# Interactions among gene regulation and expression, sequence deletion, and purifying selection following whole genome duplications in flowering plants

By

James Carey Schnable

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Plant Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Freeling, Chair

Professor Sarah Hake

Professor Mike Levine

Fall 2012

Interactions among gene regulation and expression, sequence deletion, and purifying selection following whole genome duplications in flowering plants.



by

James Carey Schnable

Abstract

Interactions among gene regulation and expression, sequence deletion, and purifying selection following whole genome duplications in flowering plants.

by

James Carey Schnable

Doctor of Philosophy in Plant Biology

University of California, Berkeley

Professor Michael Freeling, Chair

Polyploidy, or whole genome duplication, is rampant among both extant and ancient flowering plant species. Whole genome duplications create simultaneous copies of all genes contained within a genome as well as associated regulatory sequences. These duplication and the subsequent deletions of redundant coding and noncoding sequence both shape the natural evolution of plant genomes and provide a unique opportunity for researchers to characterize the regulatory sequences which determine when, in which cells and in what quantities the mRNA encoded for by particular genes will be produced.

This dissertation describes a model for explaining both bias in gene loss between parental subgenomes and the escape from preferential retention of duplicated genes between sequential whole genome duplications. Bias in gene deletion between individual duplicated segments had been previously observed by the publication of the sorghum and maize genomes provided an opportunity to demonstrate this bias was a consistent mark distinguishing whole pairs of ancestral chromosomes, and that ongoing gene loss remains consistently biased between high and low gene loss subgenomes millions of generations after a whole genome duplication. Bias in both ancestral and ongoing gene loss is shown to be correlated with biased gene expression between parental subgenomes with genes on the low gene loss subgenome tending to show higher expression levels than duplicate copies of the same genes on the high gene loss subgenome. This phenomena, originally referred to as genome dominance, although the literature has since become somewhat confused, provides an explanation both for biased gene loss between parental subgenomes and for the escape of deletion-resistant genes from the ratchet of ever increasing copy numbers through continued whole genome duplications.

This dissertation also demonstrates the use of polyploid lineage – in this case maize – as a deletion machine to rapidly characterize the function of regulatory sequences shared by orthologous genes within a clade. It was possible to develop testable hypothesis about the

specific function of individual regulatory sequences by combining conserved noncoding sequence sequence datasets, noncoding sequence deletions identified using comparative genomics with analysis and visualization of gene expression data from diverse organs, tissues, and cell types. As a test of the accuracy of this method, a putative pollen specific enhancer of expression identified using expression data from maize was cloned from the orthologous sorghum gene and used to drive the expression of a reporter construct in *Brachypodium distachyon*. Polyploid deletion machines have the potential to radically accelerate the characterization of noncoding regulatory sequences, an area of genetics previously largely untouched by advances next generation sequencing technologies.

Dedication & Acknowledgements:

I am certain I would not be writing these acknowledgements today if it were not for Tom Brutnell. Tom hired a directionless freshman who wasn't even taking a single biology course to work in a maize genetics lab. I also owe thanks to the members of his lab I worked with and learned from over my years as an undergraduate: Keven Ahern, Patrice Dubois, and Tesfamichael Kebrom. I would also like to thank Eric Vollbrecht, Mei Guo, and Brad Barbazuk all of whom gave me opportunities to learn from them and briefly become part of their research groups during my summer breaks in college.

Here at Berkeley I have had the opportunity to work with and learn from countless gifted and supportive friends and colleagues. I owe a special debt of thanks to Brent Pedersen, Haibao Tang, and Eric Lyons who were my guides on the intimidating road from pipetting to programming. I need to give a heartfelt thank you to the members of the incoming PMB graduate class of 2008. I couldn't have asked for a better bunch of people to experience everything that graduate life can throw at a person with than you guys. I also need to thank my advisor Michael Freeling. Few graduate students are lucky enough to receive the freedom Mike has allowed me in mapping out and pursuing research interests. I have done my best to live up to the trust implicit in that freedom. I hope I have succeeded.

Finally I would like to dedicate this thesis to my parents: Patrick and Katherine Schnable. It is something of a cliché to call your own parents exceptional, but how many people's memories from before kindergarten include identifying cotyledons or having photosynthesis explained with lego blocks? For the time, the teaching, the confidence, and the independence, thank you both.

*-James Schnable, December 2012*

# Table of Contents

## Chapter 1: Introduction

The details of prior scientific knowledge will be covered within the introductions to individual chapters. But before getting into those details it is important to understand the technological changes which shaped the years during which this research was conducted.

Perhaps every generation of scientists to believe they are coming of age in an era where technology was changing the very way that science would be conducted, whether that technology is the polymerase chain reaction or Sanger sequencing, or the discovery of DNA as the molecular of heredity. And in a way each generation was correct. The technologies they saw emerge did change the way biologists studied everything.

For me, the technology that changed everything was second generation sequencing. I was a senior in college when the first human genome sequenced with 454's technology (that of James Watson) was published and a first year grad student a few months later when the first genome sequenced with Illumina's (then Solexa's) technology came out. And in the last five years these technologies really have changed everything.

This thesis deals first and foremost with comparative genomics, which, just like it sounds, is the comparison of the arrangement and sequence of DNA between different organisms (or in certain cases between different regions within the genome of a single organism). Obviously then, conducting most comparative genomics requires the known sequence of the genomes of different organisms.

Upon entering grad school the count of available plant genome sequences stood at six: arabidopsis, rice, poplar, grape, and the moss *Physcomitralla patens*. Each of these genome sequences represented years of work by hundreds of scientists and the investment of tens of millions of dollars in funding. Four short years later the count of sequenced genomes has grown to nearly fifty (Figure 1). Although the cost of sequencing has plummeted, most of the genome sequences available today still reflect the efforts of large consortiums of scientists spanning multiple institutions but the way genomes are sequenced is changing. The next wave of genome sequences are being produced by single labs.

At the same time second generation sequencing technologies are providing more plant genome sequences to compare than there are comparative genomicists to make the comparisons these technologies are also providing whole new types of genome-wide data. The binding sites of individual transcription factors throughout an entire genome can now be identified by sequencing fragments of DNA isolated by chromatin immunoprecipitation (ChIP-seq). DNA methylation is now measured with single base pair resolution through the sequencing of bisulfite treated DNA (BS-seq). Next generation sequencing is also allowing measurements of global small RNA population,

epigenetic histone modifications, and even identify unlinked genomic regions which interact in vivo. However, of most relevance to the research presented here is the measurement of gene expression levels through the sequencing of tens of millions of cDNA fragments (RNA-seq).

Using the wealth of both genome sequences and gene expression data which became available as I was starting out in graduate school I set out to address two problems, one a basic question of genome evolution and the other more applied. The first question involved finding an explanation for a startling observation my advisor had made several years prior, which was that duplicated segments of the arabidopsis genome experienced unequal rates of gene loss (Thomas *et al.* 2006). The second question also began with an existing project in the lab. Previous research had identified tens of thousands of non-protein coding sequences associated with genes which evolved more slowly than expected of functionless DNA (Thomas *et al.* 2007). While these sequences appeared to be functionally constrained and evidence pointed to the fact that they generally functioned in gene regulation, it was unclear how the specific function of individual sequences might be determined without years of expensive molecular biology and the creation of transgenic lines, something the lab was not set up for.

In the following chapters I will outline a generalized explanation for how whole genome duplications create subgenomes which remain functionally distinct after millions of years in ancient polyploid species and explain how these differences, observable in the expression patterns of duplicate genes in the first generation following whole genome duplication, explain the differences in patterns of gene loss between subgenomes observable after millions of generations. I will also touch on how these mechanisms may explain why plant genomes have not been overrun with certain classes of genes which are particularly resistant to the deletion of duplicate copies following whole genome duplication. In the final chapters I will explain how we may harness the power of whole genome duplications combined with the abundance of RNA-seq datasets being generated for other purposes to demonstrate the function of specific conserved noncoding sequences.

## Figures

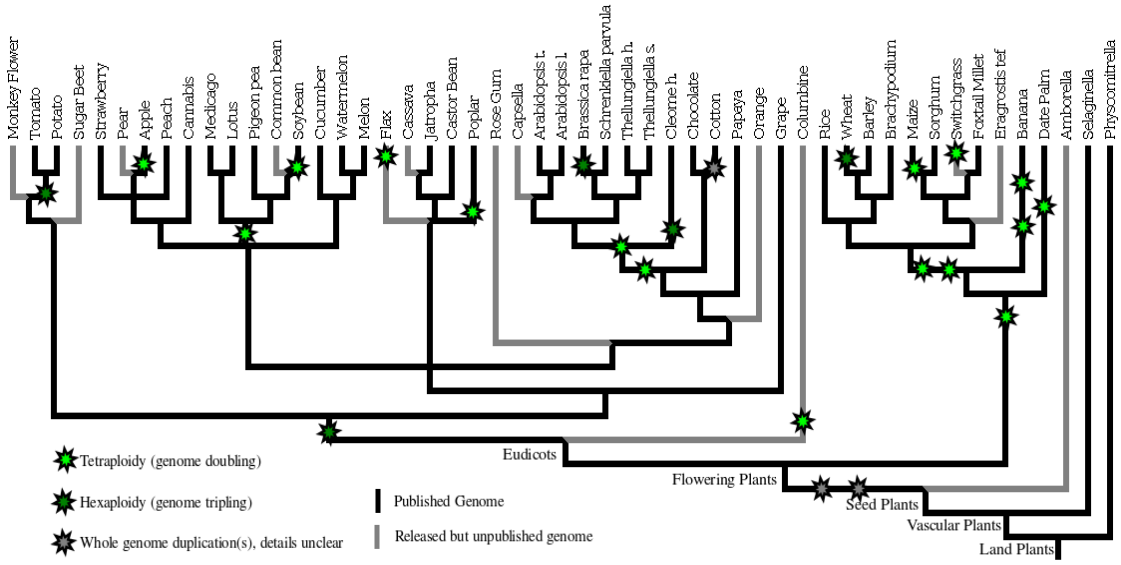



Figure 1
A phylogenetic tree showing the relationships between land plant species with sequenced genomes. Know whole genome duplications are marked with starbursts.

## Chapter 2: Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss

The following chapter (excluding the preface) has been published as a peer reviewed article in the Proceedings of the National Academy of Sciences:

Schnable JC, Springer NM, Freeling M. (2011) "Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss." Proceedings of the National Academy of Sciences, USA doi: 10.1073/pnas.1101368108



Contributions:

Nathan Springer provided pre-publication access to data on Presence Absence Variation among maize inbreds which was used in the analysis displayed in Figure 3A.

### Preface:

This paper -- and previous paper published nine months earlier in PLoS Biology of which I was a co-first author (Woodhouse *et al.* 2010) -- moved the study of whole genome duplication in maize from a comparison of individual genes or BACs to genome-wide comparisons, a jump made possible by the completion and publication of the maize genome in late 2009 (Schnable *et al.* 2009). The comparative genomics work of the Freeling lab had previously focused on Arabidopsis thaliana, the first plant species to be sequenced.

What drove the move? While the maize genome is much larger than that of arabidopsis and presented significant challenges to assemble correctly (see Chapter 3), the most recent whole genome duplication in maize was much younger than the “alpha” tetraploidy being studied in arabidopsis. As of now the maize whole genome duplication, which is actually shared by all the members of the Tripsacum and Zea genera (Bomblies and Doebley 2005), remains the most recent event represented by a species with a well assembled genome. Our hope was that the more recent maize event would provide more genomic traces of the sorts of sequence changes which immediately followed whole genome duplications. The change in model species paid off.

The previously mentioned PLoS Biology paper described for the first time the mechanism by which redundant copies of duplicated genes were removed following whole genome duplication. This paper was able to explain the observation from arabidopsis that genomic regions duplicated during whole genome duplications were

unequally gene-rich, suggesting that gene loss was unequal between duplicated regions (Thomas *et al.* 2006). In maize it was possible to reconstruct whole pairs of ancestral chromosomes both because of the younger age of the maize whole genome duplication and because, unlike arabidopsis, maize had a close relative which had diverged before the whole genome duplication to use as an outgroup (*Sorghum bicolor)*.

By comparing whole ancestral pairs of chromosomes and using multiple outgroup species to determine the ancestral gene content and order we were able to demonstrate that the bias in gene loss was consistent across each chromosome pair even when translocations and inversions had broken up these ancestral chromosomes within the modern chromosomes of maize. This ruled out many stocastic mechanisms and left me searching for some sort of genome-wide differentiation between the subgenomes.

Working in maize also provided access to a set of genes which had been determined showed Presence Absence Variation between different inbred lines (PAVs can be thought of as copy number variation which varies between one copy and zero copies). Using this dataset I determined that bias in gene loss continued to the present day in maize, which also ruled out a set of explanations which would only explain bias in gene loss in the early generations following a whole genome duplication.

Finally my interest settled on a phenomena which was, at the time, referred to as genome dominance (Flagel *et al.* 2008; Flagel and Wendel 2010) – although the literature has since become contradictory on the term (Grover *et al.* 2012). Throughout this thesis the term will be used to refer to the unequal expression of duplicate genes between two subgenomes within an ancient or recent polyploid with the majority of high expressed copies coming from one parental genome. Using RNA-seq data from four datasets I was able to demonstrate that the maize subgenomes showed genome dominance. The presence of genome dominance in maize millions of generations after the completetion of whole genome duplication a possible explanation for the bias in gene loss observed in most ancient polyploids, as explained in the body of this chapter.

Those interested in a more detailed explanation of the techniques used to identify orthologs and homeologs in chapters 2 & 3 should refer to chapter 4.

## Introduction:

Genomes that have experienced ancient polyploidy show nonequivalence between duplicated genomic regions. The most easily observed aspect of this nonequivalence is that one copy of a duplicated region will retain more genes, while the other copy of that same region will lose more genes, a phenomenon known as fractionation bias. This bias in gene loss and retention between duplicated genome segments was first observed in arabidopsis (Thomas *et al.* 2006), more recently in maize (Schnable *et al.* 2009; Woodhouse *et al.* 2010), and is probably a general characteristic of post-tetraploid eukaryotic genomes (Sankoff *et al.* 2010). While the proximate mechanism of gene loss following the whole genome duplication in maize has been shown to be a short deletion mechanism (Woodhouse *et al.* 2010), this mechanism does not explain why genes from one genome segment should be more likely to be lost than their homeologs (syn. homoeologs, ohnologs, syntenic paralogs) in the duplicate region of the genome.

A second form of nonequivalence between duplicated regions, in fact between whole genomes, has been shown in studies of more recent allotetraploid species. Wang and coworkers in the Z. J. Chen laboratory used 70-mer oligo microarrays to measure gene expression differences in a synthetic allotetraploid of *Arabidopsis thaliana* and *A. arenosa*, and, compared these results to midpoint values of gene expression in the two parents (Wang *et al.* 2006). They showed that genes originating from *A. arenosa* tend to dominate over homeologous genes from *A. thaliana* by contributing more to total gene expression in the allotetraploid. The same pattern of genome dominance was observed for the recent natural allotetraploid *Tragopogon miscellus*, a species estimated to have originated less than 80 years ago (Ownbey 1950). The Barbazuk and Soltis labs sequenced leaf RNA from *T. miscellus* and found the higher expressed members of differential expressed gene pairs were more likely to carry SNPs shared with *T. dubious* than with the other diploid parental species, *T. pratensis* (Buggs *et al.* 2010). Tetraploid cotton species originated in an allotetraploid event between diploid species carrying A and D genomes with an estimated age of 1 and 2 million years (Senchina *et al.* 2003). Data from these species provide evidence that genome dominance persists over much longer time scales. L. E. Flagel and J. F. Wendel used petal RNA hybridized to microarrays with probes specific to genes originating in the A or D cotton genomes to show that, while many gene pairs are expressed contrary to the prevailing pattern, genes originating in the D genome are more likely to contribute a majority of total gene expression than their homeologs from the A genome in five allotetraploid cotton species and a synthetic hybrid between diploid cotton species containing the A and D genomes (Flagel and Wendel 2010).

