considered, as is the case for $ra1$ ($S_{\text{maize}} = 23, S_{\text{teosinte}} = 110$); demographic issues also influence results. Thus, Tajima’s $D$, Fu and Li’s $D$, and Fay and Wu’s $H$ test statistics often cannot stand alone, to definitively determine if a signature of selection is indeed due to positive directional selection or due to population bottlenecks or expansions (Sabeti et al. 2006). However, when combined with the HKA tests for selection our analyses of Tajima’s $D$, Fu and Li’s $D$, and Fay and Wu’s $H$ support the inference that the 3’ region of $ra1$ was targeted by selection in derivation of maize landraces.

**Phylogenetic Analysis**

For a domestication locus, if a single preferred allele becomes fixed at domestication and there is little subsequent change, then in phylogenetic analysis, the extant maize haplotypes are hypothesized to be confined to a single clade (Wang et al. 1999; Clark et al. 2004). For neutral loci on the other hand, maize alleles are expected to scatter among teosinte alleles resulting in a more dispersed tree topology (Goloubinoff, Paabo, and Wilson 1993; Hanson et al. 1996; Hilton and Gaut 1998). Thus, although homoplasy, recombination, and population structure considerations may hinder accurate reconstruction of the relationships between teosinte and maize haplotypes, phylogenetic analysis may provide patterns consistent with a hypothesis of selection at domestication. Given the differences we observed in patterns of diversity at $ra1$ and the precedent of intralocus variability at the domestication locus $tb1$ (Clark et al. 2004), we separated our $ra1$ data set to assay for differing phylogenetic signals within the data, by reconstructing phylogenies in three separate analyses, of 5’ noncoding sequence, coding sequence, and 3’ noncoding sequence.

The 5’ noncoding sequence data set contained the most haplotype diversity consisting of 43 different haplotypes over 1271 nucleotide sites of which 71 were parsimony-informative. The Bayesian tree topology (Figure 4A) shows all landrace and most teosinte haplotypes to be monophyletic with 100% posterior probability. These results illustrate the close relationship between maize and the subspecies $Z. m. parviglumis$ and $Z. m. mexicana$.
due to their recent divergence. Within this clade, a secondary clade of four landrace haplotypes cluster with a *Z. m. mexicana* haplotype, which suggests these landrace haplotypes may be more closely related to extant *Z. m. mexicana* than to *Z. m. parviglumis*. In addition to the more distant members of Zea (ssp. *luxurians*, ssp. *perennis*, ssp. *diploperennis*, and ssp. *mays huehuetenangensis*), four *Z. m. parviglumis* and two *Z. m. mexicana* haplotypes were basal to the main clade. Two of these *Z. m. parviglumis* haplotypes clustered with *Z. m. huehuetenangensis*, which may attest to the age of some *Z. m. parviglumis* allelic lineages (Hilton and Gaut 1998).

In contrast, the coding sequence data set consisted of only 29 differing haplotypes over 561 nucleotide sites of which 16 were parsimony-informative. The 5’ noncoding and coding sequence trees (Figure 4B) are similar in that most landrace and teosinte haplotypes are monophyletic and the secondary clade of landrace and *Z. m. mexicana* haplotypes is present in both. Also, the same two divergent *Z. m. parviglumis* haplotypes in the 5’ tree fall to the base of the coding sequence tree with one clustering with *Z. m. huehuetenangensis*. However, support for this tree topology is low compared to that of the 5’ sequence tree, thus these results should be interpreted with caution.

The 3’ noncoding sequence data set contained the least haplotype diversity with only 19 haplotypes over 577 total nucleotide sites of which 23 were parsimony-informative. The topology of this tree is somewhat different from that of both the 5’ noncoding and coding sequence trees due to the reduction of haplotype diversity in this region (Figure 4C). This reduction of haplotype diversity can partially be attributed to the absence of the secondary clade of landrace and *Z. m. mexicana* haplotypes. Interestingly, this region also exhibits a reduction of haplotype diversity for all the teosintes, but also exhibits basal placement of the same two *Z. m parviglumis* haplotypes. The extreme sequence conservation and low haplotype diversity of this region for both landraces and teosintes suggest it may have an essential role in *ral* function and thus be under some additional functional constraint. However, functional analyses in maize have not detected such a function for this region at this time.

Phylogenetic analyses indicated that both the 5’ noncoding and the coding sequence have more haplotype diversity than the 3’ noncoding tree. Despite differing levels of
polymorphism, for both regions the maize alleles fall predominantly into one clade with a few forming a second clade; this topology is more neutral than strictly expected for a domestication locus. However, for the 3’ noncoding region, one haplotype predominates in both the maize landrace and teosinte samples. This result is common among probable plant domestication loci, where the hypothesized selected allele is found frequently in the wild progenitor population (Purugganan, Boyles, and Suddith 2000; Nesbitt and Tanksley 2002; Clark et al. 2004).

**Recombination and Linkage Disequilibrium**

Recombination rates across a region can vary significantly and thus have variable roles in generating haplotype diversity. In maize, the minimum number of recombination events averages about 2.1 in genes and in teosinte the average is slightly higher at 2.7 (Ross-Ibarra, Tenaillon, and Gaut 2009). In a previous study, no evidence of recombination was found at the *ral* locus in a diverse inbred population (Vollbrecht et al. 2005). For maize landraces we found evidence of a minimum number of two possible recombination events (R_m) using the four-gamete test (Hudson and Kaplan 1985) (Table 3). Eight possible recombination events were estimated for the teosinte population. For both populations the recombination events are putatively located in the 5’ noncoding regions. However, accurate estimation and comparison of the recombination rate at the *ral* locus and/or in partitions of it was precluded by the low level of polymorphism (only 10 informative SNPs) in the landrace populations. Similarly, linkage disequilibrium (LD), an estimate of the correlation between different polymorphisms due to shared mutations and recombination histories, was estimated for teosintes only. The analysis suggested two separate LD blocks, located in the 5’ and 3’ noncoding regions (data not shown). Using a cutoff value of 0.1 (Flint-Garcia, Thornsberry, and Buckler 2003; Palaisa et al. 2004) for r^2 (Hudson and Kaplan 1985), LD decays at ~500 bp in this region for the teosintes. This estimate corresponds with the observation that in both maize and teosinte, LD tends to decay within genes (Tenaillon et al. 2001; Flint-Garcia, Thornsberry, and Buckler 2003; Clark et al. 2004; Weber et al. 2007).
DISCUSSION

Identifying the Target of Selection

Both the 5’ and 3’ noncoding sequences of ra1 show evidence of selection in the form of reduced nucleotide diversity, suggesting a target of selection lies outside the ra1 coding sequence. The 3’ region of the gene, which includes the 3’-UTR and downstream sequences, has some of the lowest nucleotide diversity found in maize and the lowest $k$ values from our MLHKA analysis, which would be consistent with selection occurring in or near it. However, from these analyses we cannot exclude that a target of selection lies in the 5’ region or even outside the analyzed region. Extensive sequence and expression analysis provides no evidence of alternative splicing for ra1 (E. Vollbrecht, unpublished), so the 3’ region does not contain coding sequences. BLAST searches show this 3’ sequence to be unique in the maize genome. Hence, while genetic studies to date have not revealed a function for this unique genomic sequence, if the conserved region harbors a target of selection then the target may be a cis-regulatory element, or involved in mRNA metabolism. In such a case, altering this sequence would be predicted to have some consequence on gene expression and potentially on phenotype in domesticated maize. Ongoing molecular mutageneses of ra1, for example using transposon (Ahern et al. 2009), TILLING (Till et al. 2004) or related chemical approaches, should prove useful in querying the functional significance of this highly conserved sequence.

A selected locus may be contained within a selective sweep, wherein DNA that neighbors a target of positive selection also contains reduced nucleotide diversity due to hitchhiking. Selective sweeps in maize can be limited or quite extensive. For example, the unidirectional selective sweep around the teosinte branched1 (tb1) locus extends 60-90 kilobases in the 5’ direction of the gene (Clark et al. 2004) whereas an asymmetric sweep at the yellow1 (y1) locus includes a much larger region of up to 600 kilobases (Palaisa et al. 2004). In the case of y1, the selective sweep is hypothesized to be so extensive because the locus underwent strong selection more recently than did tb1. Recently, a selective sweep in maize spanning 1100 kilobases and more than 15 genes was identified, however the target of selection in this region is unknown (Tian, Stevens, and Buckler 2009). The extent of a selective sweep can be measured by sampling nucleotide diversity of low-copy genomic
regions nearby and by estimating LD in the region of selection. In the *ra1* region analyzed here, levels of polymorphism are too low for comparative analysis of LD or recombination. However, regional patterns of nucleotide diversity will soon be accessible on a genome-wide scale using genome resequencing techniques and it will be interesting to see patterns of nucleotide diversity and LD in the region surrounding *ra1*. Preliminary analysis indicates *ra1* is imbedded in a region of overall low diversity, although such regions are not uncommon in the maize genome (E.S. Buckler, pers. communication). The *ra1* gene is near the centromere so a selective sweep around it may be physically quite extensive due to low rates of recombination per kilobase (Fengler *et al.* 2007), aside from the strong LD that may accompany positive selection. Thus, an understanding of genomic patterns of nucleotide diversity around *ra1* may elucidate the boundaries of the selective sweep and therefore the strength and timing of selection at the locus (Olsen *et al.* 2006).

### *ra1* as a Candidate Domestication Locus

Several lines of evidence suggest the *ra1* region was a target of selection during maize domestication. This hypothesis is chiefly supported by the low nucleotide diversity present at the *ra1* locus and the significant HKA and MLHKA tests for maize landraces, but not teosintes. Compared to nucleotide diversity levels in teosinte, the reduction in *ra1* diversity for maize landraces is much greater than expected from the population bottleneck occurring at domestication (Tenaillon *et al.* 2004). Previously, nucleotide diversity of *ra1* was reported for a diverse panel of modern inbred lines (Vollbrecht *et al.* 2005). In those data, the inbreds retained approximately 52% of the diversity found in maize landraces. In the present study the landraces retained only 5-20% of the teosinte diversity, compared to an expected 60-80% for neutral genes (Zhang *et al.* 2002; Hufford *et al.* 2007). These data suggest most of the reduction in *ra1* genetic diversity is due to selection during domestication from teosinte, with some further reduction following an improvement bottleneck, conclusions that are also consistent with the $D$ and the $H$ test statistics. Significantly low values for $D$ are consistent with recovery from a recent bottleneck and the lack of high frequency variants estimated by $H$ suggests that a strong bottleneck has removed ancestral polymorphisms from the landrace population (Tajima 1989; Fay and Wu 2000). Therefore, it is unlikely that
significant amounts of selection occurred during the evolution of teosinte or during the improvement process.

To date, molecular genetic approaches in plants have identified on the order of ten genes as domestication loci while another two dozen or so may be classified as post-domestication, crop-diversification genes; almost all of the domestication genes encode transcriptional regulators while roughly half of the diversification genes encode structural genes like enzymes (Purugganan and Fuller 2009). Thus, a transcription factor like ramosa1 that controls branching architecture of inflorescences is intuitively a good selection candidate. The weakest known mutant alleles of the ral gene were analyzed here, and shown to result in disordered rows on the maize ear. Therefore, we speculate that prehistoric farmers may have selected for straight rows on the ear for purposes of aesthetics and/or effects on grain yield, that in doing so they selected particular, relatively high-activity alleles of ral, and that this artificial selection resulted in reduced genetic diversity for the ral locus.

Interestingly, in a study of present-day maize farmers in central Mexico, row straightness was ranked as a desirable or necessary criterion when selecting landrace seeds for propagation (Perales et al. 2003). As expected given the great genetic diversity of teosinte, we documented a large variety of ral alleles within the population that could exhibit varying effects on inflorescence branching and therefore, row formation. The function of ral in teosinte has not been studied. Since teosinte ears produce two ranks of solitary spikelets (Sundberg and Orr 1990), a trait such as crooked rows would have been inconsequential in teosinte, but may have become important as ear diameter and number of rows of spikelet pairs per ear increased during the domestication process. In an association mapping study in teosinte, markers for the zea apetala homolog1 (zap1) gene showed a significant association with inflorescence branch number but accounted for only 2.7% of the phenotypic variance (Weber et al. 2007). Since the frequency of the maize-like allele was found to be at a higher frequency in landraces than in teosintes, zap1 may have been selected upon during the domestication process (Weber et al. 2007). The additive nature of putative domestication loci involving inflorescence architecture traits, like zap1, reinforces the notion that many genes responsible for the phenotypic variance of these traits in both teosinte and maize remain undiscovered. In any case, the absence of evidence for selection on ral in teosinte suggests
that its coupled trait was not subject to natural selection during the evolution of teosinte and its single-rowed inflorescence, but was subject to it as prehistoric farmers began domesticating teosinte and increasing row number, such that straight rows became important. This hypothesis could be tested by observing the phenotypic consequences in the ear for introgression into maize of various teosinte alleles of \textit{ra1}, by themselves and in complementation tests with mutant maize alleles. Artificial selection on standing variation within the progenitor population is a feature common for plant domestication loci. Other examples include \textit{tb1} in maize (Clark et al. 2004), \textit{fw2.2} in tomato (Nesbitt and Tanksley 2002), and \textit{BoCal} in cauliflower and broccoli (Purugganan, Boyles, and Suddith 2000). To date \textit{tga1} in maize is the only domestication locus where the cultivated allele was not found in the wild progenitor population (Dorweiler et al. 1993; Dorweiler and Doebley 1997; Wang et al. 2005).

Our finding that two \textit{ra1} alleles were prevalent within the landrace population at 24% and 12% frequency suggests that more than one allele could have made it through the domestication bottleneck. Phylogenetics could hypothetically be used to address that question, but phylogenetic reconstruction for domestication loci is problematic due both to the fact that low nucleotide diversity regions must be used to build trees, and to past occurrences of introgression and hybridization events between maize and teosinte populations, which complicate accurate reconstruction.

\textbf{Evolution of the \textit{ramosa} Pathway and Maize Inflorescence Morphology}

\textit{ra1} regulates meristem function as part of a molecular genetic pathway that also includes \textit{ramosa2}, which encodes a LOB domain transcription factor (Bortiri \textit{et al.} 2006), and \textit{ramosa3}, which encodes a trehalose-6-phosphate phosphatase (Satoh-Nagasawa \textit{et al.} 2006). Thus, both \textit{ra1} and \textit{ra2} likely function as regulatory transcription factors in this pathway, while the biochemical role of \textit{ra3} is less clear at this point in time. It has been shown that some genes identified as developmental regulators in maize do affect the natural variation of complex traits in extant teosinte, although the \textit{ramosa} genes were not among the regulators tested, and that this variation may serve as a basis for directional selection during domestication (Weber \textit{et al.} 2007). As originally pointed out by Darwin, artificial selection
during domestication has many parallels with natural selection during evolution. Given that the *ramosa* genes regulate branching architecture in maize, it is tempting to speculate that expression differences among these genes or functional differences among their gene products may contribute to branching architecture variation in grass inflorescences (Vollbrecht and Sigmon 2005; McSteen 2006). Both *ra2* and *ra3* are conserved by purifying selection in the grasses (Bortiri *et al.* 2006; Satoh-Nagasawa *et al.* 2006), but putative *ra1* orthologs have only been identified from Panicoid grass species (B. Sigmon, E. Vollbrecht and E. Kellogg, unpublished data), which includes the cereal crops maize, sorghum and foxtail millet, and the Andropogoneae tribe as a subgroup that includes maize, sorghum and others (GPWG (Grass Phylogeny Working Group) 2000; GPWG (Grass Phylogeny Working Group) 2001). Notably, spikelet pairs are a defining morphological character of the Andropogoneae as most other grasses have spikelets as singlets (Kellogg 2000). Because *ra1* acts as a switch from long indeterminate branches to short determinate spikelet pairs in maize, one possibility is that *ra1* was co-opted for this role in spikelet pair development within Andropogoneae grasses.

Little is known about the genetic basis behind the morphological transformation of the maize ear following domestication. Because for *ra1*, noncoding sequences have the lowest nucleotide diversity and exhibit evidence of selection, it is possible that regulation of the gene may have been more important for domestication than protein composition. These observations are consistent with the hypothesis that altered gene regulation can be a principal genetic basis of morphological differences arising from plant domestication (Doebley and Lukens 1998), although there is perhaps equivocal evidence for the importance of protein changes as well (Hoekstra and Coyne 2007; Takeda and Matsuoka 2008). It has been suggested that genes targeted by selection are more likely to be expressed in tissues that underwent drastic modifications during periods of artificial selection (Hufford *et al.* 2007; Zhao *et al.* 2008). To date, the domestication locus *tga1*, which is responsible for the development of "naked" kernels in maize, is perhaps the only described gene that may help explain a portion of the mystery behind the dramatic changes implicit in the development of the modern maize ear. *tga1* encodes a transcription factor that is only expressed in the ear, thus simple changes in this gene can have dramatic phenotypic consequence in the ear.
without further deleterious pleiotropic effects in other tissues (Dorweiler et al. 1993; Dorweiler and Doebley 1997; Wang et al. 2005). Similarly, ra1 expression is restricted to developing inflorescences (Vollbrecht et al. 2005) which may minimize the likelihood of changes at the locus affecting other tissues in the plant. Given the examples of tga1 and ra1, it seems reasonable that many other domestication loci that have modified the morphology of the maize ear will be transcription factors or other developmental genes with expression limited to inflorescence tissues, and that identifying and characterizing these domestication loci will further our understanding of the evolution of the maize ear. Such information would be potentially very useful for plant breeding projects aimed at modifying maize and other cereal inflorescences to affect grain yields.

ACKNOWLEDGEMENTS
We would like to thank Thomas Peterson, Erica Unger-Wallace and two anonymous reviewers for constructive comments on the manuscript, and Jeffrey Ross-Ibarra for help with recombination analysis. This research was funded by Iowa State University and National Science Foundation DBI grant 0604923 (to E.V.).


TABLE 1
Frequency of Triplets in Weak *ra1* Mutants in Mo17 and B73 Backgrounds

<table>
<thead>
<tr>
<th>Genotype</th>
<th>n&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Quartile (number)&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td><em>ra1</em>-63 &gt; Mo17</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td><em>ra1</em>-63 &gt; B73</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td><em>ra1</em>-RS &gt; Mo17</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td><em>ra1</em>-RS &gt; B73</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of ears in each quartile. Quartile 0 = ears with no spikelet triplets, I = ears with 1-25% of total spikelets as triplets, II = 26-50% and III = 51-75%, respectively. No ears had >75% of total spikelets as triplets.

<sup>b</sup> Total number of ears sampled
## TABLE 2

Names and Origins of Plant Materials

<table>
<thead>
<tr>
<th>Sample</th>
<th>Origin</th>
<th>Accession</th>
<th>GenBank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sallu-yah (Cherokee)</td>
<td>United States</td>
<td>PI 213744</td>
<td>GQ891946</td>
</tr>
<tr>
<td>Bear Island Chippewa</td>
<td>United States</td>
<td>PI 213801</td>
<td>GQ891944 - GQ891945</td>
</tr>
<tr>
<td>6 Nations Obsweken</td>
<td>United States</td>
<td>Ames 2355</td>
<td>GQ891943</td>
</tr>
<tr>
<td>Eagle Corn</td>
<td>United States</td>
<td>PI 222285</td>
<td>GQ891942</td>
</tr>
<tr>
<td>Celaya</td>
<td>Mexico</td>
<td>NSL 2839</td>
<td>GQ891923 - GQ891924</td>
</tr>
<tr>
<td>Jala</td>
<td>Mexico</td>
<td>NSL 2834</td>
<td>GQ891936</td>
</tr>
<tr>
<td>Cariaco</td>
<td>Mexico</td>
<td>PI 260381</td>
<td>GQ891927</td>
</tr>
<tr>
<td>Conico</td>
<td>Mexico</td>
<td>NSL 2837</td>
<td>GQ891940</td>
</tr>
<tr>
<td>Chapalote</td>
<td>Mexico</td>
<td>PI 420245</td>
<td>GQ891939</td>
</tr>
<tr>
<td>Michoacan 286</td>
<td>Mexico</td>
<td>PI 629226</td>
<td>GQ891929</td>
</tr>
<tr>
<td>Costa Rica 370</td>
<td>Costa Rica</td>
<td>PI 498486</td>
<td>GQ891932 - GQ891933</td>
</tr>
<tr>
<td>Boyaca 476</td>
<td>Colombia</td>
<td>PI 444174</td>
<td>GQ891926</td>
</tr>
<tr>
<td>Magdalena 399</td>
<td>Colombia</td>
<td>PI 444954</td>
<td>GQ891925</td>
</tr>
<tr>
<td>Magdalena 469</td>
<td>Colombia</td>
<td>PI 445007</td>
<td>GQ891935</td>
</tr>
<tr>
<td>Cajamarca 15</td>
<td>Peru</td>
<td>PI 571896</td>
<td>GQ891941</td>
</tr>
<tr>
<td>Puno 26</td>
<td>Peru</td>
<td>PI 571808</td>
<td>GQ891938</td>
</tr>
<tr>
<td>Cuzco 63</td>
<td>Peru</td>
<td>PI 485333</td>
<td>GQ891928</td>
</tr>
<tr>
<td>ARG 2334</td>
<td>Argentina</td>
<td>NSL 6539</td>
<td>GQ891922</td>
</tr>
<tr>
<td>Argentine Pop</td>
<td>Argentina</td>
<td>PI 451691</td>
<td>GQ891937</td>
</tr>
<tr>
<td>Uruguay 756A</td>
<td>Uruguay</td>
<td>PI 477694</td>
<td>GQ891930</td>
</tr>
<tr>
<td>Uruguay 293A</td>
<td>Uruguay</td>
<td>PI 479149</td>
<td>GQ891931</td>
</tr>
<tr>
<td>White soft corn, Lenha</td>
<td>Brazil</td>
<td>NSL 20134</td>
<td>GQ891934</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>Ames 21797</td>
<td>GQ891915</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 566686</td>
<td>GQ891921</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 566688</td>
<td>GQ891916</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 566691</td>
<td>GQ891917 - GQ891918</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 566692</td>
<td>GQ891919 - GQ891920</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>Ames 21889</td>
<td>GQ891896 - GQ891898</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 384062</td>
<td>GQ891899 - GQ891902</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 384063</td>
<td>GQ891903 - GQ891904</td>
</tr>
<tr>
<td>Zea mays ssp.parviglumis</td>
<td>Mexico</td>
<td>PI 384064</td>
<td>GQ891905 - GQ891906</td>
</tr>
<tr>
<td>Zea mays ssp.mexicana</td>
<td>Mexico</td>
<td>PI 566674</td>
<td>GQ891913</td>
</tr>
<tr>
<td>Zea mays ssp.mexicana</td>
<td>Mexico</td>
<td>PI 566680</td>
<td>GQ891914</td>
</tr>
<tr>
<td>Zea mays ssp.mexicana</td>
<td>Mexico</td>
<td>PI 566687</td>
<td>GQ891891 - GQ891892</td>
</tr>
<tr>
<td>Zea mays ssp.mexicana</td>
<td>Mexico</td>
<td>PI 566691</td>
<td>GQ891893 - GQ891895</td>
</tr>
<tr>
<td>Zea mays ssp.mexicana</td>
<td>Mexico</td>
<td>PI 566697</td>
<td>GQ891911</td>
</tr>
<tr>
<td>Zea mays ssp. huehuetenangensis</td>
<td>Guatemala</td>
<td>PI 441934</td>
<td>GQ891910</td>
</tr>
<tr>
<td>Zea perennis</td>
<td>Mexico</td>
<td>Ames 21882</td>
<td>GQ891907</td>
</tr>
<tr>
<td>Zea diploperennis</td>
<td>Mexico</td>
<td>PI 441931</td>
<td>GQ891908 - GQ891909</td>
</tr>
<tr>
<td>Zea luxurians</td>
<td>Guatemala</td>
<td>PI 441933</td>
<td>GQ891912</td>
</tr>
<tr>
<td>Tripsacum dactyloides</td>
<td>United States</td>
<td>PI 421612</td>
<td>GQ891890</td>
</tr>
</tbody>
</table>
### TABLE 3