Genome dominance has not be observed in studies of any of the more ancient plant tetraploidies. Studies of the expression patterns of homeologous gene pairs originating from the the arabidopsis alpha tetraploidy, estimated to have occurred 25-40 million years ago, found no systematic pattern of dominant expression (Blanc and Wolfe 2004).

Similarly studies of gene expression patterns across homeologous regions in rice, originating from a duplication estimated to have occurred 50-70 million years ago (Paterson *et al.* 2004), report no evidence of genome dominance (Li *et al.* 2006). It appears homeologous gene pairs in both rice and arabidopsis are often differentially expressed (Blanc and Wolfe 2004; Throude *et al.* 2009). Please note that the ability of these studies to resolve subtle differences was limited by the inability to assign duplicated segments to specific ancestral genomes, so analyses were carried out on individual homeologous segments.

We use comparative analysis of the maize and sorghum genomes to examine the differentiation of duplicated genomic regions following the maize tetraploidy. Both grass species are members of the tribe Andropogoneae, and the genomes of both species have been sequenced (Paterson *et al.* 2009; Schnable *et al.* 2009). The lineage leading to maize experienced a tetraploidy sometime after the divergence of the two lineages while sorghum remained diploid. An unduplicated outgroup is essential for identifying highly fractionated duplicate genome segments as well as differentiating between recently transposed genes and genes lost from one duplicated segment but retained in the other (Kellis *et al.* 2004). The two genomes of maize split from each other approximately 12 million years ago, contemporaneous with but following the split between the maize and sorghum lineages, either as the result of autotetraploidy or allotetraploidy (Swigoňová *et al.* 2004). The maize tetraploidy, which combined both genomes within one nucleus and began the process of genome fractionation, occurred between 5 and 12 million years ago (Swigoňová *et al.* 2004; Schnable *et al.* 2009). The genome of maize shows evidence of ongoing gene loss (Woodhouse *et al.* 2010), making it an excellent model to study the mechanism of differentiation between duplicated genomic regions.

We show that fractionation bias results from the differentiation of entire ancestral chromosomes and suggest that this chromosomal differentiation reflects differences between the two parental genomes, with one genome being dominant at the level of gene deletion resistance and RNA expression. Biased loss of genes does not appear to be a result of inherent differences in deletion rates between homeologous regions because “silent” deletions, deletions in DNA that are usually without specific function -- such as those from introns and retrotransposons -- show no bias between ancestral chromosomes. Given the correlation observed between the subgenome which dominates expression in maize and the ancestral chromosomes which have experienced less gene deletion, we propose that deletions of duplicate genes from the less expressed subgenome may be less likely to result in reduced fitness. This hypothesis makes sense in light of the gene balance hyopthesis as will be discussed. Following tetraploidy, deletions from one subgenome would be more likely to be removed by purifying selection, while deletions from the opposite subgenome would be more likely to be selectively neutral.

## Results

### Reconstruction of chromosome level organization in the newly tetraploid ancestor of maize: defining two subgenomes.

It was inferred from multiple studies that the ancestral genome of the Andropogoneae consisted of ten chromosomes. The genome of sorghum is presumed to have approximately retained this ancestral arrangement, while the ten chromosomes of maize represent a reduction from a twenty chromosome tetraploid ancestor by chromosome fusion (WEI *et al.* 2007; SALSE *et al.* 2008). Given the small total divergence time between maize and sorghum and the fact that tetraploidy can temporarily increase the frequency of genome rearrangements (KASAHARA *et al.* 2007), the sorghum genome was treated as representative of the genome organization of both diploid genomes present in the initial tetraploid ancestor of maize.

Using whole genome dotplots color-coded by synonymous base pair substitution rates (Figure 1; plotted using CoGe software), it is possible to reconstructed the original duplicate regions within the maize genome on the basis of orthology to the ten sorghum chromosomes (Table S1). The synonymous substitution rates of individual gene pairs do not permit genes to be unambiguously classified as orthologs or ancient homeologs. However the median synonymous substitution rate of all gene pairs in a syntenic block between maize and sorghum can be used to unambiguously classify syntenic blocks of 12 or more genes as orthologous or homeologous (Figure S1; the teal color marks syntenic blocks derived from the ancient pregrass tetraploidy, and are avoided).
Inversions and other intra-chromosomal rearrangements are presumed to be more common than translocations between different chromosomes. Therefore, segments of a maize chromosome orthologous to the same sorghum chromosome are assumed to come from the same chromosome copy in the tetraploid ancestor maize. For five sorghum chromosomes at least both full ancestral copy can be reconstructed in the maize genome using this method. For the remaining five, one full ancestral copy was reconstructed based on all orthologous segments being present on a single maize chromosome, and and the remaining orthologous segments located on two – or in one case three – maize chromosomes were grouped together by process of elimination (Table S1). There are no cases where both duplicate copies of the region were located on the same chromosome. Our assumptions and reconstruction are largely concordant with previous ancestral reconstructions of the maize genome (WEI *et al.* 2007; SALSE *et al.* 2008).

For each pair of reconstructed chromosomes, one copy retained substantially more syntenic genes than the other. Bias in gene loss between pairs of reconstructed chromosomes was consistent across their entire lengths (Figure 2). For each pair of chromosomes, the copy which possessed a greater number of unique genes retained orthologously in both rice and sorghum was assigned to the maize1 subgenome, while the pair with fewer uniquely retained genes was assigned to the maize2 sugenome. Gene

counts and the statistical significance of the differences between copies are listed in Table S1. Maize1 and maize2 each constitute a genome orthologous to the entire sorghum genome. The distribution of these two genomes across the ten modern chromosomes of maize is displayed in Figure S2.

**Ongoing fractionation among 33 *Zea mays* accessions remains biased**

Using only maize genes with retained syntenic orthologs in both sorghum and rice, we constructed two lists of high confidence genes, the list of retained homeologs from the maize duplication and the list of genes where it was possible to say with high confidence that the duplicated copy was lost from the genome – singleton genes. These lists will be referred to as "retained homeolog" and "lost homeolog." Each of these gene lists is further subdivided into maize1 specific and maize2 specific lists of genes. A complete description of the criteria used to identify these two high-confidence gene sets is included in Methods. There is no significant difference in the annotated length of coding or non-coding sequences between homeologous copies of genes retained in both maize1 and maize2 subgenomes (Figure S3).

A recently published dataset documents presence/absence variation (PAV) of genes among 19 diverse maize inbreds and 14 teosinte lines using carefully controlled comparative genomic hybridization (Swanson-Wagner *et al.* 2010). Among our high-confidence lost homeolog gene sets, equal percentages of maize1 and maize2 genes were identified as lost from the genomes of one or more inbreds. However, among our high-confidence retained homeolog gene set significantly more of the genes located on maize2 were identified as lost from one or more inbreds than were the duplicate copies of those same genes located on maize1 ($p = .0043$, chi-square, df=2) (Figure 3A). PAV data indicates ongoing fractionation remains biased in modern maize inbreds.

Maizesequence.org has released at least two sets of gene annotations. The maize possesses two sets of gene model. The filtered gene set (FGS) contains ~32,000 genes considered to be of higher confidence, while the working gene (WGS) set contains over 100,000 genes including the genes of the filtered gene set as well as many likely pseudogenes, gene fragments, or transposon related proteins. Genes unique to the working gene set have a similar distribution to those genes which show presence/absence variation between maize inbreds. Ongoing fractionation by short deletions has been shown to produce truncated gene fragments prior to their complete removal (Woodhouse *et al.* 2010) exactly the sort of sequence which might be annotated as a gene, but excluded from the filtered gene set. The distribution of genes found only to the maize working gene set supports the conclusion that biased fractionation in the maize genome is ongoing. First, syntenically retained working set genes are more likely to possess a retained homeolog, which is presumably the undamaged full length gene copy (Figure 3B). Second, in these cases the low confidence gene found only in the working gene set is more likely to be the copy located in the maize2 subgenome (Figure 3B). This paragraph

describes the only portion of our study where we did not exclude the low confidence genes found only in the working gene set.

**Deletions within noncoding sequences show no bias between maize1 and maize2**
Maize1 and maize2 subgenomes cover significantly different fractions of the total maize genome – 1.26 gigabases and 0.75 gigabases respectively. As coding sequences of annotated genes, including the working gene set, account for less than 5% of the total maize genome -- and transposons account for 85% (Schnable *et al.* 2009)-- this bias in total genomic size would seem to imply that biased fractionation acts on all genomic DNA not simply coding sequence. However, the length of both coding sequences and noncoding sequences between high-confidence retained homeologous pairs on maize1 and maize2 are not significantly different (Figure S2). An analysis of 561 maize1 and maize2 introns that could be completely aligned to the orthologous sorghum intron identified an average of 6.03 deletions per intron in maize1 genes and 6.09 deletions per intron in the homeologous maize2 genes (Table S2). A similar analysis of deletions within copies of three of the largest families of retrotranposons within the maize genome – Huck, Opie, and Ji – which had inserted into maize1 or maize2 region of the genome found no difference in deletion frequencies for maize1 vs maize2 relative to an ancestral sequence for each family created from an alignment of multiple annotated transposon copies (Table S3, Figure S4).

**Expression differences between maize1 and maize2 homeologous genes**
Gene expression was measured for all genes included in the maize working gene set from the sequenced maize inbred B73 (Schnable *et al.* 2009) and RNA-seq data from four independent previously published datasets (Wang *et al.* 2009; Jia *et al.* 2009; Eveland *et al.* 2010; Li *et al.* 2010) (Table S4). Expression data was calculated in units of frequency of aligned reads per kilobase of exon per million reads (RPKMs) using the Bowtie and Cufflinks packages (Langmead *et al.* 2009; Trapnell *et al.* 2010). Cufflinks distributes reads that were found to aligned equally well to multiple gene models proportional to the relative expression rates for those genes calculated from reads with only one best alignment (Trapnell *et al.* 2010) This combination of programs allows us to deal with the ambiguity created by the small fraction of sequences which align equally well to both homeologs within the maize genome.

The expression of gene pairs included in the high confidence “retained homeolog” set described above were compared using each expression dataset. In each dataset the number of pairs where the maize1 homeolog dominated total gene pair expression outnumbered the number of pairs where the maize2 homeolog dominated expression. This bias was robust, appearing whether we defined dominance as any measurable difference in expression (Figure S5) at least two-fold difference in homeolog expression (Figure 4), four-fold difference in homeolog expression (Figure S6). The bias towards gene pairs dominated by expression of the maize1 copy remains consistent across a range

of cutoffs for the expression of the nondominant homeolog. At cutoffs as high as 30 reads per kilobase of exon per million reads (RPKM) for the less expressed gene copy, maize1 homeologs continued to disproportionately dominate expression in all parts of the maize plant examined (Figure S7). Biased expression is also observed when examining individual pairs of reconstructed chromosomes (Figure S8; effectively independent replicates of our experiment). The median difference in expression between homeologs ranges from 1.8 to 2.8-fold in different expression datasets. In every expression dataset, the median difference between homeologs where the maize1 gene is expressed at a higher level is marginally higher than the median difference for pair where maize2 is expressed at a higher level (Table S5).

**Discussion**

Biased gene loss is clearly not a transient phenomenon that occurred only in the early generations following the tetraploidy in maize. Rather, biased gene loss is a reflection of a significant differentiation of two complete subgenomes within a tetraploid lineage, and these differences are stably inherited over millions of generations. The link we observe between the biased gene loss and biased expression is likely not unique to the maize tetraploidy. A recent study of a 1 megabase region of the common bean (*Phaseolus vulgaris*) and the two co-orthologous regions of the soybean genome also found that the homeologous region with more syntenically retained genes tended to be expressed at higher levels (Lin *et al.* 2010). While we have shown that bias in the loss of duplicate gene copies continues in the maize lineage, as it presumably has for the last several million years, evidence from deletions in introns and retrotransposons suggest that this bias is not the result of fundamentally different frequencies of sequence deletion between maize1 and maize2 chromosomal segments. The equivalent deletion rates we observe for both subgenomes is concordant with our finding that single copy genes on either subgenome are equally likely to be identified as showing presence absence variation between inbreds.

Our data suggest a model in which deletions in both maize genomes occur at the same overall rate, but purifying selection is more likely to remove deletion alleles of higher-expressed duplicate copies from the population, while the loss of less-expressed homeologs are more likely to be selectively neutral or near-neutral when the higher expressed copy remains present in the genome. This model is consistent with selection against changes in the balance of gene products, as reviewed (Sémon and Wolfe 2007b; Freeling 2009; Edger and Pires 2009; Birchler and Veitia 2010). Our model states that smaller changes in total gene pair expression (maize1 transcript + maize2 transcript) are more likely to be tolerated than larger changes. The removal of a singleton gene, whether it is located in maize1 or maize2, involves the complete loss of that gene product. Because the effect of the loss of a singleton would be the same regardless of genomic location, no bias would be predicted for these genes and we detected no bias. Our control experiments showing deletions within transposons and most deletions within introns are

unbiased between maize1 and maize2 demonstrate maize1 and maize2 have no inherent difference in mutability. This result is consistent with our model that biased fractionation is a result of purifying selection acting preferentially against deletion alleles of gene copies that contribute more to total gene pair expression.

There are precedents for the idea that changes in total gene product dosage often lower fitness. Genes encoding proteins with more interaction partners -- such as protein kinases and phosphatases, or subunits of complex machines like ribosomes, proteasomes, and motors -- are predicted (as reviewed (Birchler *et al.* 2007)) to be more dosage-sensitive, and these are precisely classes of genes are more likely to be retained as homeologous pairs following tetraploidy (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Maere *et al.* 2005; Thomas *et al.* 2006). Greater changes in total gene product dosage have also been show to be more likely to impact fitness nagatively in the absence of tetraploidies. For example, the loss of highly expressed gene copies in yeast have been shown to be more likely to significantly impact fitness than the loss of their less expressed paralogs (Gu *et al.* 2003). Knockouts of duplicate genes in yeast with similar levels and patterns of expression – those presumed to be the most dose sensitive -- have been shown to share similar patterns of epistatic relationships, demonstrating the loss of either equally expressed duplicate gene impacts function in a similar way (Vandersluis *et al.* 2010).