Nucleotide Diversity Statistics for *ramosa1*

<table>
<thead>
<tr>
<th></th>
<th>bp</th>
<th>n$^a$</th>
<th>$S^b$</th>
<th>$\pi^c$</th>
<th>$\theta^d$</th>
<th>$R_m^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Landraces</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5'</td>
<td>1270</td>
<td>25</td>
<td>13</td>
<td>0.002</td>
<td>0.004</td>
<td>2</td>
</tr>
<tr>
<td>CDS</td>
<td>560</td>
<td>25</td>
<td>6 (3)</td>
<td>0.003</td>
<td>0.004</td>
<td>0</td>
</tr>
<tr>
<td>3'</td>
<td>576</td>
<td>25</td>
<td>4</td>
<td>0.0007</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>2408</td>
<td>25</td>
<td>23 (3)</td>
<td>0.002</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td><strong>Teosintes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>51</td>
<td>0.013</td>
<td>0.022</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>22 (6)</td>
<td>0.017</td>
<td>-</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>37</td>
<td>0.013</td>
<td>0.022</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>110 (6)</td>
<td>0.013</td>
<td>-</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Total number of sequences
$^b$ Number of segregating silent plus synonymous, and nonsynonymous (in parentheses), sites
$^c$ Nucleotide diversity per site
$^d$ Nucleotide polymorphism per site
$^e$ Minimum number of recombination events
### TABLE 4

**Tests of Selection for ramosa1**

For all columns, statistically significant values are in bold type.

<table>
<thead>
<tr>
<th></th>
<th>bp</th>
<th>outgroup</th>
<th>HKA test P values</th>
<th>Landraces</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sites</td>
<td>adh1</td>
<td>bz2</td>
</tr>
<tr>
<td>5'</td>
<td>1270</td>
<td>Z. lux</td>
<td>829</td>
<td><strong>2.4E-03</strong></td>
<td><strong>1.8E-04</strong></td>
</tr>
<tr>
<td>CDS</td>
<td>560</td>
<td>Z. lux</td>
<td>127.9</td>
<td>0.58</td>
<td>0.43</td>
</tr>
<tr>
<td>3'</td>
<td>576</td>
<td>Z. lux</td>
<td>495</td>
<td><strong>1.6E-03</strong></td>
<td><strong>3.3E-04</strong></td>
</tr>
<tr>
<td>total</td>
<td>2408</td>
<td>Z. lux</td>
<td>1451.9</td>
<td><strong>9.3E-04</strong></td>
<td><strong>5.0E-05</strong></td>
</tr>
<tr>
<td>total</td>
<td>1324</td>
<td>T. dact</td>
<td>634.6</td>
<td>0.07</td>
<td><strong>1.8E-02</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>bp</th>
<th>outgroup</th>
<th>HKA test P values</th>
<th>Teosintes</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sites</td>
<td>adh1</td>
<td>bz2</td>
</tr>
<tr>
<td>5'</td>
<td>1270</td>
<td>Z. lux</td>
<td>691</td>
<td>0.60</td>
<td>0.35</td>
</tr>
<tr>
<td>CDS</td>
<td>560</td>
<td>Z. lux</td>
<td>128.1</td>
<td>0.78</td>
<td>0.95</td>
</tr>
<tr>
<td>3'</td>
<td>576</td>
<td>Z. lux</td>
<td>492</td>
<td>0.49</td>
<td>0.29</td>
</tr>
<tr>
<td>total</td>
<td>2408</td>
<td>Z. lux</td>
<td>1311.1</td>
<td>0.64</td>
<td>0.35</td>
</tr>
<tr>
<td>total</td>
<td>1324</td>
<td>T. dact</td>
<td>620.8</td>
<td>0.84</td>
<td>0.62</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

**Figure 1.** Maize *ramosa1* mutant phenotypes in mature tassels and ears. Wild type (inbred Mo17) tassel (A) and ear (D), weak mutant (*ra1*-63.3359 in Mo17) tassel (B) and ear (E), and strong mutant (*ra1*-R in a hybrid background) tassel (C) and ear (F). Weak mutant tassels have a few additional long branches (arrowhead) when compared to wild type whereas strong mutant tassels have long branches extending up the central rachis. Wild type ears have straight rows whereas weak mutant ears have crooked rows. Strong mutants exhibit branched ears.

**Figure 2.** Ear development of weak *ramosa1* mutants. (A-D) SEMs of developing, 0.5 cm ears from the inbred B73 (A) and the *ra1*-RS weak mutant in B73 (B). Boxed regions are examined in close-up views of a row of spikelet pairs from the B73 ear (C) and the *ra1*-RS ear (D), revealing delayed timing of spikelet pair development and elongated spikelet branches in mutants. (E-G) Portion of a developing (~ 1.5 cm) ear from a *ra1*-63.3359 weak mutant. Red arrows point to developing spikelet triplets (E). Some positions on mutant ears, and all positions on ears from inbreds, bear spikelet pairs (F) but only mutants form spikelet triplets, which go on to develop triplets of florets (G). Scale bars: 200 µm.

**Figure 3.** Patterns of nucleotide diversity at *ramosa1*. A sliding window analysis of nucleotide diversity (π) at *ra1* is shown for maize inbreds (solid green line), landraces (solid red line) and teosintes (solid blue line) across both noncoding and coding sequence of the *ra1* locus compared to π for a corresponding average gene (dotted lines) (Hufford et al. 2007b). Within the gene cartoon, the blue box represents the zinc finger and the yellow boxes represent putative EAR repression motifs; narrower black rectangles indicate UTRs. The lengths of the multiple sequence alignment for each region are 1236 bp for 5’ noncoding sequence (including the 60 bp, 5’-UTR), 566 bp for CDS, and 578 bp for 3’ (including the 164 bp, 3’-UTR). Significantly reduced diversity in both the 5’ and 3’ noncoding sequence for maize landraces is indicative of past selection. Step size = 100 bases, window size = 25 bases.
Figure 4. *ramosa1* haplotype trees. Bayesian phylogenies shown for the (A) 1271 base pair 5’ noncoding sequence, (B) 561 base pair coding sequence, and (C) 577 base pair 3’ noncoding sequence of *ra1*. Posterior probabilities and bootstrap support are shown for each node when > 50%. Single values represent posterior probabilities signifying no bootstrap support for that node. *Zea luxurians* sequence was used as the outgroup to orient the tree. Landrace samples are abbreviated as LR whereas teosintes are abbreviated as PARV (subspecies *parviglumis*) or MEX (subspecies *mexicana*), respectively. Other teosintes are indicated by their full name. The frequency of each haplotype within each population is in parentheses for each taxon.
Figure 1 Sigmon & Vollbrecht
Figure 2 Sigmon & Vollbrecht
Figure 3 Sigmon & Vollbrecht

![Graph and Table]

*Graph and Table Description*

**ra1 (2.38 kb)**

<table>
<thead>
<tr>
<th></th>
<th>5'</th>
<th>CDS</th>
<th>3'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teosintes (π)</td>
<td>0.013</td>
<td>0.017</td>
<td>0.012</td>
</tr>
<tr>
<td>Landraces (π)</td>
<td>0.003</td>
<td>0.003</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

**Figure Legend**

- **Graph**: Nucleotide diversity per base pair (π) across the 2.38 kb region of ra1.
- **Table**: Comparative nucleotide diversity (π) values for Teosintes and Landraces across different regions of ra1.
Figure 4 Sigmon & Vollbrecht
CHAPTER 3. MOLECULAR EVOLUTION OF \textit{RAMOSAI} IN THE ANDROPOGONEAE (POACEAE)

Brandi Sigmon, Renata Reinheimer, Elizabeth Kellogg, and Erik Vollbrecht

ABSTRACT
Grass inflorescence branching architecture is a morphological trait that is potentially subject to natural and artificial selection as it affects reproduction and grain yield. Genes underlying this trait are sought to explain the molecular basis underlying the phenotypic diversity of these structures. The \textit{ramosa1 (ra1)} gene encodes a transcription factor that controls branching architecture in maize inflorescences (tassel and ear). Reduced \textit{ra1} activity in maize produces ears with crooked rows, a phenotype that may have been selected on during the derivation of modern maize. Previous work demonstrated significantly reduced nucleotide diversity in a panel of unimproved maize landraces when compared to maize’s wild progenitor teosinte, suggesting \textit{ra1} was selected on during the domestication of maize. Since \textit{ra1} controls the branching architecture of maize inflorescences, we hypothesize this gene may also have been important in the evolution of inflorescence architecture in other grasses. To test this hypothesis, we determined the probable origins of \textit{ra1} during the evolution of grasses and sequenced \textit{ra1} in a diverse panel of species for subsequent analyses. In these grasses, \textit{ra1} is highly conserved in known functional motifs, which have been subject to strong purifying selection, whereas other regions of the gene are highly variable. Sequence variations including frameshifts, indels, point mutations, and duplication events correlate with differences in inflorescence morphology, which suggests these variations may have functional consequences.

INTRODUCTION
A molecular basis for morphological variation between species is often the action of transcription factors, which act as switches between developmental programs. This variation may be generated by spatiotemporal changes in transcription factor expression, which alters developmental programs. Studies have shown that these changes in transcription factor
expression patterns may correlate with phenotypic variation among species (Doebley and Lukens 1998; Wray 2003). In plants, grass inflorescences are an emerging model for the study of the molecular basis of morphological variation due to the presence across the phylogeny of immense phenotypic diversity, various genomic resources, and available experimental tools. Crop models including rice, sorghum, and maize provide a starting point for comparative studies with non-model species. Genes found to be involved in the inflorescence development of these model grasses are often found to be important in its evolution including its domestication from wild relatives (Vollbrecht and Sigmon 2005; Clark et al. 2006; Jin et al. 2008; Tan et al. 2008). Given the importance of domestication loci in crop species, it is reasonable to investigate their importance in other grass species.

In maize, the ramosa pathway has been shown to control inflorescence branching (Vollbrecht et al. 2005; Bortiri et al. 2006; Satoh-Nagasawa et al. 2006). Central to this pathway is the ramosa1 (ra1) gene, which has been most studied. The gene encodes a putative EPF-type C2H2 zinc finger transcription factor that is expressed in the boundary domain near the base of spikelet pair meristems (Vollbrecht et al. 2005). The gene contains a C2H2 zinc finger DNA-binding motif and an EAR repression motif, and through its interactions with a corepressor, RA1 likely acts as a transcriptional repressor (Gallavotti et al. 2010). Morphologically, ra1 is responsible for switching from the outgrowth of long indeterminate branches to the short determinate spikelet pair branches in maize inflorescences. This function results in male inflorescences with several long branches at the base of the tassel followed by numerous short spikelet pair branches and in female inflorescences completely lacking long branches. Therefore, proper function of ra1 includes the packing of kernels in straight rows on the ear, which would affect grain yields and therefore may have been of significance during domestication and/or improvement (Vollbrecht et al. 2005; Sigmon and Vollbrecht 2010). Indeed, a signature of selection was identified at the ra1 locus, which suggests it was a target of artificial selection during its domestication from its wild progenitor (Sigmon and Vollbrecht 2010).

Since ra1 may have been a target of artificial selection during maize domestication, it is reasonable to hypothesize ra1 may perform an important role in the development of inflorescence branching architectures in other grasses. An initial comparative expression
analysis using RT-PCR was done for maize, *Sorghum bicolor*, and the ornamental grass *Miscanthus sinensis* (Vollbrecht et al. 2005). No expression analysis could be done in rice, as data from previous BLAST searches suggest rice lacks an orthologous gene (Vollbrecht et al. 2005). Unlike maize, which produces only a few long branches in the tassel, the *Miscanthus* inflorescence meristem produces indeterminate long branches that extend almost to the tip of the main axis before switching abruptly to producing short determinate spikelet pair branches. The pattern of *ra1* expression corresponds to morphology in that the transition in producing second order meristems was accompanied by the onset of *ra1* expression leading to the production of spikelet pairs, as similarly found in maize tassels (Vollbrecht et al. 2005). However in sorghum, whose inflorescence is highly branched, spikelet pair fate becomes determined later. The delay in *ra1* expression also corresponds to this inflorescence branching phenotype (Vollbrecht et al. 2005). Since *ra1* has been shown to control the transition between long indeterminate branches and short determinate spikelet pair branches in maize and appears to be important in the formation of spikelet pairs in sorghum and *Miscanthus*, *ra1* function could explain much of the natural diversity in inflorescence architecture observed in grasses.

Maize and sorghum belong to the tribe Andropogoneae within the Panicoideae subfamily of the PACMAD clade of grasses (GPWG (Grass Phylogeny Working Group) 2000; GPWG (Grass Phylogeny Working Group) 2001; Sanchez-Ken et al. 2007). A hallmark of this grass tribe is variation in inflorescence morphology due to different branching architectures. Evidence from comparative developmental morphology supports the notion that variability results from changes in branch number and arrangement (Doust 2007); however, the molecular basis underlying this phenotypic diversity has largely not been investigated. Despite having strong hypotheses of which genes are involved in controlling branching in grass inflorescences (Doust 2007), few studies have been published to date. The *FLORICAULA/LEAFY*-like genes in maize, *zfl1* and *zfl2* are hypothesized to control differences in inflorescence structure as they function in determining floral organ identity and patterning (Bomblies et al. 2003). Functional divergence of these paralogous genes suggest the *zfl2* gene is associated with ear rank (row) number and therefore may explain some of the morphological evolution of maize inflorescences (Bomblies and Doebley 2006). However, a
study of the molecular evolution of these genes in the Andropogoneae revealed no evidence of directional selection during domestication or during the evolution of grasses (Bomblies and Doebley 2005). Moreover, without comparative expression analysis it is difficult to fully understand the zfI genes’ roles in the evolution of the development of inflorescence architecture in grasses including or even within the Andropogoneae.

To test if *ra1* was involved in the evolution of grass inflorescence morphology we first cloned *ra1*-like genes in a diverse set of Andropogoneae species, and then employed sequence, phylogenetic, molecular evolution, and comparative genomic analyses to address the following questions: 1) What can we infer about the origin of *ra1* in grasses, 2) Is there significant sequence variation among species or within specific genic regions, 3) Do sequences variations correlate with observed inflorescence morphologies, 4) Is there any evidence of positive selection for the whole gene or for specific regions of the gene, and 5) Was the tandem duplication of *ra1* important in the evolution of the sorghum inflorescence?

**MATERIALS AND METHODS**

**Comparative Genomics**

An alignment of a collinear segment was constructed based on the predicted genes in rice, sorghum, and maize. This region includes 247 kb of sequence on rice chromosome nine, 226 kb on sorghum chromosome two, 457 kb on maize chromosome two, and 1143 kb on maize chromosome seven. Sequences were obtained from maizesequence.org v4a53, MIPS/JGI Sbi1.4, and MSU Release 6.0. The Phytozome v5.0 genomic database browser (www.phytozome.net) was used for comparing these chromosomal segments.

**Plant Materials**

Putative *ra1* orthologs were isolated from 27 different Andropogoneae species, a tribe of the grass family Poaceae (Table 1). One *Arundinella* species, which is sister to the Andropogoneae, was also sampled as an outgroup. The taxa were selected to provide a broad representation within the Andropogoneae with a dense sampling of the sorghum genus. *ra1*-like genes were also isolated from a diverse panel of *Sorghum bicolor* varieties including 13 wild varieties, 20 landraces, two inbreds and 2 hybrids, along with six accessions belonging
to other species in the *Sorghum* genus (Supplementary Table 1).

**Sampling Strategy and Sequencing**

*ral* from a single individual was amplified from genomic DNA through PCR using the high fidelity ExTaq enzyme (TaKaRa) and the primers RA-35 (5’-AACGACGGATCACGTCTGTGTC-’3) which corresponds to a putative conserved enhancer element and RA-49 (5’-GCCCATCTAAGCTAGATCCA-’3) which corresponds to the C-terminus of the gene (Vollbrecht et al. 2005). Amplicons ranging 780 bp to 967 bp in size were then purified using QIAquick PCR Purification Kit (Qiagen, Valencia, CA) and cloned using pCR 2.1-TOPO (Invitrogen) to separate different alleles or gene copies. At least eight clones from each PCR reaction were then sequenced by the Sanger method to control for PCR errors and heterozygosity. Sequencing from both directions was also done in order to ensure correct sequence calls. Sequence reads were assembled and edited using the PREGAP and GAP4 software from the Staden package (Staden 1996).

**Sequence Analysis**

Sequences were initially aligned using ClustalW (Larkin et al. 2007) as part of the MacVector Version 11.1 software (http://www.macvector.com) and further adjusted by hand. GC content, pairwise $d_s/d_*$ ratios, codon biases, and calculations of $R_m$ were all calculated using DnaSP, Version 3.51 (Rozas and Rozas 1999). A graphical representation of sequence conservation was produced through analysis of nucleotide alignments using the phylogenetic footprint software eShadow (Ovcharenko, Boffelli, and Loots 2004).

**Phylogenetic Reconstruction**

jModelTest Version 0.1.1 (Posada 2009) was used to determine the best-fit model of nucleotide substitution for phylogenetic reconstruction. Phylogenies were reconstructed using PAUP* Version 4.0b10 (Swofford 2003) and MrBayes Version 3.1.2 (Ronquist and Huelsenbeck 2003). Parsimony and neighbor-joining (Saitou and Nei 1987) phylogenies were reconstructed using PAUP* with outgroup rooting, and robustness was assessed by 100
replicates through the fast stepwise-addition method. Bootstrap support was assessed utilizing a full heuristic search with 1000 bootstrap replicates. Phylogenies made using MrBayes were reconstructed using a Markov chain Monte Carlo (MCMC) algorithm under an HKY+$\Gamma$ model. Parameters and priors were set to account for nucleotide frequencies, substitution rates, transition/transversion ratio, and the shape of the gamma distribution curve. These values were all provided by jModelTest. The branch length prior was set to be uniform and unconstrained. All other parameters and priors were set at default. Two independent runs of four chains each (one cold and three heated) were run for 1,000,000 generations and the first 10,000 trees were removed. All phylogenies were viewed using the software TreeView X v0.4.1 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

RESULTS

Origin of ra1 in the Grasses

Genes in the ramosa pathway including ramosa1 (ra1), ramosa2 (ra2), and ramosa3 (ra3) have been shown to control branching architecture in grass inflorescences (Vollbrecht et al. 2005; Bortiri et al. 2006; Satoh-Nagasawa et al. 2006). Given the diversity of inflorescence branching, it is hypothesized that the differential expression of these ramosa genes, in addition to other unidentified genes in the pathway, may explain the differences in branching phenotypes. Both putative ra2 and ra3 orthologs have been identified in maize, sorghum, and rice but not in Arabidopsis, suggesting these genes are monocot-specific (Bortiri et al. 2006; Satoh-Nagasawa et al. 2006). However from previous BLAST searches, ra1 is presumed to be absent in both Arabidopsis and rice due to a lack of sequence similarity (Vollbrecht et al. 2005). To gain insight into the origin of ra1 in grasses we examined the corresponding orthologous chromosomal segments in maize, rice, and sorghum.