While the maize lineage tetraploidy occurred 5-12 million years ago, the latest transposon blooms in maize occurred only in the past few million years (as reviewed (Bennetzen 2007; Schnable *et al.* 2009)). It is conceivable that the gene contents of maize1 and maize2 genomes were already significantly different at the time of this most recent transposon bloom. *Opie* and *Ji* have both been shown to preferentially insert into heterochromatin near genes (Baucom *et al.* 2009) suggesting that over time transposons insertions will tend to track total gene content. We hypothesize that transposons inserted into maize1 and maize2 in approximate proportion to the gene content of these regions. If this were indeed the case, the difference in mobile, dispensable DNA between the two genomes is simply an artifact of preexisting differences in gene content. Further experiments are necessary to fully evaluate the degree to which selection can explain the many differences between the two maize genomes, but it is remarkable that selection frequently differentiates between relatively minor levels of gene expression. The general concept of expression thresholds, so common in discussions of allelic dominance and recessiveness, has not proven useful in interpreting our data.

The explanation of biased fractionation by genome dominance leaves unanswered the question of the mechanism behind the origin and maintenance of genome dominance. The most likely candidate remains differential epigenetic marking of genomes within an allotetraploid. Allotetraploidy has been show to produce epigenetically inherited differentiation of parental genomes (Lee and Chen 2001; Wang *et al.* 2006; Chen 2007). There is no conclusive evidence to support either an auto- or allotetraploid origin for

maize, although one study found ZFL2 may be more closely related to orthologs in in the Andropogoneae genera *Coelorachis* and *Elionurus* than to the duplicate homeolog in maize ZFL1 (Bomblies and Doebley 2005). While there is currently a dearth of high quality epigenetic data for maize available in published literature, ongoing research projects are likely to remedy this situation in the near future, thereby illuminate the mechanism responsible for differentiation of maize1 and maize2 gene copies.

Whatever the mechanism, an event occurred early in the process of tetraploidy that differentiated the two parental genomes of maize, maize1 and maize2. We have shown that these differences have persisted through millions of generations, and continue to impact both gene expression and the pattern of ongoing gene loss in maize. Ongoing fractionation by the mechanism we describe here provides an explanation as to why *Zea mays* is particularly genetically diverse.

## Methods

### Identification of orthologous and homeologous genes.

Syntenic blocks were identified between and within grass genomes using the SynMap application withing CoGe, an online comparative genomic toolbox (Lyons *et al.* 2008b). Syntenic blocks were assigned to specific evolutionary events, either speciation (orthology) or whole genome duplication (homeology) based on the median synonymous substitution rates of genes within a syntenic block. Maize genes scored as orthologous to sorghum genes were assigned to reconstructed ancestral chromosomes according to the arrangement shown in Table S1.

### Identification of high confidence retained homeolog and no homeolog genes

High-confidence genes were considered to be the subset of the maize filtered gene set with annotated start and stop codons whose gene models were supported by expression data (cDNA and/or EST) (27,313 of the 32,540 genes in the maize filtered gene set satisfied these criteria). We further required that it was possible to identify a retained syntenic orthologs in both the rice and sorghum genomes (14,855 of the 27,313 genes), and possess an identifiable homeologous location within the maize genome (13,844 of 14,855 genes). Genes with a history of tandem duplication in rice, sorghum, maize1 or maize2 were eliminated from the analysis, as these genes are expected to show greater rates of copy number variation, create problems for comparative expression studies, and confuse all arguments involving selection (9536 of of 13,844 genes). Finally two-high confidence sublists were created. High-confidence retained homeologous pairs are those pairs where there are genes that satisfy all the above criteria and are present at both locations in the genome (1750 genes in both maize1 and maize2). High-confidence no homeolog genes are those that satisfy all the above criteria, excluding those genes were a homologous working set or other low confidence gene is present at the homeologous location in the genome as well as those genes where an unannotated syntenic blast hit was detected as the homeologous location in the genome. (3617 genes located in maize1

and 1577 genes located in maize2). A total of 842 genes which satisfied all criteria above were disqualified from inclusion in either the high-confidence retained homeolog or high-confidence no homeolog lists because of a homeolog  made ambiguous by being either a low confidence gene or unannotated syntenic blast hit.

**Calculation of gene expression levels**

Gene expression data was calculated from mRNA-seq data published by four different laboratories, Deng, Brutnell, Schnable and Jackson (Wang *et al.* 2009; Jia *et al.* 2009; Eveland *et al.* 2010; Li *et al.* 2010) (Table S4). For all expression sets except immature ears, reads were aligned to the maize genome using Bowtie allowing one mismatch per read and disregarding reads with more than two best alignments (Langmead *et al.* 2009). Expression values were calculated in units of reads per kilobase of exon per million reads (RPKM) using Cufflinks (Trapnell *et al.* 2010) using the published annotations of the B73 refgen_v1 working gene list (Schnable *et al.* 2009). Immature ear expression data were generated using a digital gene expression technique. For this expression dataset, collapsed reads were aligned to the genome using Bowtie, disregarding all alignments with 1 or more mismatches and all alignments with more than one unique alignment in the genome. For immature ears, expression values for each gene were calculated as the sum of the number of reads represented by each collapsed read mapping within a window starting 300 bp upstream of the start of the gene model and extending 300 bp downstream of the gene model. Final gene expression values were calculated in units of reads per million reads (RPM).

**Figures**

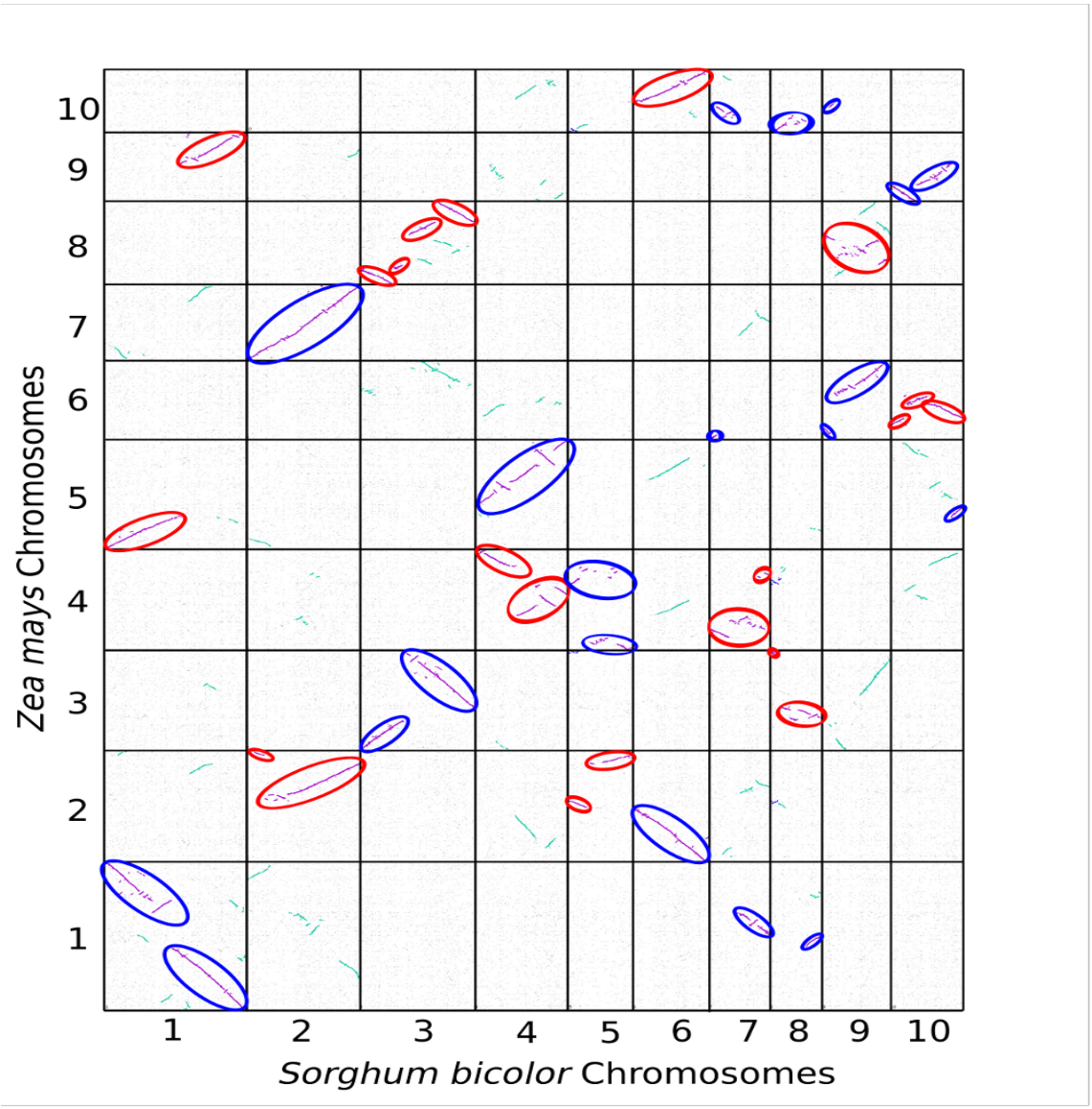



**Figure 1**
A dotplot comparison of the maize and sorghum genomes. Each dot marks a pair of genes one in sorghum and one in maize, identified as homologs in a blast comparison. Genes with conserved syntenic gene order are highlighted in color. Orthologs from the maize sorghum split were distinguished from homeologs from the pre-grass duplication by synonymous substitution rate (Ks). Orthologs are marked in purple (lower Ks), pre-grass homeologs are marked in teal (higher Ks). The regions making up one complete ortholog of each sorghum chromosome in the maize genome are circled in blue, the regions making up the other complete ortholog are circled in red. The original dotplot from which

this figure was created was produced using CoGe software, and can be regenerated at http://tinyurl.com/2am77tn by clicking "Generate SynMap".

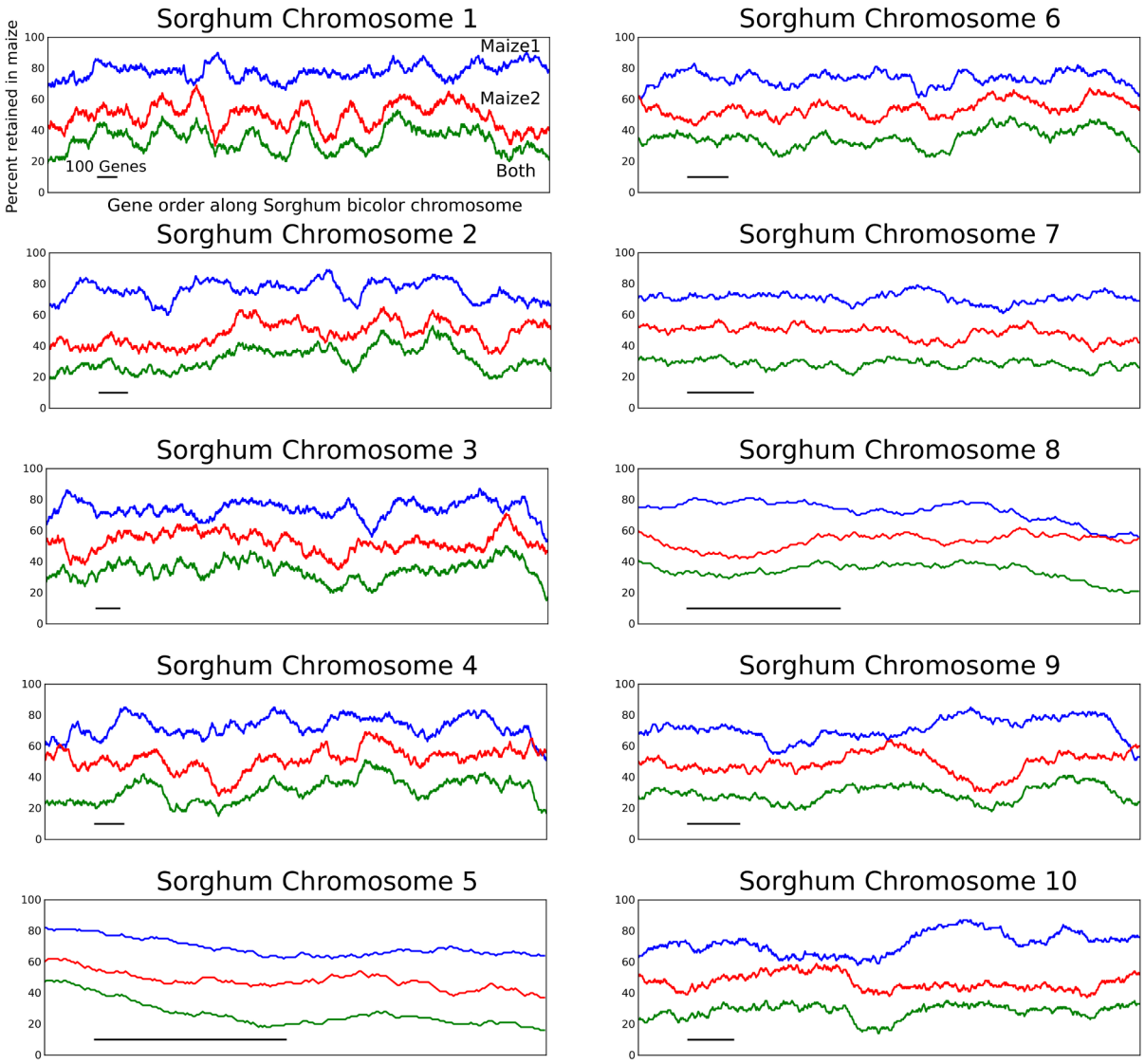



Figure 2
Biased fractionation is observed for each reconstructed, or “sorghumized”, pair of maize ancestral chromosomes. Bias is measured as the number of conserved genes out of 100 in a sliding window (black bars) of genes conserved syntenically between sorghum and rice (y-axis) and displayed based on the gene order along sorghum chromosomes (x-axis). Conservation of genes on reconstructed chromosomes assigned to maize 1 is shown in blue. Conservation of genes on reconstructed chromosomes assigned to maize2 is shown in red. The proportion of genes retained on both reconstructed chromosomes is shown in green.

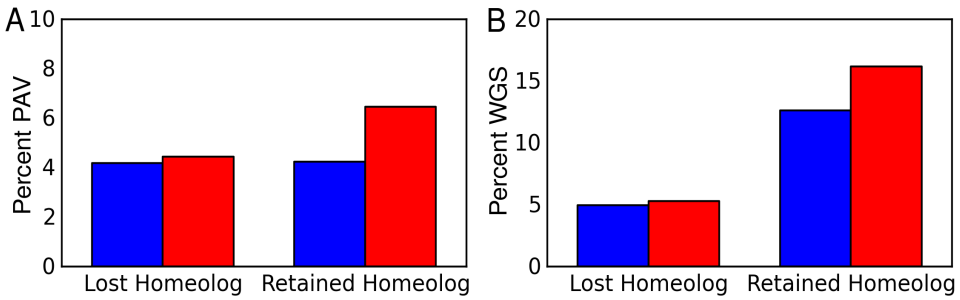



Figure 3

Multiple measures of ancient and ongoing fractionation. A. Percent of high confidence maize genes (see Methods) which exhibited presence absence variation in a study of maize inbreds and teosinte accessions. B. Percent of all annotated maize genes conserved syntenically in both rice and sorghum which are excluded from the maize filtered gene set.

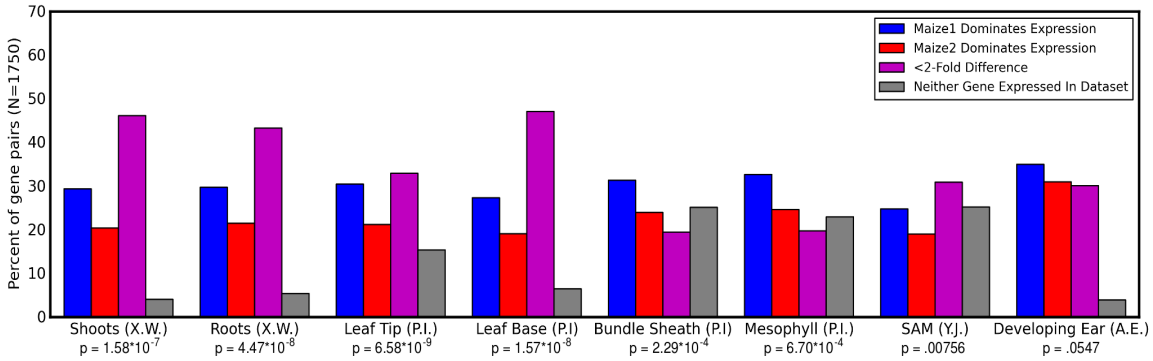



Figure 4

Patterns of expression for the 1750 best confidence (see Methods) pairs of maize homeologs in eight organ systems, organs or cell types. Homeologs were considered to be differentially expressed if the expression of one homeolog was at least twice the expression of the other. RNA-seq data was from X.W. (first author's initials) (21), P.I = (22) Y.J = (23) A.E = (24). All p-values calculated using cumulative binomial distributions assuming equal chances of gene copies on maize1 or maize2 dominating total expression for the gene pair.

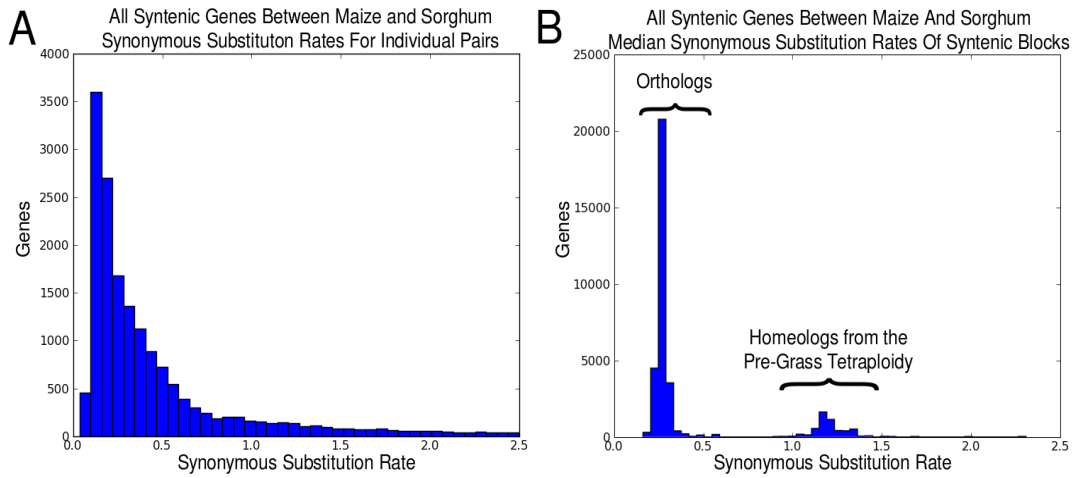



Figure S1:
Histograms showing the distribution of synonymous substitution rates when rates are calculated for individual gene pairs in syntenic locations (A) and when gene pairs are assigned a synonymous substitution rate based on the median rate for all gene pairs included in the same syntenic block (B).

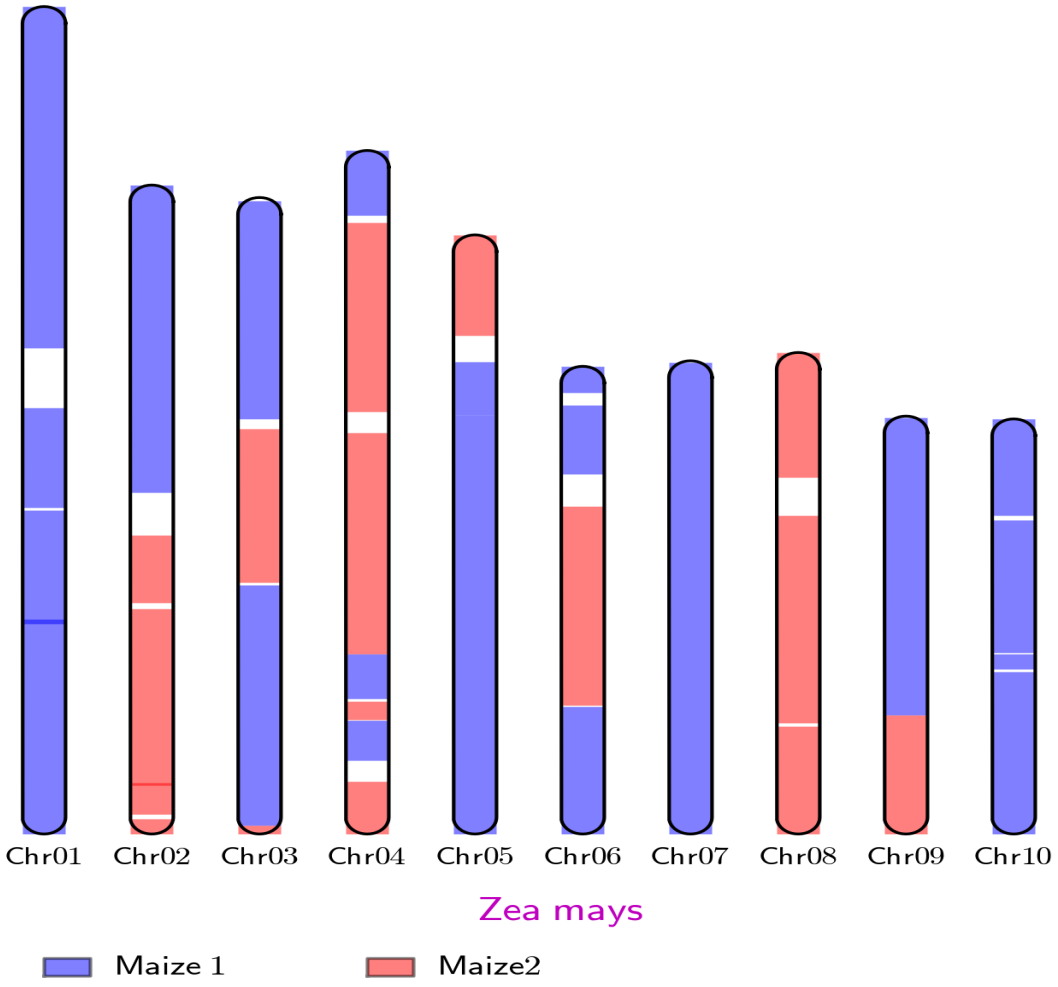



Figure S2
Distribution of the maize1 and maize2 subgenomes across the ten modern chromosomes of maize.

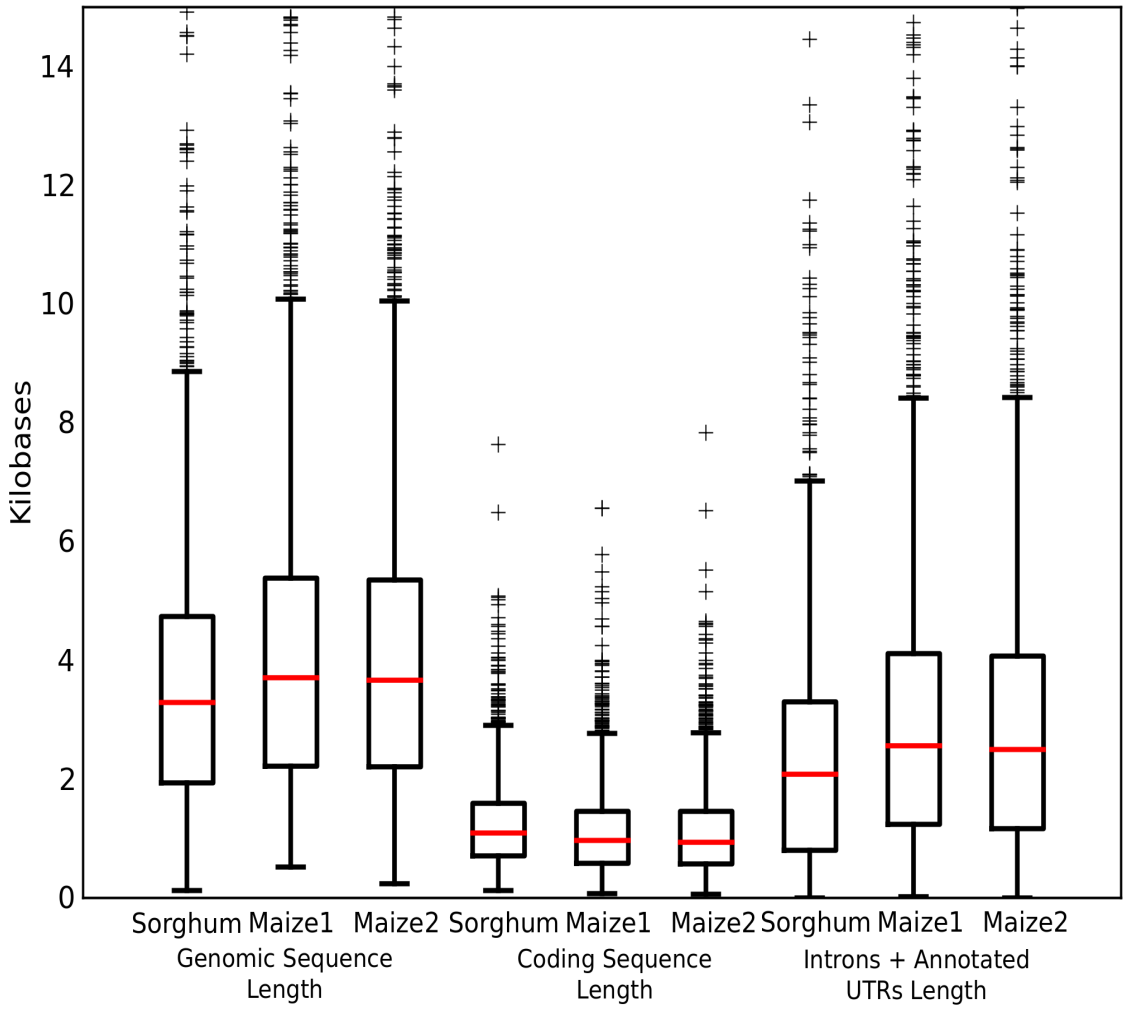



Figure S3
Distribution of gene lengths for the maize1 and maize2 copies of genes from the 1750 highest confidence maize homeologous pairs, as well as the distribution of lengths for the shared orthologs of these genes in sorghum.

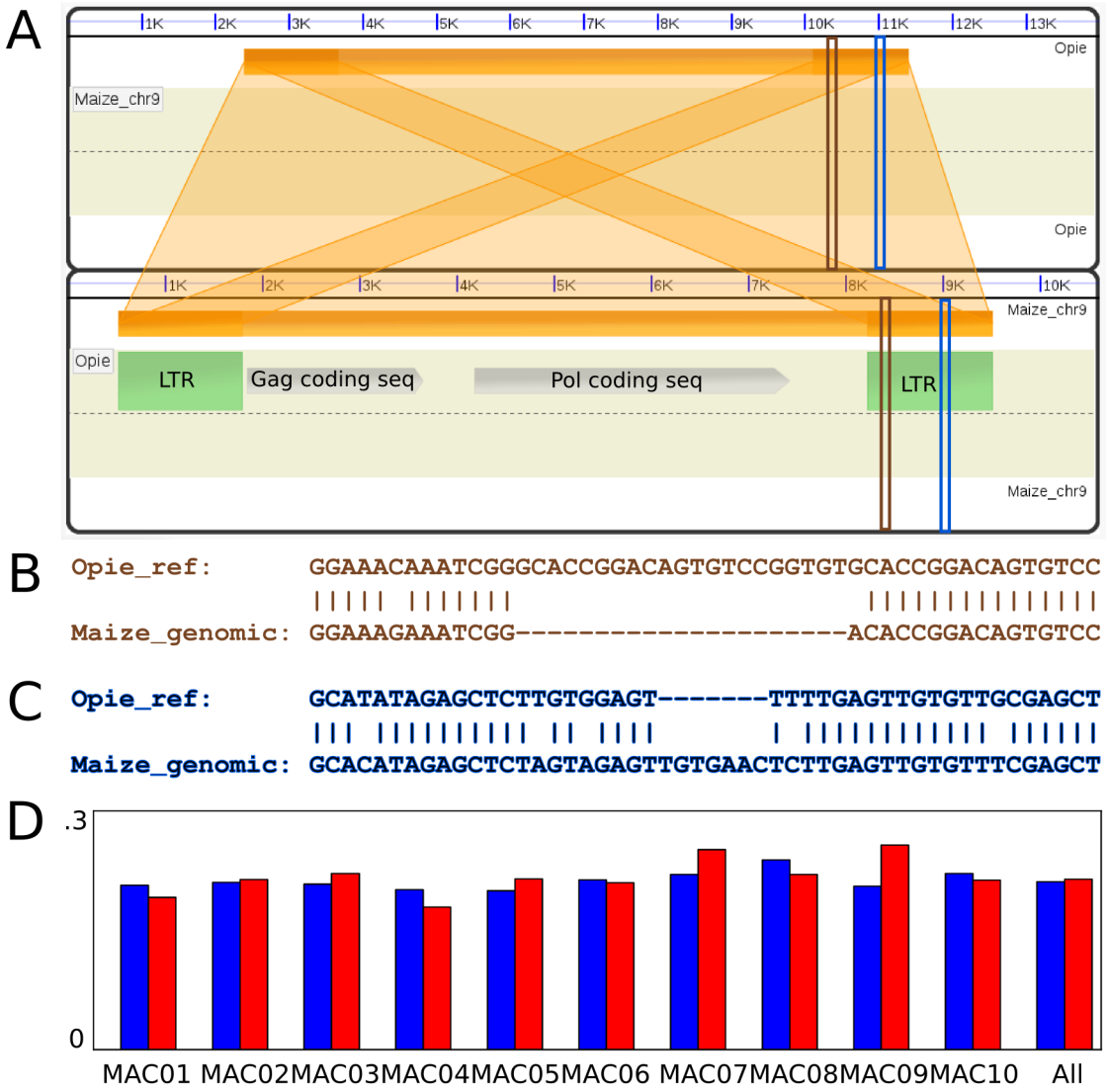



Figure S4:
Analysis of deletions from within retrotransposons. A. A schematic of the BLAST matches between a reference transposon (Opie-2 GENBANK ID: U68408.1 and a transposon of the Opie family located at between 64,485,456 and 64,499,425 of maize chromosome 9 (B73_refgen1). The comparison of the Opie reference transposon and B73 genome sequence can be examined using this link: http://genomevolution.org/r/rva B. An example of a putative deletion from the genomic maize transposon, which would be scored as a deletion. C. An example of of a putative insertion in the genomic maize transposon although it's also possible this represents a deletion from the reference sequence. This gap would not be scored as a deletion. D. Average number of deletions identified per transposon fragment for the maize1 and maize2 versions of each reconstructed ancestral chromosome.