In maize ra1 is located on chromosome 7 (Vollbrecht et al. 2005). A chromosomal segment (1143 kb) containing ra1 was found to be orthologous to segments of rice chromosome 9 (247 kb) and sorghum chromosome 2 (226 kb) and homeologous to a segment of maize chromosome 2 (457 kb). An alignment of these chromosomal segments showed largely conserved synteny between species with respect to gene content and order with a few notable exceptions (Figure 2). Consistent with previous BLAST results, a ra1-like sequence
is absent from the syntenic block in rice whereas many surrounding genes are present. The absence of *ral* and some surrounding sequence is apparently due to two noncollinear segments in rice, which contains 19 genes found in neither maize 7, maize 2, nor sorghum 2. This break in collinearity therefore provides additional evidence that *ral* is not present in rice. Also, unlike maize, sorghum chromosome 2 has two copies of *ral* due to a tandem duplication (Vollbrecht et al. 2005). Examination of the surrounding genomic sequence reveals ~6 kb of sequence was duplicated which includes the immediate 5’ sequence, the entire *ral* protein coding sequence and immediate 3’ sequence. Because adjacent genes have no apparent duplicates, the duplication event involved only *ral*. Finally, due to an allotetraploidy event 5-12 mya (Swigonova et al. 2004), segments of chromosome 7 and 2 are homeologous in maize, which suggests the corresponding syntenic region on chromosome 2 may contain a *ral*-like sequence. However, no *ral*-like gene or gene remnant was found suggesting the duplicate was deleted after the allotetraploidy event. Other genes near *ral* appear to have been deleted as well. These results are consistent with the observation that the chromosome 2 homeolog has experienced a preferential bias towards gene deletion following the allotetraploidy event compared to chromosome 7 (Xu and Messing 2008). Also the *ral* region of chromosome 7 appears to have expanded in relation to maize chromosome 2 and the corresponding orthologous regions in rice and sorghum, which is also consistent with previous observations for these chromosomes (Xu and Messing 2008).

From the totality of these data we suggest as a working hypothesis, at the earliest, *ral* originated with the PACMAD clade in grasses due to its absence in rice (Figure 1B). *ral* was apparently not lost specifically in rice, as it is absent from *Brachypodium*, a Pooideae grass, as well. Also, we hypothesize *ral* to be present throughout the tribe Andropogoneae (Poaceae) due both to its presence in a conserved syntenic region in sorghum and maize and in conjunction with the previously reported cloning and expression data in maize, sorghum, and *Miscanthus* (Vollbrecht et al. 2005). Furthermore, we conclude maize has only one *ral*-like gene, as it appears the homeolog was deleted at some point following the allotetraploidy event. As more grass genomes are sequenced the origin of *ral* will be more precisely pinpointed, which may elucidate how events in the evolution of the *ramosa* pathway may have impacted changes in grass inflorescence architecture.
**ra1 Sequence Conservation**

Given the likely presence of *ra1* throughout the Andropogoneae, we focused on this grass tribe for subsequent sequence analyses. We were able to clone and sequence *ra1* across a diverse panel of Andropogoneae species for a total of 44 *ra1*-like sequences in 27 different species (Table 1). Also, a putative ortholog in a species belonging to the sister genus *Arundinella* was cloned, which suggests the origin of *ra1* likely predates the origin of the Andropogoneae. This sequence also provides an outgroup for phylogenetic and evolutionary analyses. Since *ra1* is the only single-zinc finger EPF gene known to have the conserved alpha-helical sequence motif QGGLGGH in eudicots and monocots (Sigmon, unpublished), we were confident in distinguishing *ra1* orthologs from other gene members of the EPF subfamily of C2H2 zinc finger transcription factors. In almost all cases only one *ra1*-like gene with 1-2 alleles was identified, suggesting the tandem duplicate gene structure present in sorghum may not be present in other closely related grasses. For most species, at least 780-967 bp including the coding region and immediate 5’ non-transcribed region was amplified and sequenced, which is sufficient for phylogenetic and sequence analyses.

Initial sequence analysis shows varying levels of conservation throughout the 5’ noncoding and coding sequence. Alignments of the coding sequence show that previously described functional motifs, including the zinc finger DNA-binding motif and the C-terminal EAR repression motif (Vollbrecht et al. 2005), are highly conserved across all species. We found additional, short DNA sequences that are highly conserved including one encoding a second EAR motif, a 5’ upstream element, and the sequence immediately adjacent to the C-terminal EAR motif. All these conserved motifs are present throughout all Andropogoneae grasses sampled, but otherwise the nucleotide sequence is not well conserved (Figure 3). This pattern is similar to that of other known transcription factors such as *teosinte branched1*, where functional domains of the gene are conserved, but the sequence is otherwise inundated with indels (Lukens and Doebley 2001). Most of the conserved sequences we discovered have no known functions. However the second EAR motif was tested and shown, like the C-terminal EAR, to participate in interactions with a known corepressor protein (Gallavotti et al. 2010). Identifying these conserved sequences may lead to future experiments to determine
the role these sequences have in regulating expression, as well as, in identifying potential interacting partners.

We also found some unique sequence anomalies in some grass species. As mentioned previously, sorghum possesses two copies of ra1 due to a tandem duplication. The upstream gene copy is frameshifted due to a single base pair deletion, which results in a premature stop codon whereas the downstream copy remains intact and is putatively the fully functional gene copy. This mutation leaves the zinc finger DNA-binding motif intact, but without both EAR repression motifs, which suggests the upstream copy may have retained its DNA-binding ability. This tandem duplication may be the molecular change responsible for the relatively delayed expression of ra1 in sorghum, which results in reiterative branching in the inflorescence reminiscent of strong ra1 mutants in maize (Vollbrecht et al. 2005).

In the grass species Capillipedium parviflorum, a SNP creates a missense difference in the stop codon so that an additional six amino acids are added to the C-terminus of the protein. In the C. parviflorum inflorescence a combination of spikelet pairs and triplets are present, which is reminiscent of the phenotype displayed by weak mutant allele ra163.3359 in maize, which has an additional 17 amino acid added to the C-terminus (Sigmon and Vollbrecht 2010).

Finally, from sequencing multiple clones of ra1 in Schizachryium scoparium, we found putative gene copies exhibiting substantial divergence in nucleotide and amino acid sequence suggesting these sequence are likely not allelic. This result was found in both sampled accessions of S. scoparium sequenced and none of the gene products were identical. S. scoparium is also a polyploid, so the extra copies may be due to this mechanism, but a single gene duplication event is also possible. A hypothesis for the role these gene copies may have on the development of the inflorescence is not obvious without additional functional data.

Using sequence analysis, we have found several events that may correlate with morphological differences in inflorescence architecture in a few grass species. However, we were unable to discern events in most other species, which suggests more subtle sequence differences and changes in gene regulation may indicate a role for ra1 in the variation in inflorescence branching morphology in the Andropogoneae.
**ra1 Gene Tree**

A *ra1* gene tree was reconstructed using the 1350 bp nucleotide sequence alignment of the putative *ra1* orthologs for 27 grass species sampled (Figure 4). For this phylogeny, one allele was arbitrarily chosen for each species and putative gene duplicates were not included. A second phylogeny was also reconstructed for all putative alleles and gene copies using an alignment of the coding sequence (Supplementary Figure 1). Since the genus *Arundinella* is sister to the Andropogoneae (GPWG (Grass Phylogeny Working Group) 2000; GPWG (Grass Phylogeny Working Group) 2001), this sequence was used as an outgroup to root both phylogenies. The reconstructed gene tree architecture agrees roughly with previously published phylogenies in several key aspects, (Spangler et al. 1999; GPWG (Grass Phylogeny Working Group) 2000; GPWG (Grass Phylogeny Working Group) 2001; Lukens and Doebley 2001; Mathews et al. 2002), lending further support to the conclusion that the cloned *ra1*-like sequences are orthologous. A phylogenetic approach is also useful in the identification, positioning, and timing of molecular changes within an evolutionary context. The tandem gene copies of *ra1* in sorghum clustered together, suggesting the duplication event occurred relatively recently and may not be present outside the *Sorghum* genus (Supplementary Figure 1). Also, the putative gene copies in *S. scoparium* clustered together, but with less support than found in sorghum (Supplementary Figure 1).

In addition to being architecturally similar to other gene trees for the Andropogoneae, the *ra1* gene tree displays several unique features. It has short internal, but long terminal branches, which is typical of all Andropogoneae phylogenies based on molecular data reconstructed to date (Spangler et al. 1999; GPWG (Grass Phylogeny Working Group) 2000; GPWG (Grass Phylogeny Working Group) 2001; Lukens and Doebley 2001; Mathews et al. 2002). This topology is likely due to the rapid radiation of grasses about 35 mya (Bremer 2002), which makes accurate reconstruction and resolution of deep phylogenetic relationships difficult. However, the *ra1* gene tree provides strong support for several generic relationships including *Tripsacum* + *Zea*, *Rottboellia* + *Coelorachis*, *Urelytrum* + *Oxyrhachis*, *Chrysopogon* + *Trachypogon*, *Capillipedium* + *Dicanthium*, and *Hyparrhenia* + *Diheteropogon* (Figure 4). Also, there is universally strong support for clades with different alleles within a species, different species within the same genus, and gene copies within a
species (Supplementary Figure 1). Most notably, the $ra1$ gene tree provides strong support for several deep nodes along the backbone of the phylogeny, which is unique among phylogenies based on molecular data. This is the first reported phylogeny that strongly supports awned and awnless clades (both with 100% posterior probability), and the placement of $Chionachne$ as an early diverging species sister to the common ancestor of the awned and awnless clades. Other molecular phylogenies predict an awned clade without support and provide no resolution for awnless species (Spangler et al. 1999; GPWG (Grass Phylogeny Working Group) 2000; GPWG (Grass Phylogeny Working Group) 2001; Lukens and Doebley 2001; Mathews et al. 2002). Given this unique result, $ra1$ gene phylogeny will be of general interest to grass systematists.

**Molecular Evolution of $ra1$ Orthologs**

Inflorescence architecture is a morphological trait subject to both natural and artificial selection as it potentially impacts the reproductive success of plants. Previously, we demonstrated that $ra1$ may have been a target of artificial selection during the domestication of maize in order to maximize the straightness of kernel rows in the maize ear (Sigmon and Vollbrecht 2010). Since $ra1$ may have been a target of selection during maize domestication and given its function in determining inflorescence structure as part of the $ramosa$ pathway, $ra1$ may have been subject to different selective pressures in other grass species. To address this hypothesis, we employed several analyses to examine the molecular evolution of $ra1$. First, we calculated pairwise ratios of nonsynonymous to synonymous substitutions ($dN/dS$) for the whole gene to see if it has been targeted by selection. For the Andropogoneae, we found a range of values from $dN/dS = 0.30$ to 0.64 (mean = 0.43), suggesting varying levels of purifying selection since $dN/dS < 1$ (Table 3). Next, to test for selection with higher resolution we used a sliding window of $dN/dS$ values to determine if particular regions of $ra1$ have been subject to positive and/or purifying selection. From this analysis, we found evidence of purifying selection ($dN/dS < 1$) across the most of the middle of the gene, which includes the zinc finger DNA-binding motif. (Figure 3). These data suggest most nonsynonymous changes in these regions of $ra1$ were detrimental and were therefore eliminated through purifying selection. These results, especially for the zinc finger motif, are not unexpected
given the known importance these conserved motifs to proper gene function. Several regions of \textit{ra1} have $d_N/d_S$ values greater than one, which suggests these region of the gene may have been subjected to positive selection. Peaks suggestive of positive selection were found adjacent to the zinc finger, the first EAR, and in some conserved sequence with unknown function adjacent to the C-terminal EAR motif. Amino acid changes in this region could possibly result in a modification of the strength of repression since amino acid changes in and adjacent to EAR motifs has been shown to have this effect in other genes (Kagale, Links, and Rozwadowski 2010). Another peak was found near the N-terminus of the gene. This region contains microsatellites, which are more polymorphic in general and may explain the high rate of nonsynonymous substitutions.

Codon bias is the preference of a certain subset of synonymous codons for an amino acid in the universal genetic code. No codon biases indicate all amino acids are being used equally. Codon biases can be detected through a number of tests including the calculation of the effective number of codons or ENC (Wright 1990) and by the interpretation of GC content. For \textit{ra1}, the range of ENC values were from 44.4 in \textit{Oxyrachis gracillima} to 61 in maize with a mean value of 52. These ENC values indicate moderate to no bias suggesting selection for translational efficiency has not been a major force in the molecular evolution of \textit{ra1}. Codon bias is also correlated with synonymous GC content since a high codon bias is usually due to a preference for codons that end in G or C. For \textit{ra1} the mean GC3 (third position in the codon) content was 68% with a range of 65% in \textit{Coix lacryma-jobi} and maize to 76% in \textit{Oxyrachis gracillima}. These values, like the ENC values are not indicative of strong bias.

Finally $d_S/d_S$, ENC, and GC3 values were mapped onto the \textit{ra1} gene phylogeny to see if values were similar in closely related species, which would suggest shared evolution of bias strength along branches of the tree (data not shown). ENC and GC3 values were similar between different alleles within a species, between gene copies (ex. sorghum), between species in the same genus (ex. \textit{Andropogon} spp. and \textit{Hyparrhenia} spp.), and in some closely related genera (ex. Tripsacum and maize). Due to a lack of support for many nodes, no conclusions could be made pertaining to most generic comparisons. Although codon bias and synonymous GC content provides some information on the selective constraint acting on
protein evolution, it does not preclude constraint on regulatory sequences or in specific functional domains within proteins.

**ra1 Duplication in Sorghum**

Sorghum contains a whole gene tandem duplication of *ra1*. The duplication comprises ~ 6 kb of sequence (Figure 5A). In the reconstruction of the *ra1* gene phylogeny the tandem duplicates of *ra1* in sorghum are monophyletic (Supplementary Figure 1), which suggests the paralogous duplication event occurred relatively recently and may not be present in other grass species. To further investigate the origins of the duplication event as well as the subsequent frameshift mutation in the upstream gene copy, we cloned and sequenced *ra1* in a diverse panel of *Sorghum bicolor* varieties and subspecies. The sorghum genus is paraphyletic with 25 species organized into five taxonomic sections within two major clades (Dillon, Lawrence, and Henry 2001). We first surveyed the diversity of the section *Eusorghum*, which consists of *Sorghum bicolor* varieties and subspecies including two inbreds, 20 cultivated landraces, 13 wild accessions, *S. propinquum* and *S. halepense*. We also sequenced *ra1* in two other species including *S. laxiflorum* (*Heterosorghum*) and *S. purpureosericeum* (*Parasorghum*) (Supplementary Table 1).

All members of the *Eusorghum* diversity panel, which includes the cultivated and wild sorghum varieties, had two *ra1* copies, providing evidence that this event predates the domestication of sorghum. Further evidence is provided by the *ra1* sorghum phylogeny, which shows the upstream and downstream copies of *S. bicolor* and its close outgroup form separate monophyletic clades (Figure 6). Also, the more distantly related species from other sections of the genus representing the two major sorghum clades (Dillon, Lawrence, and Henry 2001) appear to have two copies, suggesting the duplication likely originated at least with the *Sorghum* genus. To more precisely pinpoint the timing of the duplication event, other species should be sampled including the early diverging sorghum species *S. extans*, *S. intrans*, and *S. matarankense*, as well as, others in genera sister to the *Sorghum* genus.

The vast majority of accessions, including all cultivated and most wild varieties, had the post zinc finger frameshift mutation in the upstream gene copy and a fully intact downstream copy. However, several wild accessions had a different frameshift mutation in
the upstream gene copy near the C-terminus of the gene (Figure 5B). Upon closer inspection of the haplotype structure, it is evident that recombination between the gene copies has occurred resulting in small tracks of converted sequence (Figure 5C). Evidence of recombination was mainly in the wild varieties and the more distantly related sorghum species. Of the 75 gene copies examined, 30 were downstream-like (fully functional), 27 were upstream-like (frameshift), and 18 were recombinant. Of these 18 recombinant types, only four were from cultivated varieties. The calculated minimum number of recombination events (R_m) (Hudson and Kaplan 1985) was three, meaning at least three recombination events have occurred. However, the polymorphism pattern for most of these recombinant types is unique to their accession, suggesting this R_m may be an underestimate. Regardless, this level of recombination is unusual for this gene (Vollbrecht et al. 2005; Sigmon and Vollbrecht 2010) and would be expected to impact levels of nucleotide diversity.

To address this question, we analyzed the patterns of nucleotide diversity for both copies of ra1 in cultivated and wild accessions of sorghum (Supplementary Table 1). For cultivated sorghum, nucleotide diversity (π) was not substantially different between the upstream and downstream gene copies (π = 0.0007 vs. 0.0009, respectively). These values are approximately three-fold lower than the estimated average nucleotide diversity for the sorghum genome, π = 0.0022 (Hamblin et al. 2006). For wild sorghum π = 0.0035 at ra1, which is substantially higher than cultivated sorghum, and higher than the average for the sorghum genome. The large number of recombinant-type gene copies may have increased the levels of nucleotide diversity in wild sorghum compared to cultivated varieties. Comparing π between the cultivated and wild accessions indicates that at ra1, cultivated sorghum retained 19.9% of the nucleotide diversity from its wild relatives. Previous studies have estimated that domesticated sorghum has retained about 66% of the diversity from its wild progenitor (Morden, Doebley, and Schertz 1990). The reduction of π at the locus in cultivated sorghum suggests the ra1 locus may have been subjected to artificial selection during its domestication 3000-6000 years ago. In our diversity panel, all cultivated sorghum had the early frameshift in the upstream copy and an intact coding region in the downstream copy. Because this combination predominates, it may have been preferred during selection, for example if it was responsible for a more highly branched inflorescence. Previously, it has
been suggested that selection for increased inflorescence branching occurred early during the domestication of sorghum in order to increase grain yields (Dillon et al. 2007).

To test for evidence of selection, we employed HKA tests, which compares ratios of intraspecific diversity and divergence between species to detect signatures of selection (Wright and Charlesworth 2004). For these tests we used S. propinquum as an outgroup and a reference locus comprised of pooled data from 204 loci (Hamblin et al. 2006). We tested the upstream and downstream gene copies separately for our panel of cultivated sorghum varieties. Only the upstream copy showed evidence of selection (p-value = 0.016), but this result does not preclude the possibility that the downstream copy may have been as a target of selection. Detecting evidence for directional selection in sorghum is unusual due to low diversity, population structure, sampling, and mating system issues (Hamblin et al. 2004; White, Hamblin, and Kresovich 2004; Casa et al. 2005; Casa et al. 2006; Hamblin et al. 2006). However if the reduced levels of diversity and the significant HKA test are indicative of selection, then ra1 may have been subjected to artificial selection during the domestication and/or improvement of sorghum. To further narrow the timeframe, a more comprehensive sampling wild sorghum accessions is necessary.

DISCUSSION

The Origins of ra1 and Implications for Morphological Evolution

Our examination of a chromosomal segment that is syntenic between rice, sorghum, and homeologous regions in maize, further supports the inference that ra1 is absent in rice. It is possible that ra1 is located elsewhere in the genome and at some point moved to its current location through duplication or insertion after the BEP and PACMAD clades diverged. However if this scenario is correct, it is unlikely the homologous gene would have the same function due to the large degree of sequence divergence of known functional motifs in particular. We also sequenced ra1 in a diverse set of species within the tribe Andropogoneae and in one Arundinella species. This suggests ra1 is present throughout Andropogoneae, but did not originate with the tribe since it is also present in Arundinella, which is sister to the Andropogoneae. These data do not precisely pinpoint the origin of ra1 in grasses but do narrow the timeframe to after the divergence of rice and maize some 50-70 mya (Kellogg
2001) and prior to the origins of the Andropogoneae. As more grass genomes are sequenced, synteny can be used to look for *ra1* using the maize, sorghum, and rice genomes as guides. Using such comparative genomics between multiple species, a mechanism for the origin of *ra1* in grasses may be revealed.

The presence of paired versus single spikelets is considered a defining morphological character of the Andropogoneae since most species have this trait (Clayton and Renvoize 1986). Thus, given the role of *ra1* in the switch to producing spikelet pairs in maize (Vollbrecht et al. 2005; Sigmon and Vollbrecht 2010), it is not unexpected that a putative *ra1* ortholog is present throughout the Andropogoneae, where it could perform a similar function. However, other grasses outside the Andropogoneae also have paired spikelets, suggesting this innovation is not unique to this tribe. Paired spikelets are also present in some clades belonging to the other major Panicoideae tribe Paniceae s.l., as well as, Pharoideae an early diverging grass lineage (Mathews et al. 2002). These other grasses may also have *ra1* orthologs or other modifications in the *ramosa* pathway have occurred. Examining other genes involved in inflorescence branching would help address this question. Genes that control inflorescence architecture are also being identified in other grass models such as rice. Recently, gain of function alleles of the *ABERRANT PANICLE ORGANIZATION1 (APO1)* gene, which is an ortholog of *UNUSUAL FLORAL ORGANS (UFO)* in Arabidopsis, was found to cause an increase in branching due to a delay in the program shift to spikelet production (Ikeda-Kawakatsu et al. 2009), a gene function similar to *ra1*. At this time it is unknown if the *APO1* gene is involved in the *ramosa* pathway, but it does suggest other genes that determine inflorescence architecture are yet to be identified. Also, it has been suggested that inflorescence form is an easily modified trait in evolutionary time since it is so variable and is homoplasious on the phylogeny (Mathews et al. 2002). Therefore, it is likely that many genetic mechanisms control inflorescence development and modifications to the genes involved would explain the great diversity in inflorescence form.