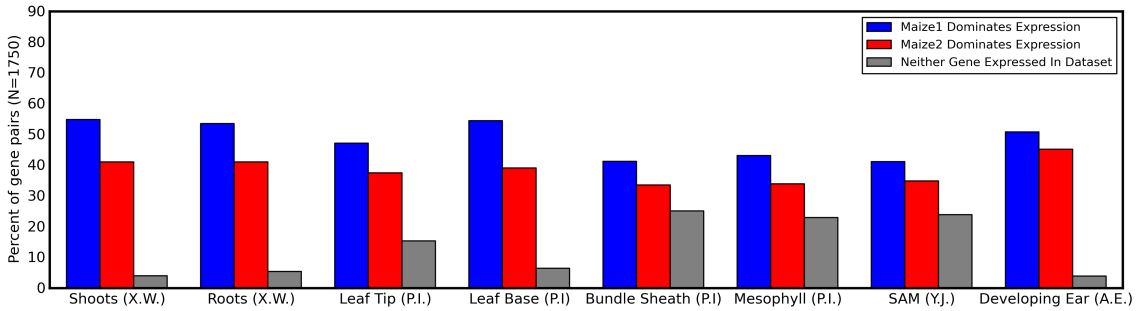



Figure S5:
Distribution of 1750 high-confidence gene pairs between unbiased, maize1 dominated and maize2 dominated when requiring a 4-fold difference in expression between homeologs to classify one homeolog as dominant.

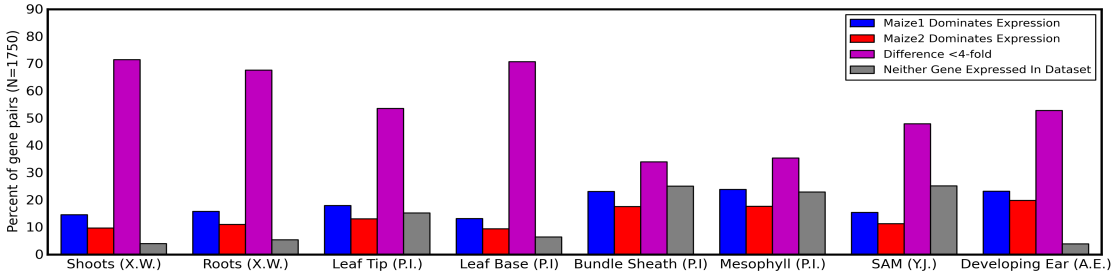



Figure S6:
Distribution of 1750 high-confidence gene pairs between unbiased, maize1 dominated and maize2 dominated when any difference in expression between homeologs to is sufficient to classify one homeolog as dominant.

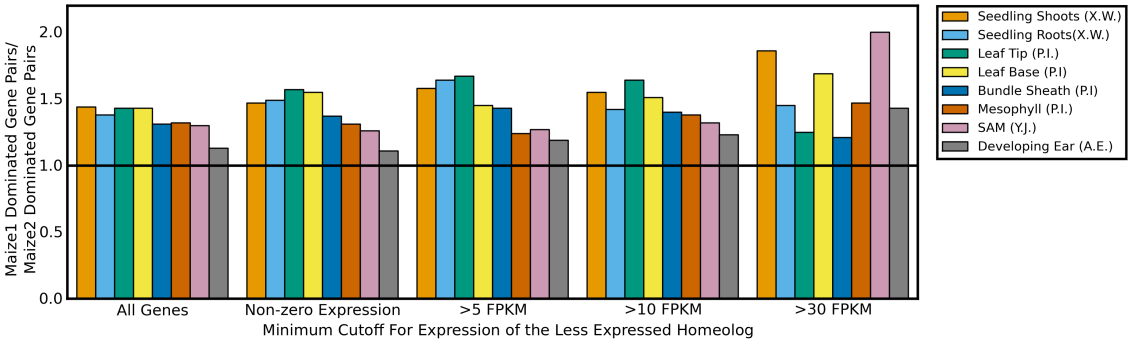



Figure S7:
Effect of applying higher cut offs for minimum expression rates of the less expressed homeolog. Data is displayed as number of gene pairs where expression of the maize1 homeolog dominates/ number of gene pairs where expression of the maize2 homeolog dominates. Expression cut offs are defined in units of reads per million for developing ears (as this was a digital gene expression experiment) and reads per kilobase of exon per million reads for all other expression datasets.

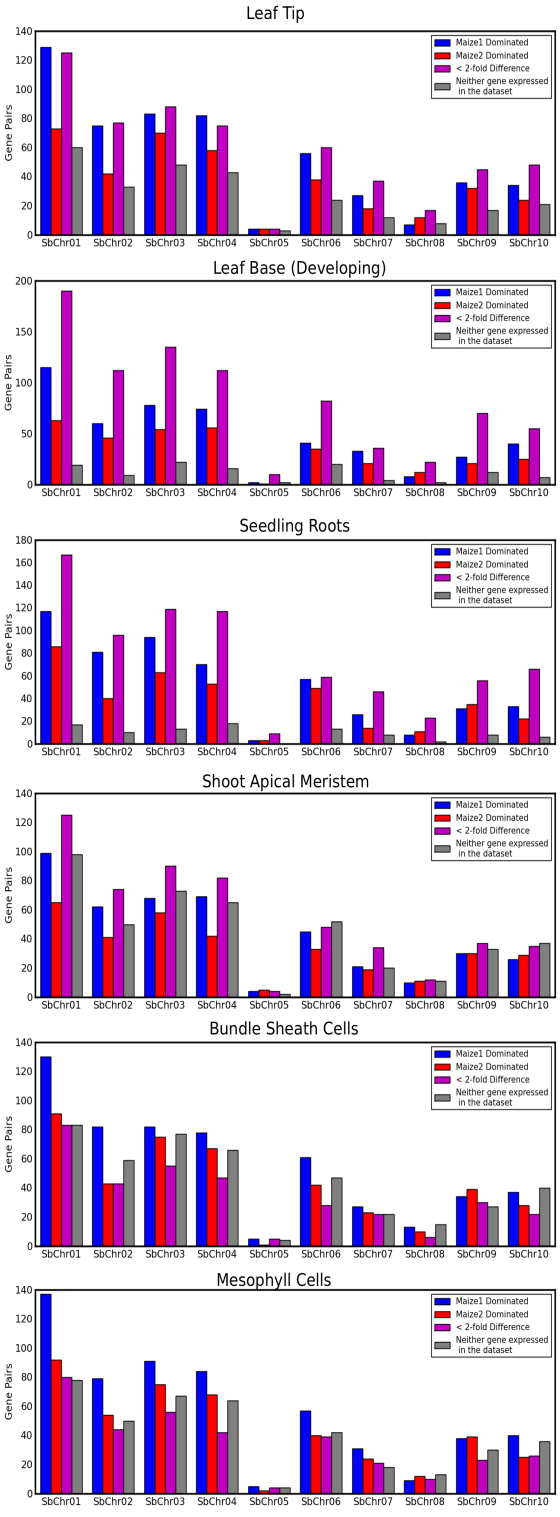

Leaf Tip
Leaf Base (Developing)
Seedling Roots
Shoot Apical Meristem
Bundle Sheath Cells
Mesophyll Cells
Gene Pairs
Maize1 Dominated
Maize2 Dominated
< 2-fold Difference
Neither gene expressed in the dataset
SbChr01
SbChr02
SbChr03
SbChr04
SbChr05
SbChr06
SbChr07
SbChr08
SbChr09
SbChr10

Figure S8:
Individual expression analysis for each pair of reconstructed ancestral chromosomes in maize. Data is displayed for six expression datasets (leaf tip, leaf base, seeding roots, shoot apical meristems, bundle sheath cells and mesophyll cells). Pairs of maize chromosomes are identified by the number assigned to the orthologous sorghum chromosomes. Gene pairs with an expression difference >= 2-fold between homeologs were classified as differentially expressed.

**Tables**

Table S1: The reconstructed ancestral chromosomes of maize

| Sorghum chromosome | Location of Maize1 Ortholog | Location of Maize2 Ortholog | Pan-grass genes retained on both | Maize1 only pan-grass genes | Maize2 only pan-grass genes | p-value |
|---|---|---|---|---|---|---|
| Sb01 | Zm01 \| MAC-01-1 | Zm05, Zm09 \| MAC-01-2 | 850 gene pairs | 1071 genes | 395 genes | 1.79e-72 |
| Sb02 | Zm07 \| MAC-02-1 | Zm02 \| MAC-02-2 | 561 gene pairs | 720 genes | 251 genes | 2.20e-53 |
| Sb03 | Zm03 \| MAC-03-1 | Zm08 \| MAC-03-2 | 688 gene pairs | 787 genes | 368 genes | 7.97e-36 |
| Sb04 | Zm05 \| MAC-04-1 | Zm04 \| MAC-04-2 | 541 gene pairs | 672 genes | 332 genes | 1.65e-27 |
| Sb05 | Zm04 \| MAC-05-1 | Zm02 \| MAC-05-2 | 95 gene pairs | 129 genes | 56 genes | 4.12e-8 |
| Sb06 | Zm02 \| MAC-06-1 | Zm10 \| MAC-06-2 | 432 gene pairs | 447 genes | 222 genes | 1.13e-18 |
| Sb07 | Zm01, Zm06, Zm10 \| MAC-07-1 | Zm04 \| MAC-07-2 | 216 gene pairs | 298 genes | 141 genes | 2.61e-14 |
| Sb08 | Zm01, Zm10 \| MAC-08-1 | Zm03 \| MAC-08-2 | 129 gene pairs | 126 genes | 69 genes | 2.71e-5 |
| Sb09 | Zm06, Zm10 \| MAC-09-1 | Zm08 \| MAC-09-2 | 287 gene pairs | 385 genes | 193 genes | 5.34e-16 |
| Sb10 | Zm05, Zm09 \| | Zm06 \| MAC-10- | 308 gene pairs | 451 genes | 191 genes | 1.73e-25 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | MAC-10-1 | 2 | | | | |

Locations on the modern maize chromosomes of the regions orthologous to each sorghum chromosome, and the designation assigned to each reconstructed ancestral chromosome. p-values calculated with a null hypothesis that genes are equally likely to be deleted from either chromosome.

Table S2: Deletions observed within 561 aligned intron triplets between maize and sorghum

| | Mean Intron Length (SD) | Median Intron Length | Mean # of Gaps When Aligned to Sorghum (SD) | Median Gaps When Aligned to Sorghum | Mean Gap Length (SD) | Median Gap Length |
|---|---|---|---|---|---|---|
| Maize 1 | 804 bp (618 bp) | 649 bp | 6.03 gaps (4.47 gaps) | 5 gaps | 16.4 bp (24.1 bp) | 8 bp |
| Maize 2 | 787 bp (593 bp) | 651 bp | 6.09 gaps (4.79 gaps) | 5 gaps | 16.8 bp (25.0 bp) | 9 bp |

Table S3:

| | Huck-Maize1 | Huck-Maize2 | Ji-Maize1 | Ji-Maize2 | Opie-Maize1 | Opie-Maize2 |
|---|---|---|---|---|---|---|
| Fragments Identified | 8392 | 5266 | 6752 | 4225 | 9059 | 5771 |
| Mean Fragment Length | 3839 bp | 3832 bp | 3329 bp | 3246 bp | 3545 bp | 3495 bp |
| Median Fragment Length | 4818 bp | 4789 bp | 3312 bp | 3147 bp | 3541 bp | 3490 bp |
| Gaps | 1371 | 827 | 1617 | 1035 | 2117 | 1407 |
| Average Gap Length | 33.43 bp | 35.96 bp | 27.20 bp | 27.66 bp | 37.23 bp | 35.92 bp |
| Median Gap Length | 14 bp | 14 bp | 10 bp | 10 bp | 14 bp | 14 bp |
| Gaps per Fragment | 0.163 | 0.157 | 0.239 | 0.245 | 0.234 | 0.244 |

Table S3: Frequency of deletions within retrotransposons located in the maize1 and maize2 genomes. Gaps (deletions) were identified from the alignments calculated by LASTZ and exclude gaps <3 bp and all in-frame deletions.

Table S4: Sources of RNA-seq data used in this experiment.

| Dataset | Experiment ID | Uniquely Aligned Reads | Citation |
|---|---|---|---|
| Shoots | SRX012380 | 17325252 (48.23%) | Xiangfeng Wang et al. *Plant Cell* 21, no. 4 (April 1, 2009): 1053-1069. |
| Roots | SRX012381 | 15441390 (44.18%) | Xiangfeng Wang et al. *Plant Cell* 21, no. 4 (April 1, 2009): 1053-1069. |
| Leaf Tip | SRX018904 | 13645312 (70.75%) | Pinghua Li et al. *Nature Genetics* http://dx.doi.org/10.1038/ng.703. |
| Leaf Base | SRX018901 | 12423811 (70.39%) | Pinghua Li et al. *Nature Genetics* http://dx.doi.org/10.1038/ng.703. |
| Mesophyll | SRX018905 | 11905642 (54.66%) | Pinghua Li et al. *Nature Genetics* http://dx.doi.org/10.1038/ng.703. |
| Bundle Sheath | SRX018907 | 122225759 (61.53%) | Pinghua Li et al. *Nature Genetics* http://dx.doi.org/10.1038/ng.703. |
| Shoot Apical Meristem | SRX014791 | 2150900 (20.18%) | Yi Jia et al. *PLoS Genetics* 5, no. 11 (November 20, 2009): e1000737 |
| Developing Ears | GSE24788 | 387628* (47.86%) | Andrea L. Eveland et al. *Plant Physiology* 154, no. 3 (November 1, 2010): 1024-1039. |

*Condensed reads with associated information on total number of copies in sequence data

Table S5: Median differences between expression for gene pairs where both homeologs are expressed

| | Leaf Tip | Leaf Base | Seedling Roots | SAM | Bundle Sheath | Mesophyll | Developing Ears | Shoots |
|---|---|---|---|---|---|---|---|---|
| All Expressed Gene Pairs | 2.13 | 1.80 | 1.98 | 1.87 | 2.57 | 2.59 | 2.76 | 1.94 |
| Maize1 Expressed Higher | 2.31 | 1.84 | 2.08 | 1.94 | 2.77 | 2.73 | 2.79 | 1.99 |
| Maize2 Expressed Higher | 1.89 | 1.72 | 1.90 | 1.72 | 2.30 | 2.53 | 2.68 | 1.85 |

## Chapter 3: Genes identified by visible mutant phenotypes show increased bias towards one of two maize subgenomes

The following chapter (excluding the preface) has been published as a peer reviewed article in PLoS One:

Schnable JC, Freeling M. (2011) "Genes identified by visible mutant phenotypes show increased bias towards one of two maize subgenomes." PLoS One doi: 10.1371/journal.pone.0017855



### Preface:

In the preceding chapter I outlined a model for explaining bias in gene loss. Briefly:

1. Otherwise equivalent duplicate genes created by whole genome duplication are unequally expressed as a result of genome dominance.
2. Short to intermediate sized deletions, now known to be the mechanism of duplicate gene removal following whole genome duplication (Woodhouse *et al.* 2010), remove genes from both subgenomes at equal frequencies.
3. The loss of the more expressed copy of a gene is more likely to have phenotypic consequences and reduce fitness, while the loss of the less expressed gene copy is more likely to be phenotypically silent and selectively neutral.
4. Purifying selection purges deletion alleles of the higher expressed gene copy while deletion alleles of the less expressed gene copy accumulate neutrally.
5. After millions of generations the dominant subgenome retains significantly more genes than the non-dominant subgenome because the highly expressed gene copies are concentrated on the dominant subgenome.

This model perfectly explained the observed data in maize (and other species), but it remained a model. How do you test an evolutionary genomics model? Setting up two polyploid species, one with genome dominance and one without, letting them reproduce naturally for thousands or millions of generations and then resequencing clearly isn't feasible. So I started thinking about other predictions of this model.

One prediction this model makes is that, for duplicate genes still present in the modern maize genome, knocking out the copy from the dominant subgenome should be more likely to result in a phenotypic change which reduces fitness than knockout out the copy from the non-dominant subgenome. Knocking out out hundreds of genes and characterizing them myself wasn't feasible during my time in grad school, but fortunately the maize community had already done the next best thing for me. The records at maizeGDB contained information on every mutant identified by forward mutagenesis or even by fortuitous discovery in the fields of maize geneticists or farmers. I predicted these genes should be found disproportionately on the dominant maize subgenome.

The barrier to taking advantage of the rich data stored on MaizeGDB was that there was no accurate map of which gene models ID assigned by the maize sequencing consortium (GRMZM2G146644 etc) corresponded to characterized genetic loci (*a1*, *kn1*, *rs2* etc). In order to facilitate the research described below I created such a list by manual annotation and it has since been incorporated into the information provided by MaizeGDB. This manual annotation also provided a glimpse into the many inaccuracies of the first version of the maize genome (see Figure 2).

**Introduction**

The grasses, the approximately 10,000 species in the family Poaceae, are one of the most ecologically and economically significant taxa the planet. Comparative mapping of diverse grass species lead to the conclusion that they are all similar in gene content and order [1, 2] to the point that it was argued grasses could be treated as a single genetic system (Bennetzen and Freeling 1993). In other words, knowledge gained from the study of any one grass species could be quickly and directly applied to all other species in the family.

Among the grasses, maize is without question the species with the longest and most comprehensively documented history of genetic investigation. The rich genetic resources found in maize are the result of over a century of genetic investigation beginning with R. A. Emerson's small but distinguished group in the early 20th century; see B. McClintock's unpublished note on this group (McClintock). The resulting set of characterized genes has the potential to be of great value in the genomics era and sets maize apart from many model systems of more recent origin. Until now the applications of this information in a genomic context have been severely limited by the lack of reliable connections between the data produced by geneticists studying individual genes and the datasets produced by genomicists who generally work at the level of whole genomes.

We curated a dataset 464 “classical” maize genes supported by citations from at least three publications, mutant phenotype data, or direct requests from the maize community using data presented in MaizeGDB: The Maize Genetics and Genomics Database (Lawrence *et al.* 2004, 2008). Using manual annotation we connected these well characterized maize loci to gene models created by maizesequence.org, the group that recently published a sequence of the maize genome. To increase the utility of this dataset we also identified orthologous genes at syntenic locations in the genomes of three other grass species with published genomes: rice (Goff *et al.* 2002), sorghum (Paterson *et al.* 2009), and brachypodium (The International Brachypodium Initiative 2010). The evolutionary relationships of these grass species and a number of other notable grasses are shown in Figure 1. This initial classical gene list was distributed to the maize community with links to software that graphically presented our pan-grass synteny data and links to the MaizeGDB locus pages where all data regarding individual maize genes is archive.

The maize lineage, a branch that included both *Zea* and *Tripsacum*, experienced a whole genome duplication an estimated 5-12 million years ago [10-12]. This duplication created two homeologs (syn. homoeologs, ohnologs, syntenic paralogs) co-orthologous to single copy genes in other, unduplicated, grass species. The nearest unduplicated outgroup species with a sequenced genome is *Sorghum bicolor*. For many genes, the two duplicated copies were functionally redundant and one copy or the other has been lost from the genome of modern maize by an intrachromasomal recombination deletion

mechanism (WOODHOUSE *et al.* 2010). Gene deletion has been significantly more common from one of the two duplicate subgenomes of maize: maize2 (SCHNABLE *et al.* 2011b).

Here we show that the genes of interest to maize geneticists are much more likely to be syntenically conserved across all grasses than the average gene supported by full length cDNA evidence. We also found that maize genes identified by a mutant phenotype are disproportionately found on maize1. The bias is true both for genes with a retained duplicate from the whole genome duplication, and singletons whose duplicate copies have been deleted. This finding was predicted by our previously published hypothesis that deletions of duplicate gene copies from the maize1 subgenome are more likely to impact fitness than deletions of copies of the same genes from maize2, as maize1 genes tend to be expressed at higher levels than their duplicates on maize2 (SCHNABLE *et al.* 2011b). We provide all our data on gene locus to gene model mapping, and identification of orthologous genes in other grasses and the homeologous gene in maize, if present, locations in the hopes that these data will be of use to others in the research and teaching community (Supplemental Information S1).

## Results

### Comparing gene models of individually cloned genes to gene models released by the maize genome sequencing consortium

Manual mapping of experimentally validated genes to the maize genome provided a chance to error-check the version_2 gene models released by maizesequence.org. Overall most gene models agreed with previously cloned gene model data (supplemental data S1). Aside from missed UTR exons and the genes which were classified as supported only by *ab initio* prediction despite being supported by sequences in GenBank, the most frequent error we identified were genes that had been split into multiple unlinked gene models by maizesequence.org. This generally resulted from apparent mistakes in the ordering of contigs within BACs. The overall error rate was substantially reduced in the B73_refgen2 release, which increased the percent of contigs with order and orientation information from 30 to 80% (WEI *et al.* 2010). However this form of error remains present in version 2. For example the coding sequence of thegene *aspartate kinase-homoserine dehydrogenase1* is split into three separate gene models (Fig. 2A).
The most dramatic example of an erroneous gene model is provided by *cytokinin oxidase1*, where the 5' and 3' regions of the coding sequence mapped to the same gene model – GRMZM2G146644 – but the gene model included apparently unrelated exons from a contig inserted between the two ends of *cytokinin oxidase1* (Fig. 2B). In an additional two cases – *male sterile45* and *ferritin homolog2* -- the entire CDS of a gene mapped to regions annotated as UTR (Fig. 2C). We provide proofing links in our master classical maize gene list so that a researcher can immediately visualize obvious annotation problems using the GEvo comparative genomics tool (a CoGe application) used to generate Figure 1 (Supplemental Data S1) (LYONS *et al.* 2008a).

**Comparing human to computational identification of maize genes using known sequences**

Subsequent to the February, 2010 release of our initial version of classical maize gene list to the maize genetics community, maizesequence.org released a list of gene models mapped to named loci in the MaizeGDB database using the Xref computational pipeline (http://www.maizesequence.org/info/docs/namedgenes.html). Comparing their machine-annotated dataset to our version 2 list, we identified 152 cases of overlapping assignment of classical maize genes and named maize genes (Supplemental Data S1). The remaining 316 classical maize genes identified by manual annotation were not caught by the computational pipeline. In 140 of the overlapping cases, both lists assigned loci to the same gene model. The remaining 12 cases were further investigated using multiple independent GenBank records, as well as genetic location data record on MaizeGDB locus pages. In two cases the Xref assignment was clearly correct and the appropriate corrections were made to our list. In nine cases sequence and genetic location data supported the manual assignment over that of Xref. No conclusion could be reached in the final case.

**Identification of orthologs of classical maize genes in other grasses**

The current release of the maize genome – B73_refgen2 – contains over 110,000 annotated genes, many of which have already been identified as gene fragments or genes encoding transposon related proteins. To develop a subset of genes comparable to our classical gene list we adopted an approach used previously (Eveland *et al.* 2010) restricting ourselves to the subset of annotated maize genes supported by sequenced full length cDNA evidence (see Methods) (Alexandrov *et al.* 2009; Soderlund *et al.* 2009). In total we identified 34,579 genes supported by full length cDNAs including 81.9% of the unique genes on our classical maize gene list and 75% of the unique genes which were originally identified by a visible mutant phenotype.

Using the online syntenic analysis tool SynMap (Lyons *et al.* 2008b), we found that, compared to the average maize gene supported by full length cDNA evidence, classical maize genes, including those with known mutant phenotypes, are much more likely to possess conserved homologs at orthologous syntenic locations – true orthologs -- in *Japonica* rice, sorghum, and brachypodium (Fig. 3).

**Distribution of classical maize genes and mutant phenotype genes between subgenomes**

The maize genome is comprised of two subgenomes maize1 and maize2 (Schnable *et al.* 2011b). Each subgenome is orthologous to the entire genomes of sorghum, rice, and brachypodium. These other grass genomes have remained unduplicated since the radiation of the grasses. The two subgenomes are distinguished by expression of retained duplicate genes and gene loss rates. Maize1 genes tend to be expressed at higher levels

than their retained homeologs on maize2, and maize2 has lost copies of more genes syntenically retained in other grass species than maize1 (Schnable *et al.* 2011b).

The distribution of syntenically retained classical maize genes between the two subgenomes of maize roughly mirrors that of all syntenically retained genes supported by full length cDNA evidence. Figure 4 plots these data for all 34,579 genes supported by full length cDNA evidence, the 468 genes of the classical gene list, and the subset of 102 genes on the classical gene list identified by mutant phenotype prior to cloning. Given the bias towards greater expression of maize1 homeologs, the slight bias towards higher numbers of maize1 genes with retained homeologs among genes supported by full length cDNA evidence was expected, but this finding is not of significant interest. However, among syntenically retained genes which were first identified by a visible mutant phenotype, the bias towards the maize1 subgenome is significantly greater than for the classical maize gene list as a whole (p=.028, Fisher Exact Test), and members of homeologous gene pairs located on maize1 were twice as likely as the duplicate copies on maize2 to be originally identified by mutant phenotype -- 29 maize1 genes with homeologs vs. 14 maize2 genes with homeologs (significantly different from a 50/50 split p=.0222, Chi-square test).

## Discussion

### The benefits of manual gene annotation

Our manual proofing of the classical maize gene list shows that, as tempting as it may be to rely primarily on inexpensive *in silico* annotation techniques, manual structural annotation provided a significant amount of important information to B73_refgen2. Tools are available that allow interested researchers to proof and improve the structural annotations of their favorite genes (Wilkerson *et al.* 2006). Having those improvements incorporated into official genome annotations would benefit the entire community.

### Syntenic conservation of classical maize genes

The idea that genetic colinearity among the grasses could be used to accelerate the research across the whole family is a venerable one [1, 2, 22]. Enthusiasm for this concept waned as the sequencing of multiple grass genomes demonstrated that a significant fraction of transcribed genes are not syntenically retained across species. Our finding that 37% of maize genes supported by full-length cDNA are not retained at a syntenic position in other grass species, and almost 50% of these genes apparently inserted into their present locations prior to divergence of the BEP clade, represented by both rice brachypodium is in agreement with previous studies. Research in the arabidopsis, using papaya as an outgroup, estimated that half of all annotated genes in that species belonged to a “gray” genome of genes which had transposed into nonsyntenic positions within the last 70 million years (Freeling *et al.* 2008). A recent study in Drosophila found that knockouts of recently inserted – with the last 35 million

years – and ancient syntenically conserved genes produced lethal phenotypes at statistically similar rates (Chen *et al.* 2010).

Genes belonging to the gray genome of maize are essentially unexplored. The genes of greatest interest historically seem to be precisely those that are retained in the same syntenic position in the genomes of all grass species. It may be that, in plants, genes essential for day to day function, such as those involved in key biochemical and developmental pathways, are by definition less likely to transpose or, when they transpose, are less likely to rise to fixation within a species. A small but significant number of mutant genes in maize were identified using map-based cloning approaches relying on rice synteny, prior to the publication of the maize genome. While map-based cloning and comparison of maize to rice certainly did occur, we think it unlikely that this explanation accounts for the magnitude of our results.

The techniques used in this paper allowed us to identif,y with high confidence, lost or transposed genes by first identifying a predicted orthologous syntenic location in the target grass genome. Even the genes which are not retained in all species can be a starting point for hypothesis driven research, a use we support via Gevo links to enable quick visual comparisons of orthologs or predicted locations in multiple grass species (Supplemental Information S1). For example, c1 and pl1 are two homeologous maize genes that regulate the biosynthesis of anthocyanin. Both genes have been studied extensively by the maize genetics community. A syntenic co-ortholog of the two genes is retained in the genomes of both sorghum and rice. However the gene is absent from orthologous region of the brachypodium genome (Fig. S1) which prompted us to investigate further and find the gene was not present anywhere in the brachypodium genome (Fig. S2). We conclude from this brief research foray that this portion of the anthocyanin biosynthetic regulatory pathway may be significantly different or completely absent in brachypodium, opening avenues for further research.

**Increased bias of towards the maize1 subgenome of mutant phenotype genes**

A bias towards maize1 for the classical maize genes was expected given the greater total number of retained genes present in that subgenome. However, when we examined the subset of the classical maize gene list identified by a mutant phenotype prior to cloning, the bias of this dataset towards the dominant subgenome – maize1 – was significantly greater than could be explained by the difference in total gene numbers between the two subgenomes. Interestingly this bias is also statistically significant for genes with a retained homeolog on the opposite, homologous subgenome, maize2. Since there is one gene copy present in each subgenome for this class of gene, *a priori* evidence of gene function, the expectation was that mutations of either copy would be about equally likely to produce a mutant phenotype. This was not the case.

Rather, our finding that maize1 is the preferred location of genes with mutant phenotypes

even when a homeologous duplicate is present suggests that the loss of maize1 copies may be more likely to result in visible impacts of the sort which might catch the eye of researchers, or farmers, in the field. As impacts on plant morphology visible to researchers are likely to have a pronounced impact on plant fitness, this finding is certainly consistent with our previously published hypothesis that the deletion a gene from maize1 is more likely to be selected against than the deletion of the same gene from maize2 (SCHNABLE *et al.* 2011b).

The corollary is even more interesting: knockout phenotypes to not appear to be behaving as if gene function was buffered by a duplicate copy of the same gene expressed in the same cells. For the moment, our working hypothesis is that maize1 gene copies have predominantly retained the ancestral function of the gene in the pre-duplication ancestor of maize, leaving maize2 copies free to potentially adopt new, or less essential functions. This prediction is fully testable on a gene-by-gene basis through investigation of the function of orthologous genes we identify in the closely related and unduplicated species sorghum.

**Conclusion**

This pilot study demonstrates the usefulness of traditional genetics data in the genomics era, and the importance of model species like maize with long histories of genetic investigation. A large number of morphological mutants in maize remain uncloned. The ability to identify high confidence orthologs in all grass species with sequenced genomes combined with the unrivaled economic and ecological significance of the Poaceae means investigation of a gene or gene family in any of one these species can quickly benefit researchers working around the world to answer a wide range of questions in different grass species. We hope that the tools, datasets, and links provided here (Supplementary Information S1), as well as our preliminary findings, will support continued insights based on pan-grass comparative genetics.