**Implications for the Andropogoneae Phylogeny**

The *ra1* gene tree, like other phylogenies of the Andropogoneae based on molecular data, has short internal branches due to a lack of substitutions along the backbone of the tree,
which confounds reconstruction (Spangler et al. 1999; Lukens and Doebley 2001; Mathews et al. 2002). This topology, found universally to date, is most likely due to rapid radiation of species instead of convergent evolution since all gene trees display this same topology (Mathews et al. 2002). However, the \textit{ra1} gene tree does provide strong support for some key internal nodes including both awnless and awned clades, which are predicted by phylogenies reconstructed using morphological characters (Kellogg and Watson 1993). Other Andropogoneae phylogenies based on molecular data leave the relationships of the awnless species unresolved and predict an awned clade, though it is unsupported (Spangler et al. 1999; GPWG 2000; GPWG (Grass Phylogeny Working Group) 2000; GPWG 2001; GPWG (Grass Phylogeny Working Group) 2001; Lukens and Doebley 2001; Mathews et al. 2002). Therefore a combined analysis of the \textit{ra1} dataset with other molecular data might be beneficial in resolving some of the relationships between groups of genera within the Andropogoneae. However, it is unclear why \textit{ra1} would have a number of substitutions supporting awned and awnless clades. The simplest hypothesis is that \textit{ra1} function is somehow related to awn development in these species. Several, but not all, of the substitutions supporting these clades are located in or near conserved functional motifs. The zinc finger and EAR repression motifs are important to \textit{ra1} gene function and would thus be under tighter selective constraint than the rest of the gene. Nonsynonymous changes in these sequences would most likely be deleterious and removed by purifying selection. However, some of these changes must have resulted in adaptive modifications to \textit{ra1} function. For example, changes in sequence could strengthen interactions with binding partners or in the EAR motif, modify levels of repressive activity.

The \textit{ra1} phylogeny also provides strong support for some relationships between and within genera and for alleles within species (Supplementary Figure 1). Some generic relationships have been predicted elsewhere, including \textit{Tripsacum} + \textit{Zea}, \textit{Capillipedium} + \textit{Dichanthium}, and \textit{Hyparrhenia} + \textit{Andropogon} (Mathews et al. 2002). Others have not been previously reported including the inclusion of \textit{Diheteropogon} in the “core Andropogoneae” and its sister relationship with \textit{Hyparrhenia}. Other well-supported generic relationships include \textit{Rotboellia} + \textit{Coelorachis}, \textit{Elionurus} sister to \textit{Oxyrachis} + \textit{Urelytrum}, and \textit{Trachypogon} + \textit{Chrysopogon}. Also, species within their respective genera were
monophyletic, which includes species belonging to *Coix*, *Andropogon*, and *Hyparrhenia*. These results are not unexpected, but the monophyly of most genera remains unaddressed to date (Mathews et al. 2002). Broader sampling within genera is needed to address this question properly.

**Function of Conserved Sequence Motifs**

From the alignment of *ra1* we found several conserved sequence motifs. Most notably we discovered an additional, highly conserved EAR repression motif (Figure 3). Recently this second EAR motif, in addition to the C-terminal EAR, has been shown to be necessary for physical interaction between the *ramosa1 enhancer locus* (*rel2*) and *ra1* gene products. RA1 and REL2 are thus components of a hypothesized transcriptional repressor complex that regulates axillary meristem fate in maize (Gallavotti et al. 2010). The weak *ra1-63.3359* allele of maize contains an additional 17 amino acids at the C-terminus that lead to a decreased interaction between REL2 and RA1 (Gallavotti et al. 2010). Interestingly, *C. parviflorum*, which we found to have an extra six amino acids added to the C-terminus, has a branched inflorescence with spikelet triplets. By analogy, the addition of six amino acids in *C. parviflorum* could hinder the interaction of REL2 with the C-terminal EAR resulting in increased indeterminacy of the spikelet pair meristems. Genes with multiple EAR motifs are present in other gene families (Kagale, Links, and Rozwadowski 2010), however their role in gene function remains unknown. We also found that the sequence immediately adjacent to the EAR motifs to be conserved (Figure 3). Sequence adjacent to EARs has been shown to be important for proper repressive function (Kagale, Links, and Rozwadowski 2010). Therefore the conservation of these sequences in *ra1* is likely due to selection against most nonsynonymous substitutions in order to maintain proper gene function. However, amino acid change in these sequences could result in a functional modification that may be adaptive in some species. In addition to the EARs we also found highly conserved sequence motifs in the immediate 5’ noncoding sequence. The most notable among these is a large motif ~350-400 bp upstream of the coding sequence (Figure 3). The large size of this conserved sequence motif suggests it may be an operator element that interacts with several other transcription factors. Several other, shorter sequence motifs were also present and there functions are also
unknown at the time. However, cis-regulatory elements involved in regulating gene function are typically small and prone to rapid turnover and therefore would not be evident from multiple sequence alignment (Wray et al. 2003). Further molecular and computational analyses may help determine the role of these sequence motifs in *ral* function.

**The Role of *ral* in the Evolution of Sorghum Inflorescence Morphology**

In our analysis of the sorghum diversity panel, which included cultivated and wild varieties of *S. bicolor* in addition to several other species representing other major taxonomic sections, we discovered that the duplication of *ral* predates the domestication of sorghum and likely originates with the *Sorghum* genus. The vast majority of the cultivated accessions and many wild varieties possess a duplicate of *ral* with a single pair base deletion after the zinc finger DNA-binding domain in the upstream gene copy, which shifts a nearby stop codon into frame. By contrast, the second, downstream copy contains an intact coding sequence (Figure 5A). However many wild accessions had recombinant-type gene copies that had in the upstream copy the predominant frameshift or a different frameshift. Since all of these sorghums have a highly branched inflorescence, it is unlikely any particular frameshift is solely responsible for the high-branch phenotype. However, at least two other explanations are consistent with such a role for the tandem duplication including 1) the upstream copy competes with the downstream copy for binding of regulatory factors, 2) the gene duplication itself alters regulation for example by changing the distance between *cis*-regulatory elements or by creating novel ones. On the other hand, some other gene or trans-acting factor may be responsible, and the duplication may not be a primary contributor to the phenotypic difference. Transgenic experiments using maize and sorghum will help distinguish between these possibilities.

The prevalence of this duplication event, which correlates with a highly branched inflorescence, suggests selection may have fixed this event due to an improvement in reproductive success. A significant increase in inflorescence branching would produce a more compact inflorescence that holds more grain, which could be an advantage favored by natural selection. This advantage could explain the prevalence of this duplication in wild sorghum. It has also been suggested that selection for increased branching early in the
domestication sorghum was key to increasing yield in cultivated sorghum varieties (Dillon et al. 2007). It seems plausible that this developmental change involved the ramosa pathway and may involve the duplication of ra1. Finding evidence of selection on a gene during sorghum domestication is hindered by difficulties inherent to the species such as low levels of nucleotide diversity and its population structure (Hamblin et al. 2004; White, Hamblin, and Kresovich 2004; Casa et al. 2005; Casa et al. 2006; Hamblin et al. 2006). However, we did find one prevalent haplotype and very low nucleotide diversity in cultivated sorghum when compared to wild varieties. If indeed ra1 was a target of selection in sorghum then the gene has been key to the domestication of more than one cereal grass, presumably through divergent selection on increased grain yields in both cases. In maize there is evidence of selection during domestication for alleles that decrease branching, which may result in an increase in yield due to more efficient packing of kernels into straight rows on the ear (Sigmon and Vollbrecht 2010). However in the domestication of sorghum, an increase of branching was preferred for increasing yield. This example would illustrate how different selection strategies can lead to the same result and may be accomplished by artificial selection for divergent phenotypic output from the same gene and/or pathway. Recently, it has been suggested that there may be few suitable targets amenable to artificial selection because generally, as seen in domesticated grasses, one selected gene impacts one trait (Sang 2009) presumably to reduce the likelihood of deleterious pleiotropic effects (Wang et al. 2005; Sigmon and Vollbrecht 2010). ra1 is expressed only in the inflorescences (Vollbrecht et al. 2005) and therefore may be particularly amenable to artificial selection in maize and sorghum to modify inflorescence architecture. It would be interesting to conduct analyses to quantify the impact these different sorghum haplotypes have on inflorescence branching. This experiment may be important to sorghum breeding programs for maximizing grain yields.

ACKNOWLEDGEMENTS
We would like to thank Patrick Brown for most of the sorghum ramosa1 sequences (Supplementary Table 1) used for the diversity analyses, as well as, Jacqueline Farrell and Chris Gillespe for conducting some of the molecular work.
LITERATURE CITED


### TABLE 1
Sources of sampled species

<table>
<thead>
<tr>
<th>Sample</th>
<th>Collection</th>
<th>Origin</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panicoideae, Arundinelleae:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arundinella nepalensis</td>
<td>PI 384059</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td><strong>Panicoideae, Andropogoneae:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andropogon chinensis</td>
<td>SM 3110</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Andropogon hallii</td>
<td>PI 477973</td>
<td>-</td>
<td>USDA</td>
</tr>
<tr>
<td>Apluda mutica</td>
<td>PI 271556</td>
<td>India</td>
<td>USDA</td>
</tr>
<tr>
<td>Capillipedium parviflorum</td>
<td>PI 301782</td>
<td>India</td>
<td>USDA</td>
</tr>
<tr>
<td>Chionachne koenigii</td>
<td>RS 97-18</td>
<td>Dharwai</td>
<td>EAK</td>
</tr>
<tr>
<td>Chrysopogon fulvus</td>
<td>PI 215587</td>
<td>India</td>
<td>USDA</td>
</tr>
<tr>
<td>Coelorachis selloana</td>
<td>SM 3059</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Coix gastenii</td>
<td>SJ 8199</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Coix lacryma-jobi</td>
<td>PI 324509</td>
<td>Brazil</td>
<td>EAK</td>
</tr>
<tr>
<td>Cymbopogon citratus</td>
<td>JB 0108</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Dichanthium aristatum</td>
<td>PI 301994</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Diheteropogon amplexens</td>
<td>RF 1819</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Elionurus muticus</td>
<td>SJ 5865</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Heteropogon contortus</td>
<td>PI 271177</td>
<td>Poland</td>
<td>USDA</td>
</tr>
<tr>
<td>Hyparrhenia hirta</td>
<td>PI 196827</td>
<td>Ethiopia</td>
<td>USDA</td>
</tr>
<tr>
<td>Hyparrhenia rufa</td>
<td>PI 203416</td>
<td>United Kingdom</td>
<td>USDA</td>
</tr>
<tr>
<td>Miscanthus sinensis</td>
<td>-</td>
<td>-</td>
<td>EV</td>
</tr>
<tr>
<td>Oxyrhachis gracillima</td>
<td>SM 3092</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Phacelurus ssp.</td>
<td>SM 3100</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Rottboellia selloana</td>
<td>PI 404763</td>
<td>Uruguay</td>
<td>EAK</td>
</tr>
<tr>
<td>Schizachyrium scoparium</td>
<td>EAK V46</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>B.Tx623</td>
<td>United States</td>
<td>EV</td>
</tr>
<tr>
<td>Themeda triaudra</td>
<td>RF 4599</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Trachypogon spicatus</td>
<td>RF 4680</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Tripsacum dactyloides</td>
<td>PI 421612</td>
<td>United States</td>
<td>USDA</td>
</tr>
<tr>
<td>Urelytrum digitatum</td>
<td>SM 3109</td>
<td>-</td>
<td>EAK</td>
</tr>
<tr>
<td>Zea mays</td>
<td>B73</td>
<td>United States</td>
<td>EV</td>
</tr>
</tbody>
</table>

---

*a EAK, Elizabeth A. Kellogg; EV, Erik Vollbrecht; and USDA, United States Department of Agriculture*
TABLE 2

Summary of gene content of the *ra1* region in Rice, Sorghum, and Maize

<table>
<thead>
<tr>
<th>Species</th>
<th>Chromosome</th>
<th>Length (Kb)</th>
<th>Average length (Kb/gene)</th>
<th>Collinear genes</th>
<th>Noncollinear genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice</td>
<td>9</td>
<td>247</td>
<td>2.0</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>Sorghum</td>
<td>2</td>
<td>226</td>
<td>2.3</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Maize</td>
<td>2</td>
<td>457</td>
<td>2.7</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Maize</td>
<td>7</td>
<td>1143</td>
<td>3.7</td>
<td>16</td>
<td>1</td>
</tr>
</tbody>
</table>
### TABLE 3
Molecular evolution of *ra*1 orthologs

<table>
<thead>
<tr>
<th>Taxon</th>
<th>ENC&lt;sup&gt;a&lt;/sup&gt;</th>
<th>GC&lt;sup&gt;b&lt;/sup&gt;</th>
<th>GC3&lt;sup&gt;c&lt;/sup&gt;</th>
<th>d&lt;sub&gt;s&lt;/sub&gt;&lt;sup&gt;d&lt;/sup&gt;</th>
<th>d&lt;sub&gt;N&lt;/sub&gt;&lt;sup&gt;e&lt;/sup&gt;</th>
<th>d&lt;sub&gt;N&lt;/sub&gt;/d&lt;sub&gt;s&lt;/sub&gt;&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panicoideae, Andropogoneae:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Andropogon chinensis</em></td>
<td>49.69</td>
<td>0.60</td>
<td>0.68</td>
<td>0.15</td>
<td>0.05</td>
<td>0.36</td>
</tr>
<tr>
<td><em>Andropogon hallii</em></td>
<td>49.20</td>
<td>0.60</td>
<td>0.68</td>
<td>0.12</td>
<td>0.04</td>
<td>0.37</td>
</tr>
<tr>
<td><em>Aphula mutica</em></td>
<td>49.03</td>
<td>0.60</td>
<td>0.68</td>
<td>0.15</td>
<td>0.05</td>
<td>0.30</td>
</tr>
<tr>
<td><em>Capillipedium parviflorum</em></td>
<td>51.78</td>
<td>0.60</td>
<td>0.69</td>
<td>0.14</td>
<td>0.06</td>
<td>0.42</td>
</tr>
<tr>
<td><em>Chionachne koenigii</em></td>
<td>51.46</td>
<td>0.60</td>
<td>0.69</td>
<td>0.13</td>
<td>0.08</td>
<td>0.62</td>
</tr>
<tr>
<td><em>Chrysopogon ssp.</em></td>
<td>54.77</td>
<td>0.60</td>
<td>0.67</td>
<td>0.11</td>
<td>0.04</td>
<td>0.42</td>
</tr>
<tr>
<td><em>Coelorachis selloana</em></td>
<td>50.98</td>
<td>0.59</td>
<td>0.69</td>
<td>0.14</td>
<td>0.06</td>
<td>0.41</td>
</tr>
<tr>
<td><em>Coix gasterii</em></td>
<td>55.49</td>
<td>0.59</td>
<td>0.68</td>
<td>0.17</td>
<td>0.07</td>
<td>0.43</td>
</tr>
<tr>
<td><em>Coix lacryma-jobi</em></td>
<td>54.26</td>
<td>0.58</td>
<td>0.65</td>
<td>0.16</td>
<td>0.08</td>
<td>0.52</td>
</tr>
<tr>
<td><em>Cymbopogon citratus</em></td>
<td>52.41</td>
<td>0.59</td>
<td>0.70</td>
<td>0.15</td>
<td>0.06</td>
<td>0.41</td>
</tr>
<tr>
<td><em>Dichanthium aristatum</em></td>
<td>52.90</td>
<td>0.60</td>
<td>0.66</td>
<td>0.15</td>
<td>0.07</td>
<td>0.51</td>
</tr>
<tr>
<td><em>Diheteropogon ssp.</em></td>
<td>51.06</td>
<td>0.59</td>
<td>0.67</td>
<td>0.17</td>
<td>0.05</td>
<td>0.32</td>
</tr>
<tr>
<td><em>Elionurus maticus</em></td>
<td>52.60</td>
<td>0.59</td>
<td>0.69</td>
<td>0.19</td>
<td>0.06</td>
<td>0.35</td>
</tr>
<tr>
<td><em>Heteropogon contortus</em></td>
<td>47.83</td>
<td>0.60</td>
<td>0.69</td>
<td>0.15</td>
<td>0.05</td>
<td>0.37</td>
</tr>
<tr>
<td><em>Hyparrhenia hirta</em></td>
<td>48.63</td>
<td>0.58</td>
<td>0.67</td>
<td>0.11</td>
<td>0.06</td>
<td>0.64</td>
</tr>
<tr>
<td><em>Hyparrhenia rufa</em></td>
<td>49.70</td>
<td>0.58</td>
<td>0.67</td>
<td>0.12</td>
<td>0.07</td>
<td>0.64</td>
</tr>
<tr>
<td><em>Miscanthus sinensis</em></td>
<td>52.72</td>
<td>0.60</td>
<td>0.69</td>
<td>0.12</td>
<td>0.05</td>
<td>0.49</td>
</tr>
<tr>
<td><em>Oxyrhachis gracillima</em></td>
<td>44.43</td>
<td>0.64</td>
<td>0.76</td>
<td>0.18</td>
<td>0.07</td>
<td>0.39</td>
</tr>
<tr>
<td><em>Phacelurus ssp.</em></td>
<td>53.79</td>
<td>0.58</td>
<td>0.66</td>
<td>0.13</td>
<td>0.06</td>
<td>0.44</td>
</tr>
<tr>
<td><em>Rottboellia selloana</em></td>
<td>53.65</td>
<td>0.59</td>
<td>0.67</td>
<td>0.14</td>
<td>0.05</td>
<td>0.39</td>
</tr>
<tr>
<td><em>Sorghum bicolor</em></td>
<td>47.33</td>
<td>0.61</td>
<td>0.72</td>
<td>0.15</td>
<td>0.06</td>
<td>0.39</td>
</tr>
<tr>
<td><em>Themeda triandra</em></td>
<td>50.05</td>
<td>0.60</td>
<td>0.70</td>
<td>0.13</td>
<td>0.06</td>
<td>0.57</td>
</tr>
<tr>
<td><em>Trachypogon spicatus</em></td>
<td>52.25</td>
<td>0.59</td>
<td>0.67</td>
<td>0.18</td>
<td>0.06</td>
<td>0.30</td>
</tr>
<tr>
<td><em>Tripsacum dactyloides</em></td>
<td>60.78</td>
<td>0.57</td>
<td>0.66</td>
<td>0.20</td>
<td>0.07</td>
<td>0.35</td>
</tr>
<tr>
<td><em>Urelytrum digitatum</em></td>
<td>53.63</td>
<td>0.60</td>
<td>0.67</td>
<td>0.16</td>
<td>0.06</td>
<td>0.40</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>61.00</td>
<td>0.57</td>
<td>0.65</td>
<td>0.22</td>
<td>0.07</td>
<td>0.32</td>
</tr>
<tr>
<td>Mean</td>
<td>51.98</td>
<td>0.59</td>
<td>0.68</td>
<td>0.15</td>
<td>0.06</td>
<td>0.43</td>
</tr>
</tbody>
</table>

<sup>a</sup> Effective number of codons  
<sup>b</sup> Total GC content  
<sup>c</sup> GC content at third position  
<sup>d</sup> Synonymous substitutions  
<sup>e</sup> Nonsynonymous substitutions  
<sup>f</sup> Ratio of nonsynonymous and synonymous substitutions
FIGURE LEGENDS

Figure 1. Andropogoneae inflorescence architectures. The inflorescences of *Rottboellia selloana* (arrowed) (A), maize (B), *Coix lacryma-jobi* (C), and sorghum (D) represent the diversity of branching architectures within the tribe Andropogoneae (Poaceae). The summarized phylogeny of grasses (E) shows the relationships of the Andropogoneae with other grasses. The phylogeny is summarized from (GPWG 2000; GPWG 2001; Sanchez-Ken et al. 2007).

Figure 2. Alignment of the *ra1* region in rice, sorghum, and maize. Gene sizes are calculated from start codon to stop codon and the vertical bars connect conserved genes. Collinear and noncollinear genes are represented by different colors. Synteny between maize, sorghum, and rice shows *ra1* is not present in rice or in a homeologous region on maize chromosome 2. The region in rice has two large sections that are not collinear with sorghum and maize (filled dark genes), one of which coincides the hypothesized location of *ra1*.

Figure 3. Conserved sequence motifs and sliding window analysis of nonsynonymous to synonymous substitution ratios (*dN/dS*). A) An eShadow (Ovcharenko, Boffelli, and Loots 2004) image from an alignment of *ra1* orthologs (C). (B) Newly indentified conserved motifs include an additional EAR repression motif (yellow box), a C-terminal motif (red box, far right), and an upstream 5’ noncoding sequence motif (red box, far left). Known sequence motifs including the C2H2 zinc finger DNA-binding motif and the C-terminal EAR repression motif were also highly conserved. D) A sliding window of synonymous and nonsynonymous ratios (*dN/dS*) shows potential patterns of positive and purifying selection across the gene. Window size = 18 bp, step size = 3 bp.

Figure 4. *ra1* gene phylogeny. A *ra1* gene tree was constructed using the nucleotide alignment of the combined immediate 5’ noncoding and coding sequence (~1350 bp total). Posterior probabilities (first number) and bootstrap support values (second number) are listed for each node.
Figure 5: *ral* duplication in sorghum. A) Structure of the *ral* tandem duplication in all cultivated varieties. Colors represent duplicate sequence as follows: For noncoding sequences, green = conserved 5’ element, orange = duplicated 3’ sequence, and black = other noncoding sequence. Within the coding sequence, black = C2H2 zinc finger DNA-binding motif, red = EAR repression motifs, and blue = other coding sequence. The light grey box represents out of frame coding sequence in the upstream copy. B) The upstream gene copy in some wild sorghums have different frameshift mutations leading to truncated proteins of varying sizes. Examples listed here include SbW6 (*S. bicolor* ssp. *virgatum*), SbW26 (*S. bicolor* ssp. *arundinaceum*), and SbW32 (*S. propinquum*). C) Recombinate gene copies contain small tracks of converted sequence in the upstream copy (red) and the downstream copy (green).

Figure 6: Phylogeny of *ral* sequences in sorghum. The phylogeny includes cultivated varieties and wild *S. bicolor* subspecies. Letters and numbers in each taxon represent the following: Sb = *Sorghum bicolor*, numbers = accession (Supplementary Table 1), C = cultivated, W = wild, H = hybrid, U = upstream gene copy, and D = downstream gene copy. *S. propinquum* was included as a close outgroup and maize as a more distant outgroup. Posterior probabilities and bootstrap support values (second value) are included were applicable for each node. A complete phylogeny encompassing all *ral*-like sequences, including the putative recombinant copies, for all *Sorghum* accessions is located in the Supplementary Data.
Figure 1 Sigmon et al.

Diagram showing the relationships of different grass clades and species, with labels for each clade and example species.
Figure 3 Sigmon et al.
Figure 4 Sigmon et al.
Figure 5 Sigmon et al.
Figure 6. Sigmon et al.
CHAPTER 4. GENOME-WIDE COMPARATIVE ANALYSIS REVEALS THE IMPORTANCE OF LINEAGE-SPECIFIC EXPANSION IN THE EVOLUTION OF THE EPF SUBFAMILY IN PLANTS

A manuscript to be submitted to BMC Genomics

Brandi Sigmon, Rebecca Weeks, and Erik Vollbrecht

SUMMARY

Background
Members of the EPF subfamily of the C2H2-type zinc finger proteins in plants have been shown to function as transcriptional regulators involved mainly in developmental processes and abiotic stress response. Most EPF genes encode an EAR motif near their C-terminus, a potent repressor found in proteins acting in diverse pathways such as auxin signaling and a variety of cellular and developmental processes.

Results
Through database searches we exhaustively identified and characterized EPF genes in seven plant species to better understand the evolution of this important plant gene family. Through a genome-wide comparative approach, we found key differences in the frequency of certain EPF gene structure types due to lineage-specific expansions. We found notable species-specific differences in the GIS clade, in EPF genes with multiple EAR repression motifs, and in EPF genes with multiple zinc finger DNA-binding motifs. The observed expansion differences are due to both whole genome duplication events as well as single gene tandem duplications. Key changes in the gene family such as the acquisition of EAR repression motifs and the presence of multiple EAR motifs per EPF gene correlate to major plant life history events.
Conclusions
The patterns of expansion due to duplication and the subsequent retention or loss of duplicates leading to lineage-specific differences have shaped the molecular evolution of the EPF gene subfamily. Given the known EPF gene functions in plant development and abiotic stress response, we suggest the gene family has played a key role in the morphological changes plants needed to adapt to new environments during their evolution. Finally, EPF genes may be especially potent transcriptional repressors where the numbers and positions of EAR motifs are of functional consequence, which may also have been particularly important to the evolution of grasses.

BACKGROUND
Multigene families evolve largely due to expansion and diversification. Expansion occurs primarily due to duplications on various scales including single gene, segmental and whole genome duplication (WGD) events. Along plant lineages, multiple genome duplication events have occurred resulting in rapid expansion and diversification of gene families (Adams and Wendel 2005). Following duplication events, retained duplicates can diversify by evolving new functions (neofunctionalization) or by partitioning original gene functions (subfunctionalization) (Force et al. 1999; Lynch and Conery 2000). Recently, it has been shown that transcription factor families typically have larger expansion rates in plants than animals, due to greater duplicability and higher retention rates (Shiu, Shih, and Li 2005). Major plant life history events such as land colonization and the acquisition of vascularity were achieved through phenotypic evolution. Although the genetic basis behind this phenotypic evolution is still unknown, it is hypothesized to be due in large part to the activity of genes encoding transcription factors. With the recent advent of plant comparative genomics, it is possible to investigate the correlation of the evolution of transcription factor multigene families with the major life history events in plants. Many recent studies have focused on genome-wide characterization and comparative genomics of several transcription factor families (Wu et al. 2005; Nakano et al. 2006; Navaud et al. 2007). However, most of these genomic studies have limited scopes focusing on differences between monocots and eudicots.
C2H2 zinc finger transcription factors comprise one of the largest transcription factor families in eukaryotes. In plants, members of this extensive gene family have been identified and categorized in Arabidopsis (Englbrecht, Schoof, and Bohm 2004) and rice (Agarwal et al. 2007). One subgroup initially named EPFs in petunia (Takatsuji et al. 1992; Takatsuji, Nakamura, and Katsumoto 1994) and variously referred to as the C1 family in Arabidopsis (Englbrecht, Schoof, and Bohm 2004) and Q-type in rice (Agarwal et al. 2007), constitutes about one-third of C2H2 zinc finger proteins in plants. These proteins consist of a single or multiple tandemly positioned zinc finger DNA-binding motif(s) separated by linker sequences of variable, often short, length. These zinc finger motifs have been shown to be capable of interacting with DNA and contain a plant-specific QALGGH alpha-helical motif, which is considered a defining sequence characteristic of this group of transcription factors (Kubo et al. 1998). In the structure solved for an Arabidopsis C2H2 protein, the QALGGH region of that alpha helix makes contact with base pairs in the major groove of DNA (Dathan et al. 2002).

Functionally, most characterized EPF proteins are putative transcriptional repressors involved in a number of developmental processes and stress responses. Their function as repressors is hypothetically due to the presence of a short, approximately 5-7 amino acid motif that is usually C-terminally located, called the EAR (ERF-associated amphiphillic repression) motif. This potent repressor motif was first described in AP2/ERF (Apetala2/Ethylene Response Factor) and C2H2 zinc finger proteins (Ohta et al. 2001; Nakano et al. 2006), but is also be present in a total of 21 different transcriptional regulator gene families (Kagale, Links, and Rozwadowski 2010). The potency of the EAR motif in transcriptional repression has been demonstrated by its ability to convert transcriptional activator proteins into repressors in vivo (Ohta et al. 2001).

Although no in planta function has been identified for most EPF genes, within the last few years many have been described in the plant model organisms Arabidopsis, maize, and rice. In maize, the only characterized EPF gene is ramosa1, which encodes an EPF with a single zinc finger that controls the inflorescence architecture of the tassel and ear (Vollbrecht et al. 2005). Currently in rice, several EPF genes have been functionally characterized. ZFP15 a two-finger EPF, is hypothesized to have a role in spike development
(Huang, Wang, and Zhang 2005) whereas the single zinc finger gene *PROSTATE GROWTH1* is responsible for the erect plant architecture of cultivated rice due to its role in altering tiller angle and number (Jin et al. 2008; Tan et al. 2008). Most functional characterization of EPFs to date has been carried out in Arabidopsis. Some two-zinc finger EPFs have been implicated in abiotic and biotic stress response and root development (Sakamoto et al. 2000; Sakamoto et al. 2004; Ciftci-Yilmaz et al. 2007; Devaiah, Nagarajan, and Raghothama 2007). Single zinc finger EPFs have been shown to be involved in various developmental roles, most notably in organ initiation and development. For example, *SUPERMAN* maintains floral whorl boundaries, thus playing a major role in floral development (Sakai, Medrano, and Meyerowitz 1995; Hiratsu et al. 2002). Also, the *SUPERMAN*-like homolog named *RABBIT EARS* is involved in petal primordium initiation and the development of petal shape (Takeda, Matsumoto, and Okada 2004). The partially redundant close paralogs *NUBBIN* and *JAGGED* define and regulate stamen and carpel growth and development (Dinneny, Weigel, and Yanofsky 2006) whereas *GIS, GIS2*, and *ZFP8* have essential roles in shoot maturation and epidermal differentiation (Gan et al. 2006; Gan et al. 2007). An EPF gene has also been implicated in stamen abscission in Arabidopsis (Cai and Lashbrook 2008).

Here we present the first comprehensive identification and characterization of EPF genes for poplar (*Populus trichocarpa*), maize (*Zea mays mays*), sorghum (*Sorghum bicolor*), *Selaginella* (*Selaginella moellendorffii*), and moss (*Physcomitrella patens*) including analyses of gene structure, phylogenetic relationships, and insights into their evolution through rapid expansion and diversification. We have combined these data with previously reported data for Arabidopsis (*Arabidopsis thaliana*) (Englbrecht, Schoof, and Bohm 2004) and rice (*Oryza sativa*) (Agarwal et al. 2007), and through a genome-wide comparative approach, we have discovered key species-specific differences in gene structure due to lineage-specific expansions. Most notably we have found grass EPF genes commonly encode multiple EAR repression motifs whereas the non-grass species typically encode a single, C-terminally located EAR motif. The presence of multiple EAR-motifs has several important implications for gene function and evolution, which we discuss.
RESULTS

To date, EPF genes have been systematically identified and characterized for rice (Agarwal et al. 2007) and Arabidopsis (Englbrecht, Schoof, and Bohm 2004). Since it is a plant-specific gene family, we limited our search of databases to plants whose genomes have been completely sequenced or otherwise have substantial amounts of sequence available. For data mining we also selected plant genomes based on phylogenetic relationships among the major lineages of plants including grass angiosperms (rice, sorghum, and maize), eudicot angiosperms (Arabidopsis and poplar), Lycopodiophyta (*Selaginella*), Bryophyta (*P. patens*), and green algae (*C. reinhardtii*). Such a broad comparative sampling of plant genomes enables us to better understand how this protein family evolved and to possibly correlate the molecular evolution of EPFs to major plant life history events.

**Identification of EPF genes from public databases**

By mining publicly available databases (see methods) the number of putative EPFs we recovered numbered 91 in maize, 71 in sorghum, 74 in poplar, 15 in *Selaginella*, and 22 in the moss *P. patens* to combine with the previously reported 78 in rice (Agarwal et al. 2007) and 60 in Arabidopsis (Englbrecht, Schoof, and Bohm 2004). Table 1 summarizes the totals for each plant genome according to the number of zinc finger DNA-binding motifs. For all plant species, single zinc finger EPFs were the most prevalent, whereas EPFs with multiple tandemly positioned zinc finger motifs were less common. There is also a rough correlation between total EPF number and genome size. For instance, *Selaginella*’s genome is approximately the smallest (~100 Mb) in size (Little et al. 2007) and has the fewest EPFs whereas maize is the largest (~2500 Mb) in size (Schnable et al. 2009) and has the most EPFs.

To pinpoint the origin of the EPF family, we also searched the sequenced green alga genome, *C. reinhardtii*. Through a variety of BLAST searches we were unable to find any putative EPFs, suggesting the EPF family originated following the divergence of green algae from the rest of plants. However a broader, more comprehensive sampling of Charales, the closest relatives to land plants, is needed to precisely pinpoint the origins of the EPF gene family in plants.
Evolution of zinc finger DNA-binding motifs in EPFs

Previously, C2H2 zinc finger gene phylogenies were reconstructed using solely the amino acid sequence of the conserved zinc finger DNA-binding motif, resulting in phylogenies with poor resolution and no support due to insufficient informative sites (Englbrecht, Schoof, and Bohm 2004; Agarwal et al. 2007). The poor resolution of the amino acid sequence tree therefore hindered testing hypotheses about the origins and evolution of C2H2 zinc finger genes. These analyses also focused on Arabidopsis or rice, which further precluded inferences on the relevance of these genes to the evolution of land plants. To better understand the origins and evolution of the EPF gene subfamily we therefore attempted to reconstruct the phylogeny of the gene family using a larger, more gene comprehensive dataset and different amino acid and nucleotide models in conjunction with more taxa.

To reconstruct the phylogeny of EPF genes with a single zinc finger motif, we aligned the nucleotide and amino acid zinc finger motif sequence of all single zinc finger EPF genes for the following taxa: moss, Selaginella, Arabidopsis, poplar, rice, sorghum and maize. We first used an amino acid alignment for phylogenetic reconstruction using Bayesian analysis utilizing the WAG model (Whelan and Goldman 2001). These phylogenies resulted in a lack of resolution and poor support as seen in the neighbor-joining trees in our own analyses and in other previous studies (Englbrecht, Schoof, and Bohm 2004; Agarwal et al. 2007). Likewise, a nucleotide alignment of the conserved zinc finger motif, when used for Bayesian and parsimony analysis, produced phylogenies similar to those based on amino acid sequence. Although these phylogenies are not well supported, we were able to use them as a guide for more stringent reconstruction of specific clades of interest.

Despite the poor resolution of the phylogenies, it was evident that there are four distinct clades that are at least as old as the land plants (Figure 1a). Each of the four clades has at least one moss representative, which places single finger EPF gene origin to around 450 mya when embryophytes originated and began to diverge from one another. It is also evident from the phylogeny that some clades have preferentially expanded in higher plants. Clade 1-III and Clade 1-IV have expanded more than both Clades 1-I and 1-II, which are the more divergent clades with only a few eudicot members.
We also reconstructed a phylogeny of EPF genes that have two zinc finger DNA-binding motifs (Figure 1b). It has been hypothesized that these genes originated more recently than EPF genes with a single zinc finger DNA-binding motif (Englbrecht, Schoof, and Bohm 2004). In our genomic survey we found EPF genes with multiple zinc finger DNA-binding motifs to be few in lower plants, but numerous in higher plants (Table 1). The phylogeny shows clades specific to the eudicots (blue and green taxa, Figure 1b) and to the monocots (red and aqua taxa, Figure 1b, which suggests expansion occurred in the angiosperms following the split between the eudicots and monocots in a lineage-specific manner (Figure 1b). The presence of multiple zinc finger EPF genes in lower plants and their subsequent expansion suggests this EPF type may have originated in the common ancestor of land plants and afterwards expanded in the angiosperms in a lineage-specific manner. However, it is also possible that these genes originated more recently in the common ancestor of angiosperms and the few representatives in moss and Selaginella were gained individually.

**Mechanisms of EPF gene expansion**

Significant expansion of EPFs has likely occurred primarily due to large-scale duplication events, such as WGD and segmental duplication. It is also likely that smaller-scale duplication events, such as tandem gene duplication has played a major role in the expansion of EPF genes. We aimed to determine the role of these mechanisms has played in the expansion of the EPF subfamily by examining their genomic positions.

We defined tandem duplicates in plants with compact genomes (Arabidopsis, poplar, and rice) as duplicates less than 100 kb apart with fewer than ten nonhomologous spacer genes (Hanada et al. 2008). To study the pattern and compare the proportion of tandem duplication among species, we classified all EPF genes as either tandem or nontandem duplicates. Previously it was determined that the proportion of tandem duplicates in the Arabidopsis and rice genomes are strikingly similar (16% vs. 14%) (Rizzon, Ponger, and Gaut 2006), despite 150 my of divergence time. For the EPFs, in Arabidopsis 21.7% of genes are tandem duplicates, which is higher though not significantly (p-value = 0.209, chi-square test), than that observed in the Arabidopsis genome (Hanada et al. 2008) (Table 2). The
proportion of tandem duplicates in rice was significantly higher (p-value = 0.001, chi-square test) at 26.9% than the observed frequency in the rice genome (Hanada et al. 2008). However, for poplar, 13.5% of EPFs were tandem duplicates, which is similar to the previously observed values for Arabidopsis and rice genomes (Hanada et al. 2008). We also found a bias in EPF gene structure type (Table 2). EPF genes with multiple zinc finger DNA-binding motifs are more often positioned in tandem arrays than genes with a single zinc finger motif. In Arabidopsis five out of seven (71.4%) tandem duplicate gene pairs are multi-zinc finger EPF genes, which however, does not significantly from the expected frequency (p-value = 0.25, chi-square test). The trend was significant for rice where nine out of eleven (81.8%) pairs are multi-zinc finger EPF genes (p-value = 0.01, chi-square test) However, we did not see the same trend in the small sample size from poplar, where two out of five (40%) pairs are multi-zinc finger EPFs (p-value = 0.16, chi-square test). We also observed that the tandem duplicates tend to be in the same orientation (p-value = 0.05, chi-square test) for Arabidopsis, poplar, and rice (Table 2). For Arabidopsis, six out of seven (85.7%) pairs of tandem duplicates were in the same orientation, as well as, all five (100%) pairs in poplar. Also, of rice’s eleven pairs, eight (72.7%) are in the same orientation.

These data show that in addition to WGD, single gene tandem duplications have been an important mechanism of expansion, primarily for multi-zinc finger EPFs. The higher frequency of tandem duplicates for EPFs in Arabidopsis and rice versus the estimate for their respective genomes suggests a selective pressure for a greater generation rate and/or retention of those duplicates. The selective pressures of generation and retention of multi-zinc finger EPF genes may also be different between species, as evidenced by the differing proportions between Arabidopsis, poplar, and rice. Despite the differences in tandem duplication of EPF genes in these three genomes, these values are likely underestimates because of the inability to detect ancient tandem duplications due to subsequent genomic rearrangements.

**Identifying conserved motifs**

Transcription factors are known to be modular proteins wherein domains and motifs important to gene product function are highly conserved across species, but the sequence in
between these motifs is not conserved (Lukens and Doebley 2001). This knowledge enables the use of amino acid alignments in identifying new conserved motifs.

Two zinc finger EPFs
Previously identified conserved motifs in two zinc finger EPF genes include the “B-box” (putative nuclear localization signal), the “L-box” and “CLMLL” motifs (putative protein-protein interaction) (Englbrecht, Schoof, and Bohm 2004). These motifs are specific to EPFs with two zinc fingers and are positioned upstream of the zinc fingers (Figure 2a). When we looked for these conserved motifs in our dataset, we found no B-box, L-box or CLMLL motifs in either moss or Selaginella. This finding is not unexpected due to the low number of two zinc finger EPF genes in lower plants (Table 1). All three motifs were present in Arabidopsis, poplar, rice, maize, and sorghum EPFs. Comparative analysis of these eudicot and grass sequences revealed some lineage-specific differences in these two-finger EPF genes. First, two zinc finger EPFs lacking the B-box, L-box, and CLMLL motifs have greatly expanded in the grass lineage compared to the eudicots (Figure 2b,e); this gene structure comprises 76-86% of grass two zinc finger EPFs compared to 14-18% for the eudicots. Alternatively, the eudicots have a higher percentage of the other gene structures. In Arabidopsis 59% contain the CLMLL motif and 23% contain the B-box and L-box (Figure 2b,c). The trend appears opposite for poplar, which 21% of two finger EPFs have the CLMLL motif and 64% have the B-box and L-box (Figure 2b,d). These data suggest different patterns of expansion in the two-zinc finger EPFs between the grass and eudicot lineages as well as between different eudicot species. The relatively recent divergence of Arabidopsis and poplar, some 100 mya, connotes that significant (p-value = 0.02, chi-squared test) lineage-specific differences can occur within modest amounts of evolutionary time.

GIS clade
Using amino acid alignments from specific clades, we identified several novel, conserved motifs. First we found three conserved sequence motifs in a well-supported clade containing the functionally characterized Arabidopsis genes GLABROUS INFLORESCENCE STEMS (GIS), GLABROUS INFLORESCENCE STEMS2 (GIS2), and ZINC FINGER PROTEIN8
(ZFP8) (Gan et al. 2006; Gan et al. 2007), which will hereafter be called the GIS clade. This clade contains members from all sampled angiosperms, but homologous genes were not found in either moss or Selaginella. A search for GIS genes in the gymnosperm Pinus taeda (P. taeda), for outgroup purposes, revealed two clade members, which suggests the common ancestor of angiosperms and gymnosperms likely had two GIS genes. From an alignment of 24 proteins, we found two novel conserved sequences upstream of the zinc finger motif and a third motif between the zinc finger and the C-terminal EAR motif (Figure 3). Most GIS genes contain all of these motifs but some genes have lost part or all of one or more of these motifs, perhaps due to subfunctionalization following a duplication event. For example, OsEPF30 is missing the first motif and GIS is missing part of this motif whereas it is intact in all other members of the GIS clade. Unlike the first motif, the second novel motif is conserved across the entire clade. Finally, the third motif is missing in P. taeda, in GIS and GIS2 of Arabidopsis, as well as, in some of the grasses (Figure 3). This pattern of presence and absence suggests this motif may have been absent in the common ancestor of gymnosperms and later acquired in the angiosperms, then genes with this motif expanded in the grass lineage. The role of these amino acid sequence motifs remains unknown and a search of the PLACEv30.0 database (Higo et al. 1999) revealed no known transcription factor binding sites within the conserved nucleotide sequences of these genes. The phylogeny of the GIS clade suggests the common ancestor of grasses may have had two GIS genes, then acquired a third through a single gene duplication prior to the ancient WGD (Paterson, Bowers, and Chapman 2004). Of the three genes, two expanded from the ancient WGD whereas the third did not (Figure 4).

**EPFs with multiple EAR repression motifs**

In addition to the C2H2 DNA-binding zinc finger motif, the presence of an EAR repression motif is a defining signature for this gene family (Kubo et al. 1998; Englbrecht, Schoof, and Bohm 2004). To better understand the distribution of EAR motifs in EPF genes, we searched single zinc finger EPF protein sequences using the consensus sequence LxLxL to find putative EAR repression motifs. We detected the presence of multiple EAR repression motifs with the consensus sequence LxLxL, especially in the grasses. Previous surveys report only
C-terminally located EARs in EPF genes (Englbrecht, Schoof, and Bohm 2004; Agarwal et al. 2007). However, it has been recently shown that in several transcription factor gene families some genes may have multiple EAR motifs per gene (Kagale, Links, and Rozwadowski 2010). We observed, in grasses especially, that EAR motifs are positioned in several locations including near the N-terminus and just after the zinc finger in addition to their most common position at the C-terminus (Figure 5a). Although there is no functional data on alternatively positioned EAR motifs in EPF genes, EARs are sometimes not C-terminally located in other gene families (Kagale, Links, and Rozwadowski 2010). Also, the amino acids between the leucines are almost always different between the two EARs encoded by any one gene. However, like in the canonical C-terminal EARs, these motifs primarily contain charged and/or polar amino acids with functional groups capable of hydrogen bonding (Figure 6). This suggests these additional EARs could interact with target proteins in a similar way, although the variability suggests the notion that within a gene the targets and/or strength of the interactions may vary at different EARs.

**Position of EAR motifs**

The number and position of EARs differed depending on the species (Figure 5b). For the moss *P. patens*, of the total 18 single-finger EPFs, ten (55%) have a single C-terminally located EAR whereas eight (45%) have no EAR. Since it is known that the presence of an EAR motif confers a repressive function (Ohta et al. 2001), EPF genes with an EAR motif likely function as transcriptional repressors, but those without an EAR may have some other function. EPFs with no evident EAR motifs are present, though not as prevalently, in all other species examined. This observation suggests either some EARs were lost in mosses or that after the divergence of vascular plants from their common ancestor with mosses, EPF genes with EARs became more prevalent through expansion, for example due to preferential retention of duplicates of those with C-terminally located EAR motifs. The latter scenario seems more likely given the ubiquity of this gene structure throughout plants.