**Materials and Methods**

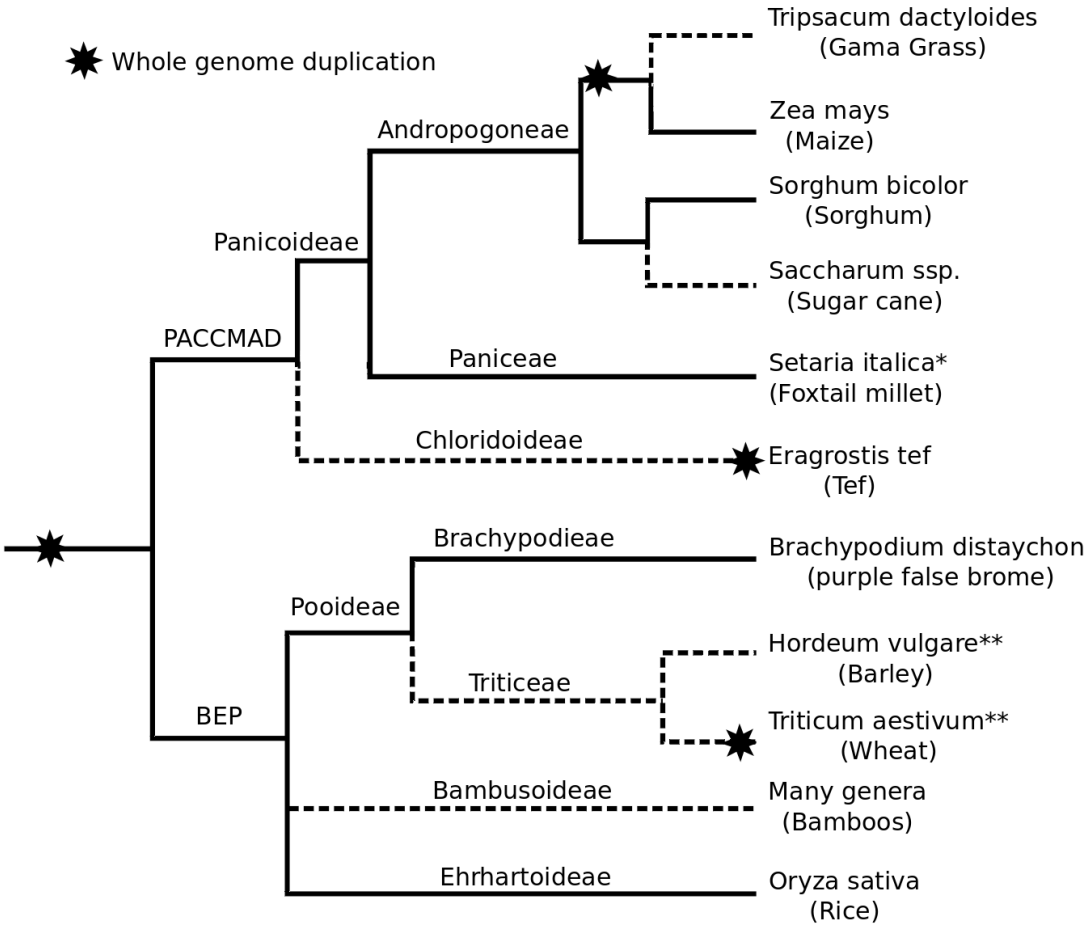
Classical maize genes were identified from the list of maize loci maintained by MaizeGDB (LAWRENCE *et al.* 2004, 2008) and include genes with associated GenBank sequence records with greater than three referencing papers in the database, additional cloned genes with known mutant phenotypes, as well as genes added after soliciting community input. Genes were initially mapped to the sequenced maize genome using LASTZ, and then visually proofed and corrected using GEvo part of the CoGe comparative genomics platform (http://genomevolution.org/CoGe/) (LYONS *et al.* 2008a). These GEvo links are provided to aid continued research and permit proofing and verification of our results.

The full length cDNA supported gene set was constructed using the 'semi-strict assembly' collection of full length cDNAs provided by the maize cDNA project

(http://www.maizecdna.org) (SODERLUND *et al.* 2009). Full-length cDNAs were aligned to B73_refgen2 gene models using LASTZ, and those models supported by a full length cDNA with > 95% identity and > 90% coverage were included in the set.

Homeologous genes in maizes and orthologous genes in other grasses were identified using SynMap (LYONS *et al.* 2008b) with the optional Quota Align filters (TANG *et al.* 2011); SynMap is a web based tool available at http://www.genomevolution.org/CoGe/SynMap.pl. When no syntenic gene was identified, a predicted location was generated based on syntenically conserved flanker genes. Predicted orthologous locations longer than 1 MB were excluded as were predicted homeologous locations in maize longer than 2 MB. Our classical maize gene list provides a GEvo link that permits quick visual comparisons among grass orthologs and the predicted locations of deleted grass genes.

**Figures**
Figure 1:




Branch lengths not to scale. *The genome sequencing of foxtail millet by the joint genome institute is complete, but has not yet been published. Therefore it is not included in our analyses (SI 1). **Projects to sequence the genomes of barley and wheat are announced or in progress.

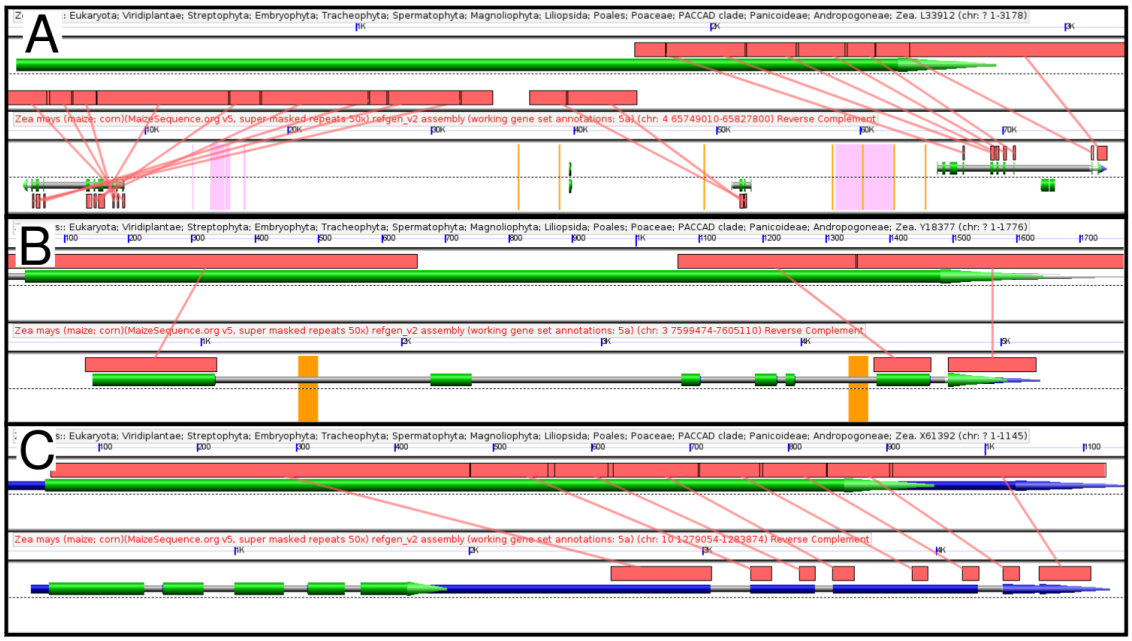



Figure 2: Examples of manually identified errors in maize gene annotations

Graphics from GEvo comparative sequence alignment tool. Annotated cDNAs from GenBank are compared to regions of the maize B73_refgen2 genome. Features on the forward strand are displayed above the dotted line, and features on the reserve strand are displayed below the line. Grey lines mark the extent of gene models with CDS sequences in green and UTR sequences in blue. Orange bars mark the gaps between assembled contigs of the maize genome (stretches of N's). Red boxes connected by lines show sequences identified as homologous by blastn. A. A comparison of the coding sequence of *aspartate kinase-homoserine dehydrogenase1* to the region of maize chromosome 4 that contains the three gene models –from left to right, GRMZM2G365423, GRMZM2G389303, and GRMZM2G437977 -- among which the exons of this gene have been divided. An interactive version of this graphic can be regenerated in GEvo using the following link: http://genomevolution.org/r/25xh B. A comparison of *cytokinin oxidase1* to GRMZM2G146644, a gene model which includes the 5' and 3' ends of *cko1* but has also incorporated unrelated exons from another maize genome contig. Regenerate analysis: http://genomevolution.org/r/25s5 C. The coding sequence of ferredoxin homeolog2 which maps to a region of the maize genome annotated as the 3' UTR of GRMZM2G147266. Regenerate analysis: http://genomevolution.org/r/25s7

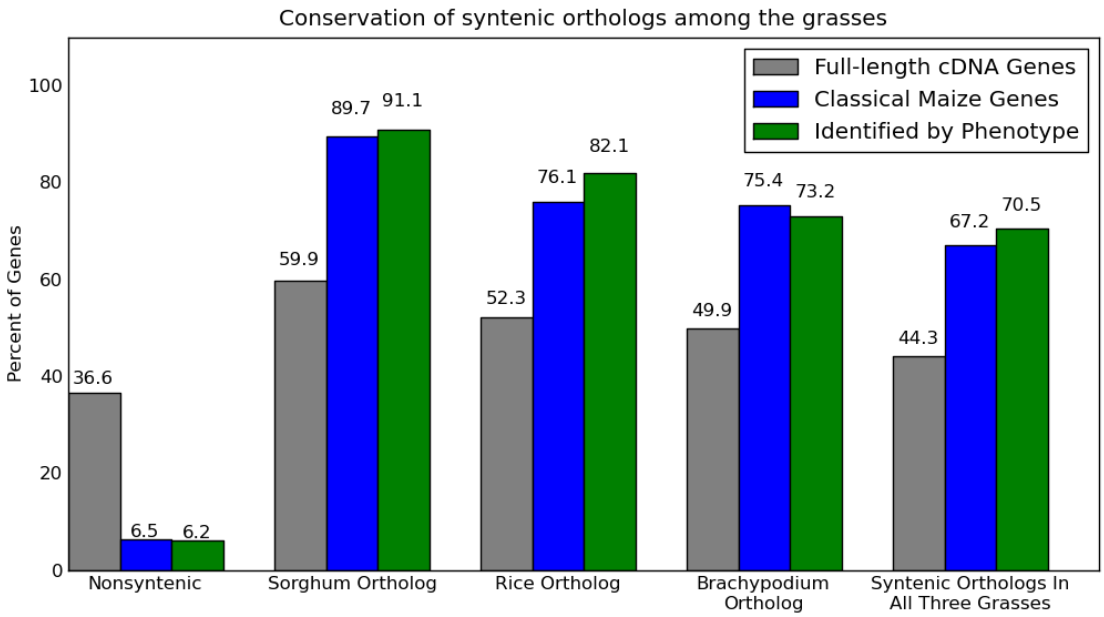



Figure 3:

Comparison of the proportion of genes identified by a mutant phenotype prior to cloning (N=111) , all classical maize genes (N=464), and all maize genes supported by full length cDNA evidence (N = 34579) for which syntenic orthologs could be identified in the other three grass species with sequenced genomes: sorghum, rice, and brachypodium.

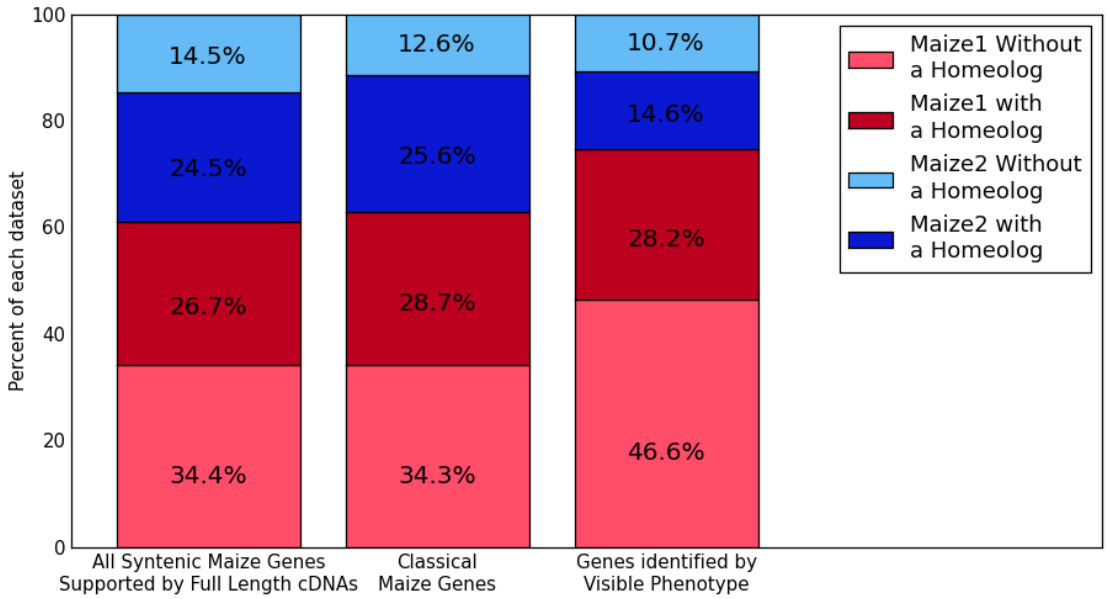



Figure 4
Comparison of the distribution of genes retained syntenically in at least one other grass species between the two subgenomes of maize as well as whether genes possess retained homeologs from the maize whole genome duplication. For syntenically retained maize genes with full length cDNA support N = 17956. For the subset of the classical maize gene list that are syntenically retained N=429. For the subset of genes that were first identified by mutant phenotype and are syntenically retained N = 102.