*Selaginella’s* proportion of single C-terminally located EAR-containing genes is similar to *Physcomitrella’s* (58% vs. 56%), but *Selaginella* has fewer EPF genes with no EAR motifs (17%) (Figure 5b). Unlike in moss, 25% of *Selaginella* EPFs have two putative
EAR repression motifs. In addition to the typical C-terminally located EAR, a second EAR sequence is located near the N-terminus, prior to the zinc finger. All other sampled plants also have EPF genes with multiple EAR repression motifs though at differing frequencies and positions (Figure 5b). In eudicots, 7% of Arabidopsis and 10% of poplar EPF genes have two EAR motifs. Most of these genes have a similar EAR positioning as Selaginella, but a couple have two C-terminally located EARs. Genes that encode multiple EAR motifs are most prevalent in grasses where 43% of rice, 49% of sorghum, and 54% of maize EPFs have two EAR motifs (Figure 5b). Like eudicots, most of these have one N- and one C-terminally located EAR motif, although a small percentage have two C-terminally located EARs. A plausible explanation for the significantly higher proportion of multiple EAR-containing EPF genes in grasses is a rapid expansion through duplication events and retention of such genes along the grass lineage after its divergence from the eudicot lineage.

To address this hypothesis we attempted to reconstruct the evolutionary relationships between EPF proteins using phylogenetic analysis. However reconstruction was hindered due to the lack of conserved nucleotide and amino acid sequence as reported previously (Englbrecht, Schoof, and Bohm 2004; Agarwal et al. 2007). Neither parsimony nor Bayesian methods produced a resolved phylogeny since only the sequence of the zinc finger motif is conserved enough to produce a reliable alignment. Therefore, we used a neighbor-joining clustering algorithm of 72 bp nucleotide alignments of the conserved zinc finger motifs to create grass and eudicot phylograms to demonstrate basic relationships between groups of EPFs (Figure 5c,d). From these phylograms, several grass clades show evidence of expansion of EPFs with multiple EAR motifs whereas those with only one EAR motif have expanded less so. In eudicots, the opposite was observed where clades with one C-terminally located EAR motif have greatly expanded. These observations suggest the prevalence of these motifs in grasses is likely due to rapid expansion through duplication along the grass lineage following their divergence from the eudicots.

The phylogenetic distribution of EPFs as well as the number and pattern of EAR motifs correlate with major plant life history events. First, the origin of the EPF subfamily is early in the plant lineage, after divergence from C. reinhardtii and possibly Chara, the closest relative to land plants. Strikingly, there may have been a shift or specialization in
function after plants colonized land since a significant proportion may not function as transcriptional repressors in mosses due to the absence of EAR repression motifs. Furthermore, the first instance of multiple EAR motifs per EPF gene corresponds to the innovation of a vascular system in plants, since the dual-EAR gene structure is present in Selaginella and all angiosperms sampled in this study. Finally, a significant expansion of EPFs with multiple EAR motifs occurs along the grass lineage, which is likely due largely to the WGD ~70 mya (Paterson, Bowers, and Chapman 2004) followed by retention of these duplicates.

Amino acid composition of EAR motifs

In a recent study that analyzed the amino acid composition of EAR motifs in 21 different gene families, a high incidence of charged and polar amino acids in and around EAR motifs was noted (Kagale, Links, and Rozwadowski 2010). When we looked at the amino acid composition of EAR motifs in EPF genes, we independently found that residues between the conserved leucines are almost always charged and/or polar (Figure 6). In EPF genes, C-terminal EAR motifs typically contain four leucines with three intervening amino acids \( (L_x1L_x2L_x3L) \). For these intervening residues, \( X_1 \) is usually acidic whereas \( X_2 \) tends to be polar and \( X_3 \) basic (Figure 6a,b). This pattern of composition is conserved across all species. The amino acid composition of the N-terminal EAR motifs, found mostly in grasses, differs somewhat from the C-terminal EARs (Figure 6c). All the intervening amino acids are polar, but are not necessarily charged. It has been suggested that the amino acid sequence context around the EAR motifs may play a role in exposing the motif surface(s) to potential interactors (Kagale, Links, and Rozwadowski 2010). Also the presence of amino acids known to be important in posttranslational modification such as serine, threonine, and leucine are common in EAR motifs (Kagale, Links, and Rozwadowski 2010). The EAR motifs in EPF genes show some variation in the presence of these three amino acids, which may suggest potential differences in EAR function. For instance, N-terminal EARs preferably have a serine or threonine in the \( X_3 \) position whereas the C-terminal EAR tends to have a basic amino acid in this position. Finally, the number of leucines may also be important when
considering posttranslational control as they also vary in number from three to six (data not
shown).

DISCUSSION

Origins of EPF gene subfamily
Through database searches of sequenced plant genomes we have determined the EPF gene
family originated at least 450 mya since they are present in the moss _P. patens_, but
apparently absent from the green alga _C. reinhardtii_. To more precisely pinpoint the origins
of this gene family a search should be performed in _Chara_ (Charales), a close outgroup to all
land plants. If no EPF genes are present in _Chara_, then the origin of this gene family would
coincide with the colonization of land by plants, a major life history event in plant evolution.
This result would be somewhat unusual as only about 14% of gene families are unique to
land plants (Jody Banks, personal communication), as it is more typical for gene families to
originate with the plant kingdom. Since some known functions of EPFs are in plant
developmental processes and response to environmental stimuli their uniqueness to land
plants suggests the EPFs may have been important in creating some of the innovations
necessary to colonize land. It is also likely EPF genes played an important role in more
recent plant evolution as plants acquired new developmental structures and complexities,
such as flowers, that enabled them to thrive in almost all environments. Recently, evidence
has surfaced that some EPFs in crop plants have been targeted by artificial selection during
their domestication in order to increase their agronomic value to humans. The EPF gene
_PROSTRATE GROWTH1_ determines tiller angle in rice, a key trait manipulated during rice
domestication (Jin et al. 2008; Tan et al. 2008). Also, _ramosa1_ is a maize EPF gene that has
been shown to affect the packing of kernels on the maize ear and is also a key domestication
locus candidate (Sigmon and Vollbrecht 2010). Given the role EPFs play in plant
development and in stress response, they may be of great agronomic importance and could be
utilized in the future for the improvement of all kinds of plants.

There has been some debate in the literature on whether single zinc finger or multiple
zinc finger proteins arose first (Kubo et al. 1998; Englbrecht, Schoof, and Bohm 2004). From
our comparative genomic analysis we found that multi-finger EPFs are of potentially ancient
origins given their presence in lower land plants, but appear to have expanded more recently since there are lineage specific differences in higher plants (Figure 1a). Because multi-finger EPFs number so much fewer in lower plants it seems more probable that they are of more recent origin than the more diverse and numerous single zinc finger EPFs. However, further evidence is required to determine the mechanism of their origin.

**EPF expansion and diversification**

It is apparent that duplication events have played a major role in the expansion of the EPF gene subfamily in plants. Comparative genomic analysis focusing on three clades: the GIS clade, a two EAR clade, and the multi-finger EPF genes, has revealed that the expansion of EPFs has occurred through both large-scale WGD and small-scale tandem duplication events. Following duplication events, it is assumed most duplicates are lost (Lynch and Conery 2000). However, it has been shown that frequently duplicates of transcription factors are preferentially retained in plants (Shiu, Shih, and Li 2005). Gene copies may be retained following duplication through partitioning of original functions (subfunctionalization) or through acquiring a new function (neofunctionalization). Indeed, evidence of this may exist in the GIS clade where several genes appear to have lost conserved sequence motifs (Figure 3). Experimental evidence on such a gene pair would test the hypotheses that subfunctionalization or neofunctionalization has occurred.

EPF genes are prime candidates to be preferentially retained following duplication events given their small size and simple modular structure. In particular, the highly conserved alpha helix QALGGH of the zinc finger motif is capable of sequence-specific recognition and interaction with just three nucleotides, AGT (Takatsuji, Nakamura, and Katsumoto 1994; Takatsuji and Matsumoto 1996; Kubo et al. 1998). Given the large number of binding opportunities in a genome, a duplicated EPF gene may be more easily co-opted for new gene functions than a duplicated transcription factor that recognizes a longer sequence motif. Also the regions outside of the conserved zinc finger and EAR motifs appear to diversify very quickly, which may also influence their rate of retention following duplication.

In this study we focused on three particular clades, which appear to have expanded substantially in at least one lineage. By comparing genomes and inspecting phylogenies we
found both WGD and single gene duplication events have played a role in the expansion of the gene family, which has presumably led to some of the lineage-specific differences observed. The seemingly imminent availability of more total genome sequences will better pinpoint the timing of major expansions in the EPF subfamily.

*Expansion of EPF genes with two tandem zinc finger DNA-binding motifs*

Through comparative genomics and phylogenetics we found evidence for lineage-specific differences between Arabidopsis, poplar, and grasses in the expansion of EPF genes containing two tandemly positioned zinc finger motifs (Figure 2). The species-specific expansion in Arabidopsis and poplar suggests these expansion events occurred primarily after they diverged from one another. On the other hand, because all grasses have similar frequencies of the three main types of these EPF genes, the primary expansion in the grasses most likely occurred before rice, maize, and sorghum diverged from one another. Therefore the probable mechanism of expansion would have been the WGD event in the grass lineage ~70 mya (Paterson, Bowers, and Chapman 2004). Also, the low number of multi-finger EPF genes in moss and Selaginella suggests the common ancestor of all angiosperms also had few of these genes and they later expanded in a lineage-specific manner.

By examining the genomic positions of the multi-finger EPF genes, it is evident that tandem duplication has also played a major role in the expansion of these genes, especially in rice (Table 2). The tandem duplicates also tend to be in the same orientation (Table 2), which suggests unequal crossover as a driving mechanism in producing them. Tandem duplicates may also be produced by intrachromosomal recombination between direct and indirect repeats, which would produce gene duplicates in opposite orientations (Rizzon, Ponger, and Gaut 2006). Since we found fewer duplicates in opposite orientations, this mechanism hypothetically has played a lesser role in the expansion of these genes.

In recent studies it has been observed that gene families involved in the response to environmental stimuli are more likely to have expanded in a lineage-specific manner through tandem duplication (Rizzon, Ponger, and Gaut 2006; Hanada et al. 2008). Tandem duplication occurs more frequently than WGD and thus provides a mechanism for response to rapidly changing environments (Hanada et al. 2008). Since many multi-finger EPFs are
involved in stress response, these genes may have played a role in the adaptive evolution of plants in response to changing environments.

GIS clade expansion
From our genomic surveys of diverse plant species, we found an expanded clade of EPFs with three novel conserved sequence motifs in angiosperm species (Figures 3, 4), which we named the GIS clade. To test if these EPFs are confined to angiosperms, we BLASTed the Pinus taeda EST database and found two GIS clade members indicating their origin predates the common ancestor of the angiosperms. The presence of GIS clade genes in angiosperms and gymnosperms and their absence in Selaginella and moss suggests the ancestor of this specific clade of EPF genes was present in the common ancestor of all angiosperms and gymnosperms, which coincides with the developmental innovation of seeds in plants. Following the origin of this clade of EPF genes, it then expanded through duplication and retention in the grass lineage, but not in the eudicot lineage (Figure 4). It is also evident some duplicates have possibly subfunctionalized following duplication since not all conserved motifs are present in all genes. These genes in Arabidopsis have been shown to be involved in various aspects of shoot maturation including most notably epidermal differentiation (Gan et al. 2006; Gan et al. 2007), but the role these conserved motifs play in these functions remains unknown. Outside of Arabidopsis, the functions of GIS clade genes have not been investigated.

Two-EAR EPF expansion
We found that EPF genes with multiple EAR motifs have significantly expanded in grasses at some point after their divergence from eudicots. To pinpoint the timing, a survey of the genomic sequence available for the gymnosperm Pinus taeda and the basal eudicot Aquilegia formosa would be useful. A comprehensive survey of Pinus could provide evidence of whether a major expansion has occurred prior to or following the divergence of gymnosperms and angiosperms. Aquilegia is a member of the basal-most eudicot clade and thus can be used in comparative studies to determine when important genomic changes occurred leading to the major diversification of the flowering plants (Kramer 2009). In
addition to providing insight into the eudicot ancestor, the phylogenetic position of *Aquilegia* is roughly intermediate between the highly divergent model systems in the monocot grasses and core eudicots (Kramer 2009). This allows *Aquilegia* to provide an important third data point in genomic studies comparing these groups. As related to the apparent expansion of EPF genes with multiple EARs in the grass lineage (Figure 5), we found only a couple such genes in the *Pinus* and *Aquilegia* EST databases, suggesting they may not be prevalent outside the monocots (data not shown). These results are inconsistent with expansion the of two-EAR single finger EPF transcription factors in the common ancestor of all angiosperms or in the common ancestor of gymnosperms and angiosperms with a subsequent loss in the eudicots. Instead these data support expansion in the grass lineage after the grasses and eudicots diverged from each other. We suggest this expansion was first due to a WGD event at the base of the grass lineage (Paterson, Bowers, and Chapman 2004) and preferential retention of a majority of duplicates, followed by further expansion through single gene duplications. Complete sequencing of the *Aquilegia* genome will further inform these inferences regarding the timing of the expansion of EPF genes with multiple EAR repression motifs.

**Role of EAR repression motifs in EPF function**

EARs are regulatory motifs involved in transcriptional repression (Ulmasov et al. 1997; Ohta et al. 2001; Hiratsu et al. 2002; Tiwari, Hagen, and Guilfoyle 2004). In addition to the EPF subfamily, EAR motifs are present in at least 21 plant gene families, most of which are transcription factor families (Kagale, Links, and Rozwadowski 2010). In a few of these gene families, there are genes with multiple EARs. However, in most families they are not as pervasive as observed in the grass EPF genes we sampled. Given the frequency of these genes in grasses it appears they too have expanded in the grass lineage following the split from the eudicots, presumably due to the WGD that occurred ~70 mya (Paterson, Bowers, and Chapman 2004) followed by additional single gene duplication events.

Experimentally, EAR motifs have been shown to function as active transcriptional repressor motifs (Ulmasov et al. 1997; Ohta et al. 2001; Hiratsu et al. 2002; Tiwari, Hagen, and Guilfoyle 2004). Many genes that contain EAR motifs are involved in plant defense and
stress response where their expression is induced by signals from the pathway that they regulate. Several different mechanisms have been proposed for their action (Kazan 2006). It has been suggested the gene products that contain EARs act as adaptor proteins forming complexes with general plant co-repressors to act in various signaling pathways. Recently, it has been shown that the EAR motif mediates the interaction in Arabidopsis between *IAA12* and *TOPLESS (TPL)*, a co-repressor, to act as an auxin-dependent transcriptional repressor (Szemenyei, Hannon, and Long 2008). Similarly *NINJA*, an AFP gene that contains an EAR motif, interacts with *TPL* co-repressors to act as a repressor in jasmonate signaling (Pauwels et al. 2010). These two examples represent the function of genes with a single EAR repression motif. An investigation of genes with multiple EAR motifs is needed to determine their role in the biological function of these genes. Such an investigation using experimental evidence may reveal new functions and repression mechanisms yet to be discovered. For instance multiple EARs could suggest a stronger repressor function or could be a means of modulating the strength of repression. In EPFs with multiple EARs, there are subtle differences in amino acid composition patterns between N- and C-terminally positioned EAR motifs, suggesting either a difference in binding partners or variation in the strength of the protein interactions with the same partner(s) (Figure 6). Regardless, multiple EARs in a gene may have evolved and been retained as a way to modify gene function through altering the strength of repression or by recruiting additional interactors. However there is no intuitive reason why this device would be so prevalent in the grasses and not the eudicots. Knowledge of the protein structure in addition to knowing their binding partners would help clarify the function of multiple EAR motifs in EPF genes.

Due to shifts in EAR number and position in EPFs through evolutionary time, there may have been several changes in the primary function of the gene subfamily. For example, in mosses a significant percentage (44%) of single-finger EPFs have no EAR motif suggesting a large proportion of the gene family may not function as transcriptional repressors (Figure 5). The functions of most EPF genes with no EARs in other species are currently unknown. *ZFP15*, a two-zinc finger rice EPF, contains no EAR repression motif and does not appear to be regulated by abiotic stresses like most other two zinc finger EPFs (Huang, Wang, and Zhang 2005). Our data suggest that the evolutionarily earliest instance of
multiple EAR motifs per EPF gene occurred with the origins of vascular plants, followed by the expansion of these genes in the grass lineage (Figure 5). Although the mechanism by which multiple EAR motifs first originated in EPF genes is uncertain, one possibility is through domain recruitment by intergenic recombination. Since the EPF gene family is quite expansive and the zinc finger DNA-binding motifs share so much sequence homology, such a scenario is intuitively possible. As is the case for EPFs with no EAR motif, the function or mechanism of action for these EPF genes with multiple EAR motifs has largely not been investigated. However, recently, it has been shown that both EAR motifs in the maize EPF ramosa1 (ra1) are necessary for physical interaction with a corepressor ramosa1 enhancer locus2 (rel2) (Gallavotti et al. 2010). RA1 and REL2 are thus components of a hypothesized transcriptional repressor complex that regulates axillary meristem fate in maize.

**Role of EPF genes in plant evolution**

The sessile nature of plants necessitates they possess an ability to respond and adapt to environmental stressors including drought, salinity, and nutrient deficiencies. Several EPFs have recently been implicated in the response to these stimuli (Sakamoto et al. 2000; Sakamoto et al. 2004; Devaiah, Nagarajan, and Raghothama 2007). It has also been suggested that genes involved in stress responses are preferentially retained following duplication events (Rizzon, Ponger, and Gaut 2006; Hanada et al. 2008). Given the stress response function of many EPFs, particularly those with multiple zinc finger DNA-binding motifs, it is intuitive that these genes following their duplication would be visible to natural selection as plants entered new environments, thus leading to their retention and subsequent diversification. As plants colonized new environments, it also became essential to evolve developmental innovations to become better adapted to these new environments. These innovations include but are not limited to vascular tissues, seeds, and flowers. Understanding how these plant innovations evolved can be achieved by identifying genetic developmental programs responsible for growth, patterning, and differentiation and determining how these programs have been modified and elaborated on to produce novel morphologies (Bowman, Floyd, and Sakakibara 2007). As these programs are elucidated, it will be interesting to see if members of the EPF subfamily of zinc finger transcription factors are central players.
CONCLUSIONS

Plants occupy numerous environments and have many morphological differences that imply diverse developmental programs, which implies an ability to adapt to different abiotic stressors. Given the known functions of EPF genes in plant development and stress response as well as the evidence of lineage-specific expansion of certain types of EPFs, we suggest these genes may underlie some of these differences in morphology and stress response. Many single-EAR-containing EPFs have been functionally characterized, but few EPF genes that lack an EAR motif or contain multiple EARs, have. Functional characterization of these genes will help elucidate the role this gene family may have played in the evolution of plant growth and developmental innovations, as well as, uncover mechanisms behind EAR function.

METHODS

Data mining

Sequences of known EPF genes in rice and Arabidopsis were acquired from previously published work (Englbrecht, Schoof, and Bohm 2004; Agarwal et al. 2007). For *Populus trichocarpa* (poplar), *Zea mays* (maize), *Sorghum bicolor* (sorghum), *Selaginella moellendorffii* (Selaginella), *Physcomitrella patens* (moss), and *Chlamydomonas reinhardtii* (*C. reinhardtii*) putative EPF genes were acquired through TBLASTN searches of public databases using the conserved zinc finger motif. For maize, searches were performed using databases at the National Center for Biotechnology Information (NCBI) and at www.maizesequence.org. For sorghum and poplar, sequences were acquired from the website: www.phytozome.net. For *Selaginella* sequences were acquired from the Joint Genome Institute (www.jgi.doe.gov). Searches for moss and *C. reinhardtii* EPFs were conducted solely using NCBI. For all BLAST searches, no restrictions were imposed on e-values in order to detect all EPF genes possible. All non-EPF genes and duplicate entries were manually deleted. Since EPF proteins generally lack introns and gene annotations were often incorrect, ORFs were found by translating nucleic acid sequences in six frames followed by searches for the highly conserved QALGGH alpha-helical motif to identify the correct frame. Where applicable, new annotations were deposited into public databases.
**Phylogenetic analysis**

Amino acid and nucleotide sequences were aligned using ClustalX 1.83 (Thompson et al. 1997) with the default parameters and alignments were further adjusted by hand. Phylogenies were reconstructed using PAUP* Version 4.0b10 (Swofford 2003) and MrBayes Version 3.1.2 (Ronquist and Huelsenbeck 2003) using a 32 amino acid or 72 nucleotide alignment of the conserved zinc finger motifs. Unrooted parsimony and neighbor-joining phylogenies were reconstructed using PAUP* and robustness was assessed by 100 replicates through the fast stepwise-addition method. Bootstrap support was assessed using a full-heuristic search with 1000 bootstrap replicates. Bayesian analyses were performed using a Markov chain Monte Carlo (MCMC) algorithm with a mixed model of evolution used to estimate the fixed rate model for the data set. In nearly all cases, the WAG model (Whelan and Goldman 2001) made the largest contribution to the results with a posterior probability greater than 95% for phylogenies reconstructed using amino acid alignments. Phylogenies reconstructed from nucleotide alignments utilized the GTR model and gamma distribution. Two independent runs of two chains each (one hot and one cold) were run for $1 \times 10^6$ generations and the first 25,000 trees were removed. Phylogenies were viewed and edited using the HyperTree Version 12.0.0 (Bingham and Sudarsanam 2000).

**Description of gene structures and conserved motifs**

Searches for the presence and number of zinc finger and EAR repression motifs were done using grep commands in the text editor, BBEdit Version 7.1.1. Other conserved motifs were found through use of multiple sequence alignments. Searches for potential regulatory element binding sites for conserved motifs were done using the PLACEv30.0 database (Higo et al. 1999). Graphical representation of the amino acid composition in known motifs was done using WebLogo version 2.8.2 (Crooks et al. 2004).

**List of abbreviations**

C2H2, Cys2/His2-type zinc finger

EAR, ERF-associated amphiphilic repression

NCBI, National Center for Biotechnology Information
WGD, whole genome duplication

Authors’ contributions
B.S. and E.V. drafted the manuscript. B.S. and R.W. carried out the data mining, gene annotations, alignments, and phylogenetic analyses. B.S. identified and characterized the distributions of conserved sequences motifs. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS
We would like to thank Elizabeth Kellogg, Clinton Whipple, Jeanne Serb, and Dennis Lavrov for invaluable discussions and suggestions.


TABLE 1
Total number of EPF genes by number of zinc finger DNA-binding motifs

<table>
<thead>
<tr>
<th>Plant Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>&gt;3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physcomitrella patens</td>
<td>18</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Selaginella moellendorffii</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>27</td>
<td>22</td>
<td>8</td>
<td>3</td>
<td>60</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>52</td>
<td>13</td>
<td>8</td>
<td>1</td>
<td>74</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>40</td>
<td>25</td>
<td>13</td>
<td>0</td>
<td>78</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>39</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>Zea mays mays</td>
<td>57</td>
<td>27</td>
<td>7</td>
<td>0</td>
<td>91</td>
</tr>
</tbody>
</table>
### TABLE 2
Tandemly duplicated EPF genes

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Distance (kb)</th>
<th>Intervening&lt;sup&gt;a&lt;/sup&gt; genes</th>
<th>Orientation&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Type&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtEPF21</td>
<td>AtEPF33</td>
<td>25</td>
<td>7</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>AtEPF23</td>
<td>AtEPF27</td>
<td>7</td>
<td>0</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>AtEPF41</td>
<td>AtEPF42</td>
<td>2</td>
<td>0</td>
<td>direct</td>
<td>2,2</td>
</tr>
<tr>
<td>AtEPF42</td>
<td>AtEPF43</td>
<td>3</td>
<td>0</td>
<td>direct</td>
<td>2,2</td>
</tr>
<tr>
<td>AtEPF52</td>
<td>AtEPF48</td>
<td>22</td>
<td>5</td>
<td>opposite</td>
<td>2,2</td>
</tr>
<tr>
<td>AtEPF66</td>
<td>AtEPF67</td>
<td>3</td>
<td>1</td>
<td>direct</td>
<td>3,3</td>
</tr>
<tr>
<td>AtEPF60</td>
<td>AtEPF55</td>
<td>1</td>
<td>0</td>
<td>direct</td>
<td>3,1</td>
</tr>
<tr>
<td>PtEPF42</td>
<td>PtEPF25</td>
<td>6</td>
<td>0</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>PtEPF5</td>
<td>PtEPF27</td>
<td>19</td>
<td>1</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>PtEPF31</td>
<td>PtEPF32</td>
<td>16</td>
<td>1</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>PtEPF58</td>
<td>PtEPF59</td>
<td>12</td>
<td>0</td>
<td>direct</td>
<td>3,3</td>
</tr>
<tr>
<td>PtEPF57</td>
<td>PtEPF73</td>
<td>12</td>
<td>0</td>
<td>direct</td>
<td>3,3</td>
</tr>
<tr>
<td>OsEPF16</td>
<td>OsEPF23</td>
<td>36</td>
<td>4</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>OsEPF24</td>
<td>OsEPF11</td>
<td>19</td>
<td>0</td>
<td>direct</td>
<td>1,1</td>
</tr>
<tr>
<td>OsEPF67</td>
<td>OsEPF49</td>
<td>35</td>
<td>5</td>
<td>opposite</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF60</td>
<td>OsEPF60</td>
<td>5</td>
<td>0</td>
<td>direct</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF51</td>
<td>OsEPF71</td>
<td>13</td>
<td>1</td>
<td>opposite</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF71</td>
<td>OsEPF54</td>
<td>3</td>
<td>0</td>
<td>direct</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF41</td>
<td>OsEPF69</td>
<td>9</td>
<td>1</td>
<td>opposite</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF72</td>
<td>OsEPF42</td>
<td>4</td>
<td>0</td>
<td>direct</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF48</td>
<td>OsEPF46</td>
<td>3</td>
<td>0</td>
<td>direct</td>
<td>2,2</td>
</tr>
<tr>
<td>OsEPF76</td>
<td>OsEPF58</td>
<td>19</td>
<td>1</td>
<td>direct</td>
<td>3,3</td>
</tr>
<tr>
<td>OsEPF75</td>
<td>OsEPF56</td>
<td>7</td>
<td>0</td>
<td>direct</td>
<td>3,3</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of nonhomologous spacer genes between tandem duplicates

<sup>b</sup> Indicates whether the duplicates are on the same strand (direct) or different strand (opposite)

<sup>c</sup> Indicates the number of zinc finger motif per EPF for each tandem duplicate
FIGURE LEGENDS

**Figure 1.** Phylogenies of EPF genes. (A) For single zinc finger EPF genes a neighbor-joining tree was reconstructed using a 72 bp nucleotide alignment of the conserved zinc finger motif of each single zinc finger EPF gene. The analysis supports the hypothesis that there are four main clades that are at least 450 million years old, which corresponds to the time since the colonization of land by the embryophytes. Clades 1-III and 1-IV have expanded more so than clades 1-I and 1-II. (B) For EPF genes with two tandemly positioned zinc finger motifs, a neighbor-joining tree was reconstructed using the 144 bp nucleotide alignment of both zinc finger motifs. The colors represent species as follows: blue = Arabidopsis, green = poplar, red = rice, and aqua = sorghum.

**Figure 2.** Differential expansion of EPF genes with two zinc fingers. (A) EPFs with two zinc fingers typically have one of three different gene structures. Some contain two tandem zinc finger motifs and an EAR and/or DLN-box repression motif (marked green elsewhere in the figure). Some have other conserved motifs including a B-box and an L-box, which are believed to be involved in nuclear localization and protein-protein interactions respectively (red). Others lack the B-box and L-box but have a CLMLL motif upstream of the first zinc finger motif (blue). (B) The prevalence of different gene structures demonstrates differential expansion among plant species as shown by the plotting the percentage of each type within each species. The Arabidopsis phylogeny (C) shows expansion of EPF genes containing the CLMLL motif whereas the other eudicot, poplar, shows expansion of B-box and L-box type EPF genes (D). Finally, the rice phylogeny shows expansion of EPF genes lacking all the conserved upstream sequence motifs (E), which is also the case in sorghum and maize.

**Figure 3.** Conserved sequence motifs in the GIS clade. By aligning of all the GIS-like genes in Arabidopsis, poplar, rice, sorghum, maize, and *P. taeda* we identified three novel sequence motifs in addition to the DNA-binding zinc finger motif and a C-terminal EAR repression motif. The WebLogos show the amino acid composition of these motifs. Subfunctionalization and/or neofunctionalization may have occurred following duplication as evidenced by the pattern of presence and absence of these conserved motifs.
Figure 4. Phylogeny of GIS-like genes. Both posterior probability support values from Bayesian analysis (first value) and bootstrap support (second value) are shown for each node. The phylogeny shows the main mode of expansion in grasses is due to the WGD before the divergence of all grasses with further expansion in maize due to an additional WGD. The phylogeny comprises all GIS-like genes in Arabidopsis (At), poplar (Pt), rice (Os), sorghum (Sb), and maize (Zm) and uses the Pinus taeda homologs as outgroups.

Figure 5. Proportion and phylograms of single zinc finger EPFs genes based on EAR number. Single zinc finger EPF genes have four main gene structures (A) which are represented as follows: blue, no EARs; red, one C-terminally positioned EAR; green, one N- and C-terminally positioned EAR; aqua, two C-terminally positioned EARs; and black, unknown number of EAR motifs. The proportion of EPFs of each gene structure type (B) shows an increase in the proportion of EPF genes with multiple EARs in grasses compared to the other plant species sampled. Numbers within bars indicate the total count of each respective gene structure type. (C) and (D) Neighbor-joining phylograms were reconstructed using an alignment of the conserved 72 bp nucleic acid sequence of the zinc finger DNA-binding motif. The same colors are used to indicate EAR number and position for each gene. The grass phylogram (C) shows certain clades of multi-EAR-containing EPFs have significantly expanded in grasses but this is not seen in the eudicots (D). The species are represented as follows: P. patens; Pp, Selaginella; Sm, Arabidopsis; At, rice; Os, sorghum; Sb, and maize; Zm.

Figure 6: The amino acid composition of EAR repression motifs in single zinc finger EPF genes. WebLogos show similar amino acid composition for C-terminally positioned EAR motifs across species. C-terminal EARs in Arabidopsis (A) and rice (B) are used as examples. Most EAR motifs consist of four leucines and intervening amino acids to be acidic at X₁, polar at X₂, and basic at X₃. (C) N-terminally positioned EARs differ somewhat in amino acid composition. Positions X₁ and X₃ are polar, but not necessarily acidic and basic as seen in C-terminal EAR repression motifs.
Figure 1 Sigmon, Weeks, & Vollbrecht
Figure 2 Sigmon, Weeks, & Vollbrecht

A

- C2H2 C2H2 E
- B L C2H2 C2H2 E
- CLMLL C2H2 C2H2 E

B

green
red
blue

C

D

E

Arabidopsis
Poplar
Rice
Figure 3 Sigmon, Weeks, & Vollbrecht
**Figure 4** Sigmon, Weeks, & Vollbrecht

![Tree Diagram]

- **Pinus taeda EPF1**
- **Pinus taeda EPF2**
- **AtEPF8**
- **AtEPF9**
- **AtEPF10**
- **PtEPF28**
- **PtEPF51**
- **OsEPF22**
- **SbEPF28**
- **ZmEPF53**
- **OsEPF33**
- **SbEPF27**
- **ZmEPF55**
- **ZmEPF81**
- **OsEPF30**
- **SbEPF25**
- **ZmEPF54**
- **ZmEPF56**
- **OsEPF35**
- **SbEPF26**
- **ZmEPF58**
- **OsEPF36**
- **SbEPF23**
- **ZmEPF57**

- 0.1
Figure 5 Sigmon, Weeks, & Vollbrecht
Figure 6 Sigmon, Weeks, & Vollbrecht
CHAPTER 5. GENERAL CONCLUSIONS

SUMMARY AND DISCUSSION

To expand our knowledge of the role of \textit{ral} in the derivation of modern maize, we used developmental analyses, population genetics, and phylogenetics to test the hypothesis that the \textit{ral} locus was a target of artificial selection during domestication. We found that the developmental basis for the crooked rows in weak \textit{ral} mutant ears was the production of extra spikelets from increased indeterminacy of the spikelet pair meristems (SPMs). This effect on row ordering may decrease yield, which suggests a phenotypic rationale behind testing \textit{ral} for selection during the domestication of maize. Additionally, previous results presented evidence that artificial selection had occurred at some point in the history of maize, but did not pinpoint when the selection occurred (Vollbrecht et al. 2005). The developmental data combined with previous evidence of selection justified further inquiry into the role of \textit{ral} in the history in the development of modern maize. Population genetics and phylogenetic analyses showed nucleotide diversity was significantly reduced in maize landraces when compared to teosintes and that one haplotype predominates in the landrace population. Also, most tests for selection strongly suggested positive selection had occurred in the landraces, but not in the wild progenitor teosinte. Combined, these data suggest the \textit{ral} locus was indeed a target of artificial selection during the domestication process and therefore important in the evolution of architectural changes in the maize ear.

Given the function of \textit{ral} in the development of maize inflorescence architecture and its importance in evolution of the maize ear, we next asked if changes in amino acid sequence and/or expression patterns of \textit{ral} could explain some of the great phenotypic diversity of inflorescence architecture in other grasses. Using comparative genetics, we found \textit{ral} is likely absent in rice but present in sorghum, supporting previous results (Vollbrecht et al. 2005). Since maize and sorghum are both members of the grass tribe Andropogoneae, we focused on this group of grasses for subsequent analyses. Through sequence analysis we identified conserved motifs, which may be important for proper gene function. We also found molecular changes including duplications, frameshifts, and missense mutations that correlate with observed inflorescence architecture phenotypes. Using tests for selection we
found some conserved motifs are under tight functional constraint and that the gene, as a whole, is under varying levels of purifying selection. By combining these data with phylogenetic analysis, we found evidence that several nonsynonymous changes in \textit{ra1} sequence occurred early in the evolution of the Andropogoneae and are preserved in their respective lineages. This result suggests changes in \textit{ra1} function may have been important early in the evolution of the Andropogoneae. Finally, by examining \textit{ra1} in a diverse panel of sorghum varieties we found unusually low diversity compared to wild relatives, which suggests \textit{ra1} may have been selected on in the context of increased inflorescence branch number in sorghum.

The discovery of a second conserved EAR motif in \textit{ra1} spurred further investigation into the evolution of the EPF subfamily of C2H2 transcription factors. Overall, we found this subfamily has greatly expanded in higher plants due to both large and small-scale duplications with preferential gene retention and that lineage specific differences exist, especially between eudicots and monocots. More specifically, we found the subfamily likely originated with land plants and that EAR repression motifs were present at this time-point during the evolution of plants. However, EPFs with multiple EAR motifs were first found in vascular plants and remain few in number in eudicots as well. Expansion of EPFs, especially those with multiple EAR motifs, has occurred along the grass lineage, suggesting duplication and preferential retention of these genes. The totality of these data suggest that key events during the evolution of the EPF subfamily, some of whose members function in plant development and abiotic stress response, correlate with developmental innovations during the life history of plants and therefore the gene family has likely been important in the acquisition of these innovations.

The summation of this work provides a study of the role of a gene in the development and evolution of a key trait in grasses using a variety of tools including comparative genomics, population genetics, phylogenetics, molecular evolution, and developmental analyses. Comprehensive research of this kind has implications for multiple disciplines. First, compared to animals relatively few studies have been done that focus on the evolution of plant development. This point is especially true for grasses. These comprehensive studies are particularly valuable for grass inflorescence traits due to their importance as crops. The study
of domestication loci provides a unique opportunity to identify genes that may have been important during the evolution and development of inflorescence traits (Doust 2007). However in maize, currently only *teosinte glume architecture1* (*tga1*) and *ra1* show evidence of artificial selection for modifications to ear structure during the domestication of maize (Wang et al. 2005; Sigmon and Vollbrecht 2010). Since the maize ear is central to grain yield, as worldwide demand continues to increase, it is particularly important to understand the genetic basis underlying its development and evolution. We also found evidence that *ra1* has been important during the evolution of other grasses, particularly sorghum. In sorghum, *ra1* showed evidence of artificial selection, presumably for an increase in branching. Therefore, *ra1* may have been selected on in both crops to increase grain yield, but through different means. By analyzing the molecular evolution of *ra1* in the Andropogoneae, both changes in expression pattern due to gene regulation and molecular changes that may alter protein structure and function have been found that correlate with inflorescence morphology (Vollbrecht et al. 2005). Therefore for *ra1*, both of these mechanisms have been important during the evolution of inflorescence architecture in grasses, suggesting there are multiple ways this highly diverse trait can be modified through time.
LITERATURE CITED


APPENDIX A. FASCIATED PHENOTYPES IN THE MAIZE LANDRACE VERACRUZ 85

Brandi Sigmon and Erik Vollbrecht

ABSTRACT

The maize landrace Veracruz 85 displayed a variety of defects in both vegetative and inflorescence development including increased organ production and fasciation. Phenotypic characterization and developmental analysis of these plants suggest the most likely explanation for the observed phenotypes is an enlargement of all meristem types. Initial genetic mapping results suggest at least one candidate locus for the effects in bin 5.05 on chromosome five.

INTRODUCTION

In plants, meristems are responsible for organogenesis and for maintaining a pool of undifferentiated cells, which allows the meristem to maintain itself. Types of meristems differ in their level of determinacy and identity, both of which correlate with position. Indeterminate meristems have potential to generate organs indefinitely whereas determinate meristems will only produce a limited number of structures before they terminate. The shoot apical meristem (SAM) gives rise to the vegetative shoot and is indeterminate in many species. Axillary meristems (AM) located in the axils of leaves produce branches when indeterminate and, for example, flowers when determinate. Several types of AMs are present in the maize inflorescence and they differ in their determinacy, which is the developmental basis for the architecture of inflorescences (Vollbrecht and Schmidt 2009). In inflorescence development the apical inflorescence meristem (IM) is indeterminate and it produces axillary branch meristems (BM). These meristems are also indeterminate as they go on to produce the long branches in the tassel. After the BMs have been initiated, the IM switches to producing spikelet pair meristems (SPM), which are determinate as they produce only two spikelet meristems (SM). The SMs are also determinate as they produce two determinate floral meristems (FM), which go on to produce a fixed number of floral organs.
Analysis of mutants has identified at least two classes of genes that control meristem function. In the first class, loss of function mutants have defects in meristem formation and/or maintenance. *WUSCHEL (WUS)* mutants, which fail to maintain the SAM (Laux et al. 1996; Mayer et al. 1998; Kieffer et al. 2006), are well studied in Arabidopsis. In maize, the *SHOOTMERISTEMLESS (STM)* homolog *knotted1 (kn1)* is expressed in the SAM and is required for meristem shoot meristem establishment and maintenance (Kerstetter and Hake 1997; Kerstetter et al. 1997; Vollbrecht, Reiser, and Hake 2000). A second class of meristem mutant causes meristem enlargement, which results in overproliferative meristems. In maize this class of mutants includes several fasciation mutants, such as *fasciated ear (fae1 and fae2), fascicled1 (fas1), and thick tassel dwarf (td1)* (Hake and Veit 1988; Orr, Hass, and Sundberg 1997; Taguchi-Shiobara et al. 2001; Bommert et al. 2005). In general, these mutants produce supernumerary organs and in extreme cases, fasciated inflorescences due to enlarged IMs (Orr, Hass, and Sundberg 1997; Taguchi-Shiobara et al. 2001; Bommert et al. 2005). *td1* and *fea2* are the putative orthologs of *CLAVATA1 (CLV1)* and *CLAVATA2 (CLV2)* in Arabidopsis (Clark, Running, and Meyerowitz 1993; Clark, Williams, and Meyerowitz 1997; Kayes and Clark 1998; Jeong, Trotochaud, and Clark 1999), respectively, suggesting the *CLV* signaling pathway may be conserved in grasses (Taguchi-Shiobara et al. 2001; Bommert et al. 2005). However differences in expression exist, such as the absence of expression in the SAM, which suggests the pathway may be more complicated in maize.

In the maize landrace Veracruz 85 we found fasciation of both inflorescence and vegetative structures, which appears to be due to the enlargement of multiple meristem types, including the vegetative SAM. These developmental abnormalities could be due to defects in *CLV* signaling. Regardless, determining the genetic basis of the widespread fasciation in this accession could provide further insight into the regulation of meristem size in maize.

**MATERIALS AND METHODS**

**Plant Growth and Phenotype Characterization**

Seed for the maize landrace Veracruz 85 (Ames 19759) was obtained from the USDA, ARS, National Genetic Resources Program, Germplasm Resources Information Network (http://www.ars-grin.gov/cgi-bin/npgs/). Both BC1 and F2 populations were
produced in B73 and Mo17 inbred backgrounds. Analysis of phenotypes was carried out for all populations on plants grown in Ames, IA during the summer of 2010. Plants were scored for several vegetative and inflorescence phenotypes. In genetic mapping populations, The B73 and Mo17 parents were considered normal. Individual plants called as normal were defined as plants with no discernable phenotypic abnormalities or those with only a very weak phenotype in either vegetative or inflorescence structures, but not both. Individuals with strong phenotypes were defined as plants with severe abnormalities in both vegetative and inflorescence structures. Intermediate individuals were classified as having weak phenotypes.

**Scanning Electron Micrographs (SEMs)**

Immature ears and tassels (0.2 to 1.5 cm) were collected from plants grown during the summer 2010 in Ames, IA. Samples were fixed with 2% paraformaldehyde and 3% glutaraldehyde in 0.1M cacodylate buffer at 4C for 24 hours. Samples were rinsed in deionized water and post-fixed in 2% aqueous osmium tetroxide followed by dehydration in a graded ethanol series up to 100% ultra-pure ethanol and dried using a Denton DCP-2 critical point dryer (Denton Vacuum, LLC, Moorestown, NJ). When dried, the samples were placed onto adhesive coated aluminum stubs, sputter coated (Denton Desk II sputter coater, Denton Vacuum, LLC, Moorestown, NJ) with palladium/gold alloy, and imaged using a JEOL 5800LV SEM (Japan Electron Optics Laboratory, Peabody, MA) at 10kV with a SIS ADDA II for digital image capture (Olympus Soft Imaging Systems, ResAlta, Golden, CO).

**Bulk Segregant Analysis (BSA)**

Pooled genomic DNA for strong mutant and normal plants was prepared for all populations and submitted to the Genomic Technologies Facility at Iowa State University for BSA. The number of individuals in each pooled sampled varied depending on the population as follows: V85>B73 BC1, 14 normal vs. 15 mutant; V85>Mo17 BC1, 36 normal vs. 26 mutant; V85>B73 F2, 8 normal vs. 22 mutant; and V85>Mo17 F2, 10 normal vs. 12 mutant (Table 1).
RESULTS

Phenotypes Include Vegetative and Floral Developmental Defects

Extra organs and fasciation were found in both inflorescence and vegetative structures in all populations. In tassels, the most obvious phenotype was clusters of spikelets, which all appear to be sessile, up to eight in number separated by elongated internodes (Fig. 1C and E). Also the spikelets often contained extra florets, up to six in total number, and some of these florets produced extra floral organs (Fig. 1F, G and H). However, each floret had one lemma and one palea. Supernumerary glumes were also observed in some spikelets (Fig. 1H). More rarely the main rachis or the long primary branches were bifurcated and/or fasciated (Fig. 1B).

A range of phenotypes was also present in the ears. The most common phenotype observed in all populations was row disorder, usually due to the presence of extra or sometimes absence of spikelets (Fig. 2B and F). The ears of some individuals were bifurcated and the more severely affected individuals were fasciated at the apex (Fig. 2C and D). In the V85>Mo17 BC1 population, many individuals had extra glume-like appendages extending from the glume-proper (Fig. 2E). Other observed aberrations include extra silks and sex determination phenotypes in the F2 populations (data not shown).

Vegetative development was also affected in all populations. Commonly observed differences in plant architecture include tillering (Fig. 3A), production of multiple ears at the nodes on the shank of the primary ear (Fig. 3C), and one individual had two tassels (Fig. 3B). Also, extra leaves were observed, especially husk leaves (Fig. 3D, F, and G). The leaves themselves often appeared to be broadened and comprised of multiple, fused leaf-like structures (Fig. 3G). More rarely, some leaves had two mid veins (Fig. 3E). Finally, twin seedlings were sometimes produced with the same axis of symmetry (Fig. 3H and I).

The degree or severity of the developmental abnormalities differed between populations and organ types. In general, populations with a B73 inbred background had more severe developmental aberrations than those with a Mo17 inbred background. This result may be explained by the fact that B73 has larger meristems than Mo17, and is therefore a more permissive background for phenotypes related to meristem size (Vollbrecht, Reiser, and Hake 2000). Also the F2 populations had more severe phenotypes than the BC1 populations.
Among the inflorescences, the tassel was more severely affected than the ear. There were also positional differences. For instance, on the primary tassel branches the tips were more severely affected than the base.

**Developmental Basis of Some Observed Phenotypes**

Most of the observed phenotypes in both inflorescence and vegetative development were due to supernumerary organs, which may be caused by enlarged meristems. In inflorescence development, many types of meristems are produced. Enlargement of different types of meristems could have different phenotypic consequences. For instance, enlarged IMs may result in extra rows of spikelets in addition to bifurcation and/or fasciation of the inflorescences. The effect on SPMs would be production of supernumerary SMs, which would result in increased spikelet production. Enlarged SMs would result in extra glumes and florets. Similarly, FMs would produce supernumerary floral organs. We observed all of these phenotypes in both tassels and ears, though more severely in the tassels. These results are in contrast with other fasciated mutants in maize, which typically impact ear IMs but other meristem types are only mildly affected (Taguchi-Shiobara et al. 2001; Bommert et al. 2005).

In addition to inflorescences, we observed defects in vegetative development. Vegetative development depends on the vegetative SAM, which produces lateral organs. Therefore enlargement of the SAM may change plant architecture due to increased production of axillary structures such as ears and tillers. An enlarged SAM could also bifurcate or fasciate in severely affected individuals. These phenotypes were present in all populations, though more so in the B73 inbred background. We also observed putative SAM bifurcations early in development, which resulted in twin plants with the same axis of symmetry. The range of observed vegetative developmental defects suggest that changes may start early in ontogeny and continue throughout the lifecycle of the plant. These finding are also in contrast to other fasciated maize mutants, which typically have no or only mild changes in vegetative development (Taguchi-Shiobara et al. 2001; Bommert et al. 2005).

To determine the developmental basis of some of the changes in inflorescence morphology, immature ears and tassels were visualized through their development using scanning electron microscopy (SEM). Compared to normal siblings, affected inflorescences
generally had meristems of various sizes (Fig. 4A and B), though enlargement was most commonly observed. Spikelet density also appeared to be increased compared to normal individuals. Later in development, the long primary branches of tassels showed the production of supernumerary spikelets and glumes (Fig. 4C-F). The summation of these data suggests altered regulation of meristem size, which results in the production of supernumerary organs.

**BSA Reveals a Candidate Region in the Maize Genome**

Pooled gDNA from developmentally normal and abnormal individuals was collected and submitted for BSA mapping for all six populations. BSA is a quantitative Sequenom-based SNP assay that can be used to genetically map mutants (Liu et al. 2009). Using this method on our four populations, only the V85>B73 F2 population produced significant results, for two regions located on chromosome five. For one region, the shape of the curve appears bimodal, which suggests two linked loci may contribute to the observed phenotypes (Fig. 5). The other region is located at the beginning of the chromosome and the shape of the curve suggests one locus may be located there. Collaborators have tentatively mapped a fasciated mutant to the same location as the bimodal curve on chromosome five (Dave Jackson, personal communication). No previously described mutant involved in meristem regulation maps to this region, but we found this chromosomal segment to be homeologous to a segment containing *fasciated ear2* (*fea2*) on chromosome four. Using the collinearity between these segments and the orthologous regions in sorghum and rice, no obvious homeolog of *fea2* was found on maize chromosome five. However given the mapping results, the coincidence with a previously mapped a fasciated mutant, in conjunction with the synteny with the chromosomal segment containing *fea2*, gene(s) that contribute to these fasciated phenotypes in Veracruz 85 are likely located on chromosome five.

**CONCLUSIONS**

The fasciated phenotypes found in Veracruz 85 share some similarities with other faciated mutants, most notably *fea2* and *tdl*. *fea2* causes overproliferation of the ear IM and has only a modest effect on SMs and FMs (Taguchi-Shiobara et al. 2001). Also, tassels are
only mildly affected and no discernable abnormalities were found during vegetative
development. *td1* mutants, like *fea2* most strongly affect the ear, but also have some mild
defects in tassel and vegetative development (Bommert et al. 2005). This pattern is somewhat
opposite to the observed phenotypes in Veracruz 85 since tassels are more affected than ears
and abnormalities are present throughout vegetative development. Also it appears that all
meristem types are affected in Veracruz 85, which is unlike fasciated mutants in maize as
they generally impact primarily a subset of meristem types. The *CLAVATA (CLV)* genes in
Arabidopsis are expressed in all meristem types (Clark, Williams, and Meyerowitz 1997),
unlike their putative orthologs in maize, *td1* and *fea2* (Taguchi-Shiobara et al. 2001;
Bommert et al. 2005), which presents a paradox concerning the role of *CLV* genes in maize
development. Since meristem enlargement is hypothetically due to a deficiency in *CLV*-type
signaling, finding the genetic basis responsible for developmental abnormalities in Veracruz
85 may provide further insight into the role of *CLV* genes in regulation of meristem size in
maize.

ACKNOWLEDGEMENTS
We would like to thank Tracey Pepper’s assistance in SEM imaging and Rebecca Weeks for
field assistance and processing the BSA data.
LITERATURE CITED


TABLE 1
Frequency and degree of abnormal developmental phenotypes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Normal</th>
<th>Weak</th>
<th>Strong</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>V85 &gt; B73 BC1</td>
<td>14</td>
<td>21</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>V85 &gt; Mo17 BC1</td>
<td>36</td>
<td>45</td>
<td>26</td>
<td>107</td>
</tr>
<tr>
<td>V85 &gt; B73 F2</td>
<td>8</td>
<td>30</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>V85 &gt; Mo17 F2</td>
<td>10</td>
<td>50</td>
<td>12</td>
<td>72</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

**Figure 1.** Tassel phenotypes in Veracruz 85. (A) A normal maize tassel (V85>B73 F2). (B) A tassel with bifurcation and fasciation of the central rachis (arrowed) (V85>B73 F2). (C) A tassel showing highly clustered spikelets (V85>B73 F2). (D) A normal primary branch (V85>B73 F2). (E) A primary branch with multiple clusters of spikelets separated by elongated internodes (V85>B73 F2). (F) A normal spikelet with two florets, each with three stamens (V85>Mo17 F2). (G) A spikelet with six florets with a total of 20 stamens (V85>Mo17 F2). (H) A spikelet with 11 stamens and three glumes (arrowed) (V85>Mo17 F2).

**Figure 2.** Ear phenotypes in Veracruz 85. (A) A normal maize ear (V85>B73 BC1). (B) An ear showing row disorder (V85>B73 F2). (C) An ear with a single bifurcation at the base of the inflorescence (V85>B73 BC1). (D) An ear with fasciation at the apex (V85>B73 BC1). (E) An ear with excess growth of the glume tissue (arrowed) (V85>Mo17 BC1). (F) An ear with abnormalities in kernel size and arrangement (arrowed) (V85>B73 F2).

**Figure 3.** Observed changes in vegetative development. (A) A developing plant with tillers (V85>Mo17 F2). (B) A plant with a single bifurcation at a node (arrowed) resulting in the formation of two tassels (V85>B73 BC1). (C) A plant with several husk leaves removed to show an elongated shank and supernumerary ears (V85>B73 F2). (D) A plant with an overproliferation of husk leaves (V85>B73 F2). (E) A leaf blade with two mid veins (V85>B73 BC1). (F) A plant with two extra sheath-like leaves at the base of the tassel (arrowed) (V85>Mo17 F2). (G) A fasciated husk leaf (V85>B73 F2). (H) A seedling with an early bifurcation of the SAM (arrowed) (V85>B73 BC2). (I) Twin seedlings have the same axis of symmetry (V85>B73 BC1).

**Figure 4.** SEMs of immature tassels and ears. (A) Part of an ear row with normal pairing of spikelets. (B) Part of an ear row with unequal spikelet sizes. (C and D) Supernumerary spikelets and glumes are produced on the primary tassel branches. (E and F) Inset of portions
of the primary tassel branches shows detail of supernumery spikelets and glumes, some are arrowed. Scale bar = 250 uM.

**Figure 5.** Candidate regions on chromosome 5. Transformed values for markers were averaged over 10 cM bins and plotted. For unlinked markers, values trend towards zero whereas for linked markers values increase as they approach a peak, which therefore marks the position of a candidate locus. The graph shows two peaks at ~80 cM and 110 cM, suggesting two linked loci may be responsible for the observed fasciated phenotypes. The arrow points to a previous hit for a fasciated mutant, which maps to the second peak. A third peak is located at ~10cM, which may also be responsible for the observed phenotypes.
Figure 1 Sigmon & Vollbrecht
Figure 2 Sigmon & Vollbrecht
Figure 3 Sigmon & Vollbrecht
Figure 4 Sigmon and Vollbrecht
Figure 5 Sigmon and Vollbrecht
APPENDIX B. SUPPLEMENTARY DATA

TABLE 1
Complete list of *Sorghum* accessions

<table>
<thead>
<tr>
<th>ID</th>
<th>Species</th>
<th>Race</th>
<th>Type</th>
<th>Accession</th>
<th>Origin</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>SbC1</td>
<td><em>S. bicolor</em> ssp.</td>
<td>caud./kafir</td>
<td>inbred</td>
<td>B.Tx623</td>
<td>-</td>
<td>PB</td>
</tr>
<tr>
<td>SbC2</td>
<td><em>S. bicolor</em> ssp.</td>
<td>guinea</td>
<td>inbred</td>
<td>IS3620C</td>
<td>-</td>
<td>PB</td>
</tr>
<tr>
<td>SbC3</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>NSL 56003</td>
<td>Kenya</td>
<td>PB</td>
</tr>
<tr>
<td>SbC9</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>NSL 92371</td>
<td>Swaziland</td>
<td>PB</td>
</tr>
<tr>
<td>SbC17</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>PI 585454</td>
<td>Ghana</td>
<td>PB</td>
</tr>
<tr>
<td>SbC33</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>PI 329353</td>
<td>Ethiopia</td>
<td>BS</td>
</tr>
<tr>
<td>SbC34</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>PI 520795</td>
<td>Kenya</td>
<td>BS</td>
</tr>
<tr>
<td>SbC35</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>PI 536007</td>
<td>Cameroon</td>
<td>BS</td>
</tr>
<tr>
<td>SbC36</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>PI 537086</td>
<td>Niger</td>
<td>BS</td>
</tr>
<tr>
<td>SbC38</td>
<td><em>S. bicolor</em> ssp.</td>
<td>bicolor</td>
<td>landrace</td>
<td>PI 591036</td>
<td>Nigeria</td>
<td>BS</td>
</tr>
<tr>
<td>SbC10</td>
<td><em>S. bicolor</em> ssp.</td>
<td>caudatum</td>
<td>landrace</td>
<td>PI 221607</td>
<td>Ghana</td>
<td>PB</td>
</tr>
<tr>
<td>SbC16</td>
<td><em>S. bicolor</em> ssp.</td>
<td>caudatum</td>
<td>landrace</td>
<td>PI 152702</td>
<td>Sudan</td>
<td>PB</td>
</tr>
<tr>
<td>SbC20</td>
<td><em>S. bicolor</em> ssp.</td>
<td>caudatum</td>
<td>landrace</td>
<td>NSL 87666</td>
<td>C.A.R.</td>
<td>PB</td>
</tr>
<tr>
<td>SbC4</td>
<td><em>S. bicolor</em> ssp.</td>
<td>dura</td>
<td>landrace</td>
<td>NSL 77217</td>
<td>Chad</td>
<td>PB</td>
</tr>
<tr>
<td>SbC8</td>
<td><em>S. bicolor</em> ssp.</td>
<td>dura</td>
<td>landrace</td>
<td>NSL 55243</td>
<td>Algeria</td>
<td>PB</td>
</tr>
<tr>
<td>SbC21</td>
<td><em>S. bicolor</em> ssp.</td>
<td>dura</td>
<td>landrace</td>
<td>NSL 87902</td>
<td>Cameroon</td>
<td>PB</td>
</tr>
<tr>
<td>SbC7</td>
<td><em>S. bicolor</em> ssp.</td>
<td>guinea</td>
<td>landrace</td>
<td>NSL 50875</td>
<td>Chad</td>
<td>PB</td>
</tr>
<tr>
<td>SbC15</td>
<td><em>S. bicolor</em> ssp.</td>
<td>guinea</td>
<td>landrace</td>
<td>PI 267408</td>
<td>Uganda</td>
<td>PB</td>
</tr>
<tr>
<td>SbC18</td>
<td><em>S. bicolor</em> ssp.</td>
<td>guinea</td>
<td>landrace</td>
<td>NSL 51030</td>
<td>Mali</td>
<td>PB</td>
</tr>
<tr>
<td>SbC19</td>
<td><em>S. bicolor</em> ssp.</td>
<td>guinea</td>
<td>landrace</td>
<td>NSL 51365</td>
<td>India</td>
<td>PB</td>
</tr>
<tr>
<td>SbC13</td>
<td><em>S. bicolor</em> ssp.</td>
<td>kafir</td>
<td>landrace</td>
<td>NSL 56174</td>
<td>Ethiopia</td>
<td>PB</td>
</tr>
<tr>
<td>SbC14</td>
<td><em>S. bicolor</em> ssp.</td>
<td>kafir</td>
<td>landrace</td>
<td>PI 267539</td>
<td>India</td>
<td>PB</td>
</tr>
<tr>
<td>SbW24</td>
<td><em>S. bicolor</em> ssp.</td>
<td>aethiopicum</td>
<td>wild</td>
<td>L-WA67</td>
<td>Sudan</td>
<td>PB</td>
</tr>
<tr>
<td>SbW5</td>
<td><em>S. bicolor</em> ssp.</td>
<td>arundinaceum</td>
<td>wild</td>
<td>L-WA55</td>
<td>Benin</td>
<td>PB</td>
</tr>
<tr>
<td>SbW26</td>
<td><em>S. bicolor</em> ssp.</td>
<td>arundinaceum</td>
<td>wild</td>
<td>L-WA59</td>
<td>South Africa</td>
<td>PB</td>
</tr>
<tr>
<td>SbW11</td>
<td><em>S. bicolor</em> ssp.</td>
<td>verticilliflorum</td>
<td>wild</td>
<td>L-WA28</td>
<td>Angola</td>
<td>PB</td>
</tr>
<tr>
<td>SbW12</td>
<td><em>S. bicolor</em> ssp.</td>
<td>verticilliflorum</td>
<td>wild</td>
<td>L-WA42</td>
<td>South Africa</td>
<td>PB</td>
</tr>
<tr>
<td>SbW22</td>
<td><em>S. bicolor</em> ssp.</td>
<td>verticilliflorum</td>
<td>wild</td>
<td>L-WA15</td>
<td>Sudan</td>
<td>PB</td>
</tr>
<tr>
<td>SbW23</td>
<td><em>S. bicolor</em> ssp.</td>
<td>verticilliflorum</td>
<td>wild</td>
<td>L-WA23</td>
<td>Angola</td>
<td>PB</td>
</tr>
<tr>
<td>SbW25</td>
<td><em>S. bicolor</em> ssp.</td>
<td>verticilliflorum</td>
<td>wild</td>
<td>L-WA13</td>
<td>Sudan</td>
<td>PB</td>
</tr>
<tr>
<td>SbW27</td>
<td><em>S. bicolor</em> ssp.</td>
<td>verticilliflorum</td>
<td>wild</td>
<td>L-WA38</td>
<td>South Africa</td>
<td>PB</td>
</tr>
</tbody>
</table>

Source of sorghum sequences: Patrick Brown (PB) and Brandi Sigmon (BS)
TABLE 1. (continued)

<table>
<thead>
<tr>
<th>ID</th>
<th>Species</th>
<th>Race</th>
<th>Type</th>
<th>Accession</th>
<th>Origin</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>SbW40</td>
<td><em>S. bicolor</em> ssp. vermicilliflorum</td>
<td>wild</td>
<td>PI 153867</td>
<td>Kenya</td>
<td>BS</td>
<td></td>
</tr>
<tr>
<td>SbW6</td>
<td><em>S. bicolor</em> ssp. virgatum</td>
<td>wild</td>
<td>L-WA88</td>
<td>Egypt</td>
<td>PB</td>
<td></td>
</tr>
<tr>
<td>SbW28</td>
<td><em>S. bicolor</em> ssp. virgatum</td>
<td>wild</td>
<td>L-WA89</td>
<td>Egypt</td>
<td>PB</td>
<td></td>
</tr>
<tr>
<td>SbW39</td>
<td><em>S. bicolor</em> ssp. drummondii</td>
<td>wild</td>
<td>PI 208578</td>
<td>India</td>
<td>BS</td>
<td></td>
</tr>
<tr>
<td>SbH29</td>
<td><em>S. bicolor</em> ssp. drummondii</td>
<td>wild/cult hybrid L-WA72</td>
<td>Sudan</td>
<td>PB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SbH30</td>
<td><em>S. bicolor</em> ssp. drummondii</td>
<td>wild/cult hybrid L-WA81</td>
<td>Sudan</td>
<td>PB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SxA45</td>
<td><em>Sorghum x alnum</em></td>
<td>-</td>
<td>PI 190579</td>
<td>Argentina</td>
<td>BS</td>
<td></td>
</tr>
<tr>
<td>S.hal41</td>
<td><em>S. halepense</em></td>
<td>-</td>
<td>PI 209217</td>
<td>South Africa</td>
<td>BS</td>
<td></td>
</tr>
<tr>
<td>S.pro31</td>
<td><em>S. propinquum</em></td>
<td>-</td>
<td>UGA propinquum</td>
<td>-</td>
<td>PB</td>
<td></td>
</tr>
<tr>
<td>S.pro32</td>
<td><em>S. propinquum</em></td>
<td>-</td>
<td>TAMU propinquum</td>
<td>-</td>
<td>PB</td>
<td></td>
</tr>
<tr>
<td>S.lax42</td>
<td><em>S. laxiflorum</em></td>
<td>-</td>
<td>PI 562254</td>
<td>Sudan</td>
<td>BS</td>
<td></td>
</tr>
<tr>
<td>S.pur44</td>
<td><em>S. purpureoserciceum</em></td>
<td>-</td>
<td>PI 536008</td>
<td>Cameroon</td>
<td>BS</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

**Supplementary Figure 1.** *ral* phylogeny reconstructed using only coding sequence. The Bayesian phylogeny includes all gene products including alleles and putative gene duplicates and was reconstructed using the HKY+gamma model of nucleotide substitution and included empirically estimated base frequencies and transition/transversion ratios. The first 75,000 trees were removed from the analysis.

**Supplementary Figure 2.** *ral* sorghum phylogeny. The Bayesian phylogeny includes all *ral* gene products from all accessions sampled (Supplementary Table 1). In the taxon names, “D” designates the downstream gene copy, “U” the upstream copy, and “M” the recombinant copies. Also the lowercase letters a, b, and c designate multiple alleles cloned within an accession.
Supplementary Figure 1 Sigmon et al.
Supplementary Figure 2 Sigmon et al.
ACKNOWLEDGEMENTS

I would like to thank everyone who has supported me throughout my graduate studies including my mentors, collaborators, family, friends, and lab mates. Especially, I would like to express my appreciation to my Ph.D advisor, Dr. Erik Vollbrecht, for his mentorship in my professional development and for the freedom to explore my own ideas and projects. Thanks also to the rest of my committee members Drs. Philip Becraft, Volker Brendel, Lynn Clark, and Jeanne Serb for their invaluable advice and guidance. I also would like to express my gratitude to my undergraduate advisors Drs. Joanne Croom, William Hutt, and Frank Quick for sparking my interest in science through their unbridled enthusiasm. To my lab mates both past and present, thanks for your friendship, support, and advice throughout these six years. Thanks especially to Becky for not only being a great lab mate, but also a wonderful friend and collaborator. I am also grateful to my best friend Kelley for keeping me laughing and for being more like a sister than a friend. Finally, I would also like to express my gratitude to my family. To my parents Donna and Alan Sigmon, thanks for your unconditional love and for instilling a strong work ethic in me. No doubt, I wouldn’t be here without you! Thanks also to my brother Keith for being my friend and partner in crime. Most of all, I am grateful to my husband Benny for his unwavering love, encouragement, and support. Thank you for being so patient!