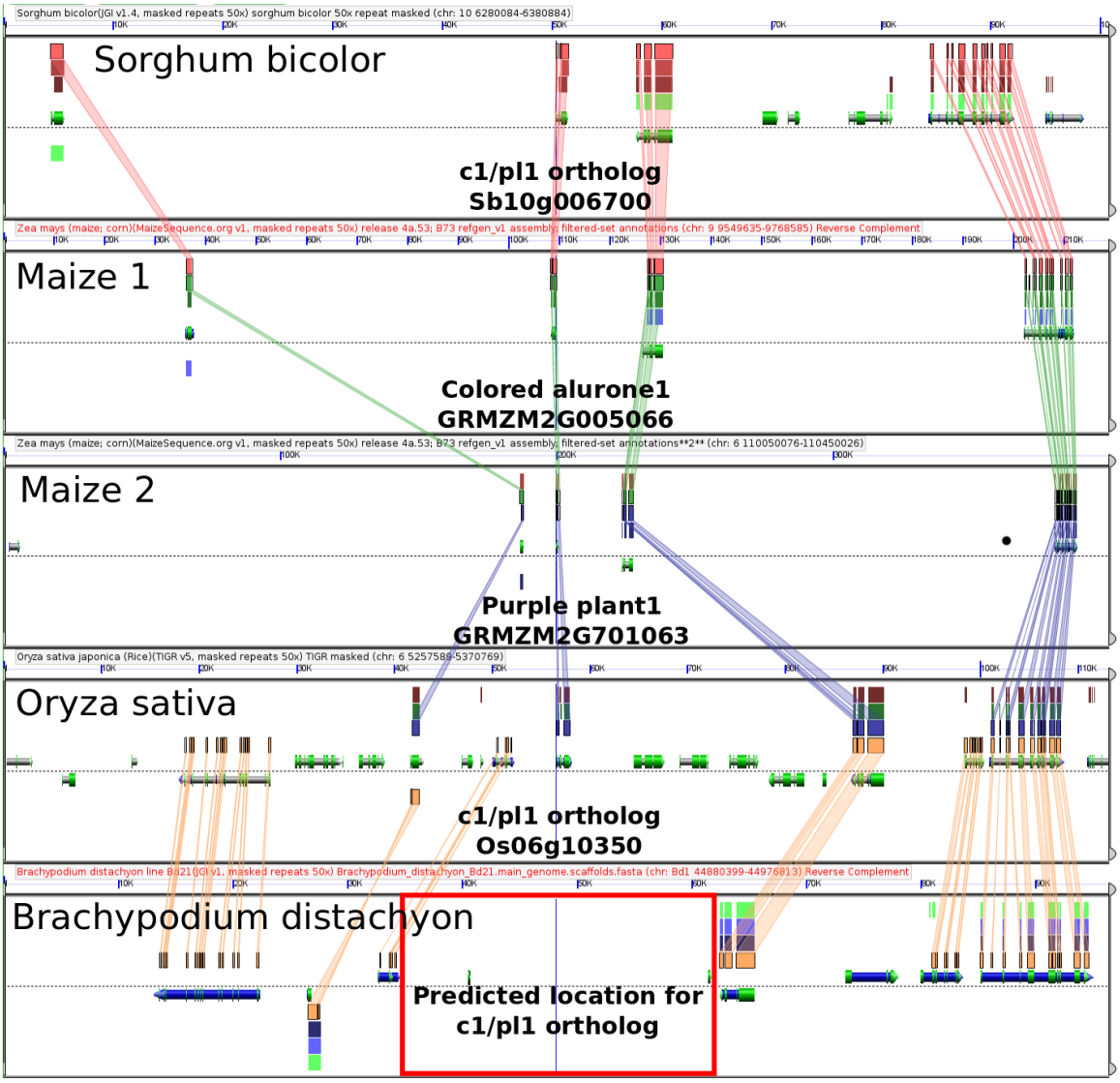



Figure S1 An example of a well studied classical maize gene which has been deleted from the genome of brachypodium.

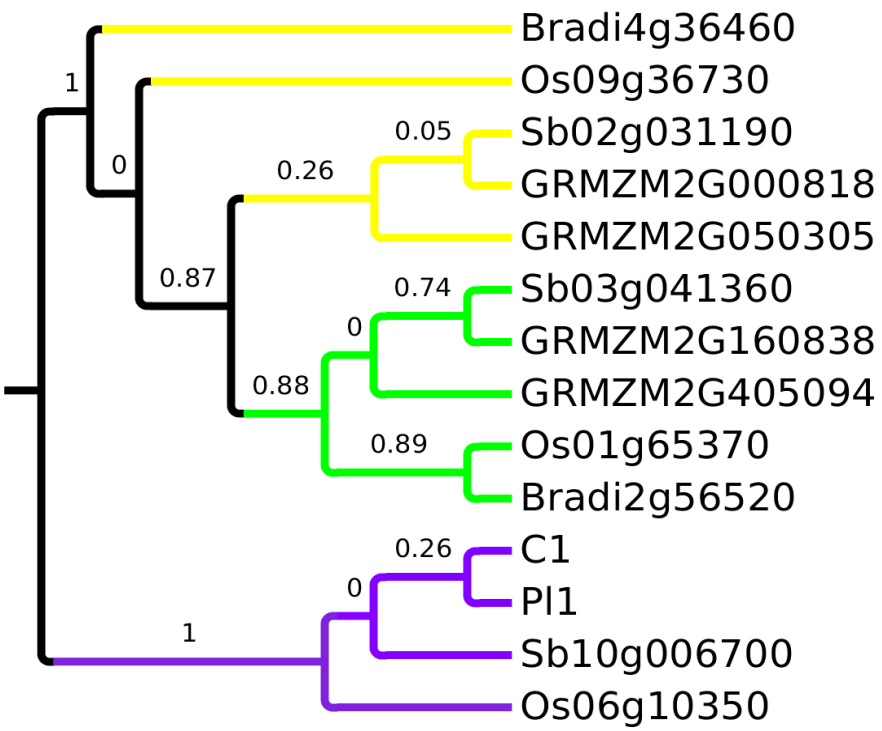



Figure S2 A phylogenetic tree confirming genes showing sequence similarity to C1 and Pl1 belong to other clades of genes and are not orthologs.

## Chapter 4: Genome-wide analysis of syntenic gene deletion in the grasses

The following chapter (excluding the preface) has been published as a peer reviewed article in the journal Genome Biology and Evolution:

Schnable JC, Freeling M, Lyons E. (2012) "Genome-wide analysis of syntenic gene deletion in the grasses" Genome Biology and Evolution doi:10.1093/gbe/evs009

Contributions:

Eric Lyons assisted in the direction of research and design of experiments. Experiments and analysis were all carried out by myself.



**Preface:**

A core requirement of the analyses presented in every chapter of this thesis is that it is possible to accurately identify orthologous genes in different species. Different research groups use different methods to identify orthologs. These methods can be largely grouped into methods based on the sequence of individual genes, and methods based on the order of homologous genes in different species. Both methods have problems. Sequence based methods can be thrown off by genes experiencing high rates of positive selection or orthologs which have become pseudogenes and are accumulating large numbers of nonsynoymous and frameshift mutations. Synteny based methods must distinguish between syntenic regions produced by orthology and homeology. The latter is a particular problem in plant genomes where whole genome duplications and the resulting syntenic homeologous regions are common.

In this paper I explained the details of the method I employed to combine the two methods. By looking at the aggregate synonymous substitution rates across all the genes in a syntenic block it was possible to identify which blocks diverged during speciation and which blocks diverged in a previous or subsequent whole genome duplication. Using the orthologous genomic location predicted by syntenic blocks, it becomes possible to zero in on a specific location in the genome to search in detail for an orthologous gene, unannotated homologous sequence (indicating either a gene missed by genome annotated or a gene fragment/pseudogene), a gap in the genome assembly indicating the orthologous gene may have been missed, or a high confidence deletion of an orthologous gene.

Within this paper these data were used to make comparisons of the rates and types of

gene loss following the pre-grass whole genome duplication in three grass species. The same methodology was used to produce the orthologous and homeologous gene sets used in all analyses presented in other chapters (unless otherwise noted).

**Introduction:**

Evidence of ancient polyploidies, or WGDs, are found throughout all eukaryotic lineages (Dehal and Boore 2005). These duplications, whether auto- or allopolyploidy events, instantly create copies of all genes and associated regulatory sequences contained within the nuclear genome of a species. Interestingly, ancient whole genome duplications tend to be associated with adaptive radiations of multiple lineages, although the causality of this relationship remains controversial (Soltis *et al.* 2009).

Multiple explanations for the association between whole genome duplication and species radiations have been proposed. At a mechanistic level, the reciprocal loss of duplicated genes from one of the multiple subgenomes of a polyploid organism, a process known as fractionation -- or in older literature as "diploidization" -- could increase the speed with which hybrid incompatibly develops between populations (Lynch and Force 2000). It has also been suggested that ancient whole genome duplications tend to be contemporaneous with major extinction events (Van de Peer *et al.* 2009); polyploid species that survived such events would be expected to radiate into the abundant newly vacated niches left by the wave of extinctions. Finally, it may be that whole genome duplications, by creating a new source of redundant genes suitable for co-option for novel functions or subfunctionalization, increase the potential for niche specialization (De Bodt *et al.* 2005) or morphological innovation (Freeling and Thomas 2006).

Following a whole genome duplication, redundant copies of many genes are removed from the genome by fractionation (Langham *et al.* 2004). A study of synthetic Bassica allotretraploids have reported major genomic rearrangements and deletions within as few as five generations (Osborn *et al.* 2003). In addition, duplicate gene deletion continues at significant levels in maize 5-12 million years after polyploidy (Woodhouse *et al.* 2010; Swanson-Wagner *et al.* 2010; Schnable *et al.* 2011b); only forty-seven percent of maize-sorghum syntenic genes are still represented by genes or gene fragments at both duplicate locations within the maize genome (Woodhouse *et al.* 2010). A study in yeast documented ongoing loss of duplicate gene copies throughout the entire period since whole genome duplication in that lineage (Scannell *et al.* 2006).

The loss of duplicate genes is biased in multiple ways. The first is biased retention of both duplicate copies of certain classes of genes following WGD. These classes include genes encoding members of large multi-protein complexes, transcription factors (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Scannell *et al.* 2006; Freeling *et al.* 2007), and genes associated with large numbers of conserved noncoding regulatory elements (Schnable *et al.* 2011a). The loss of genes is also biased between duplicated regions. After first being observed in arabidopsis and maize (Thomas *et al.* 2006; Woodhouse *et al.* 2010), this bias was found to be a general property of eukaryotic whole genome duplications (Sankoff *et al.* 2010). The bias in gene loss is a property of whole parental

genomes in both maize and *Arabidopsis suecica* (Chang *et al.* 2010; Schnable *et al.* 2011b) and therefore may represent a useful mark for reconstructing ancestral subgenomes across organisms with ancient polyploidy.

All grass species sequenced to date share an ancient whole genome duplication tentatively dated to approximately 70 million years ago (Paterson *et al.* 2004; Yu *et al.* 2005) contemporaneous with the emergence of phytoliths representing extant grass families in the fossil record (Prasad *et al.* 2005). Of all plant families, the grasses are represented by the most published sequenced genomes – brachypodium, maize, rice and sorghum -- representing three subfamily-level grass lineages (Goff *et al.* 2002; Paterson *et al.* 2009; Schnable *et al.* 2009; The International Brachypodium Initiative 2010). It is likely the grasses will retain this distinction for the foreseeable future with genome assemblies for additional grass species currently available under pre-publication restrictions, in the process of being sequenced, or in the planning stages of being sequenced. Given the economic and ecological significance of the grasses and the demand for fast porting of functional information among grass species, there is a need for automated, yet accurate, tools to identify and classify orthologs and homeologs in many-to-many genomic comparisons. However, a number of known genomic events complicate the assignment of orthologous genes between grass species.

In addition to the previously mentioned ancient WGD shared by all grasses, a relatively recent whole genome duplication is found within maize, dated to 5-12 million years ago, just subsequent to the divergence of this lineage from the common ancestor of sorghum and the core of its tribe (Swigoňová *et al.* 2004). As a result, there are two homeologous locations within the maize genome co-orthologous to any single location in the genomes of rice, sorghum, and brachypodium (Figure 1). The genomic relationships created by the pre-grass whole genome duplication and the second duplication in the maize lineage are summarized in Figure 1. The size of the maize genome is also more than twice the next largest sequenced grass, largely as a result of multiple waves of transposon amplification in the last several million years (Baucom *et al.* 2009; Schnable *et al.* 2009). Syntenic analysis of the grasses has also detected evidence of more ancient whole genome duplication events shared by most, if not all, monocot species (Tang *et al.* 2010). These more ancient duplicated blocks must be identified and removed from genomic comparisons aimed at identifying duplicates from the more recent tetraploidy shared by all grasses. Finally, duplicated regions in all grasses – located on chromosomes 11 and 12 of rice, and chromosomes 5 and 8 of sorghum – have a peculiar evolutionary history and have evolved in concert since the pre-grass tetraploidy (Wang *et al.* 2011). These highly similar duplicate regions pose significant issues for some methods of automated ortholog/homeolog classification based on average sequence similarity or evolutionary distance.

==========

======
Definitions Text Box:

**Genomes and Genomic Regions:**
Whole genome duplication: Abbr. WGD. The duplication of an entire genome. WGDs generate polyploid organisms. May be subclassified as auto- or allo- denoting a single parental genome or multiple parental genome origin respectively.
Diploid: Denotes that a genome has two homologous copies of each chromosome.
Polyploid: Denotes that a genome has more than two homologous copies of each chromosomes.
Subgenome: The constituent genomes within a polyploid species, each of which is derived from the entire genome of a parent or ancestral species and prior to fractionation, contained all the genes found throughout the clade within which the polyploid species falls.
Syntenic region: Two or more homologous genomic regions descended from a common ancestral genomic region. Syntenic regions are evidence by homologous genes arranged in a collinear order.
Fractionation: The loss of one or the other duplicated gene copy following a whole genome duplication. (near synonym: diploidization)
Fractionation bias: The uneven distribution of gene deletions between duplicated genomic regions following WGD.
Underfractionated: The copy of a duplicate chromosomal region from which fewer genes were lost.
Overfractionated: The copy of a duplicate chromosomal region from which more genes were lost.

**Evolutionary Relationships and Types:**
Homolog: Of common ancestry. Homologous genes and genomic regions are derived from a common ancestral gene or genomic region.
Orthologs: Homologous genes or genomic regions derived from the divergence of lineages.
Paralog: Homologous genes or genomic regions derived from their duplication within a lineage.
Homeologs: The subset of paralogs created by whole genome duplication. (synonyms: ohnolog; syntenic paralog)
Pan-grass gene: A gene present in the ancestral pre-duplicated genome of the grasses remaining at its ancestral position. Pan-grass genes are detected though comparison of syntenic region within and among grass genomes.
Ancestral gene: A gene hypothesized to be present in the ancestral genome at its current extant location. Ancestral genes are defined by their conserved genomic position in multiple lineages or subgenomes.
==========

======

Many previously published methodologies for ortholog identification use some variation of best BLAST hit . In order to identify whole genome duplication events, the evolutionary distances of homologous gene pairs are often calculated using synonymous mutation or 4DTV values, and the histogram of values interrogated for distinct peaks (Tuskan *et al.* 2006; Barker *et al.* 2008)
A number of tools do incorporate identification of syntenic blocks as discussed: (Soderlund *et al.* 2011). In comparisons between multiple flowering plant species, all with extensive histories of whole genome duplication, it is necessary to distinguish between more recent and more ancient syntenic blocks (Tang *et al.* 2011).

In this paper we demonstrate a method for ortholog/homeolog classification based on the identification of syntenic blocks of genes in inter- and intra-species genomic comparisons followed by the calculation of aggregate divergence data for all gene pairs within the block. Our method permits the subsequent identification of high confidence gene loss/transposition events which are crucial for the study of genome evolution following polyploidy. In addition, this method permits the identification of two subgenomes shared by all sequenced grasses--a low gene loss under-fractionated subgenome (GrassA) and a high gene loss over-fractionated subgenome (GrassB)-- as previously demonstrated for the much younger maize tetraploidy (Schnable *et al.* 2011b). We use this method to identify orthologs and homeologs between four grass species with published genome sequences and reconstruct the ancestral subgenomes comprising each grass' modern genome. We assign gene loss events to nodes on the grass phylogenetic tree and search for reciprocally lost duplicated genes which might have contributed to reproductive isolation during the radiation of the major grass lineages.

**Methods:**

**Generating Lists of Syntenic Orthologs/Homeologs**
Lists of syntenic gene pairs were initially generated for all pairwise comparisons -- including self-self comparisons -- using the SynMap utility of CoGe (Lyons *et al.* 2008b) with the parameters described in supplementary table S3 of this paper. Individual stretches of syntenic genes were merged into larger syntenic blocks using the method described in (Yang 2007).

Synonymous substitution rates between individual syntenic gene pairs were calculated within the SynMap utility for aligned coding sequences of gene pairs guided by the alignment of the the translated coding sequences of gene pairs by nwalign (http://pypi.python.org/pypi/nwalign/). Synonymous substitution rates for these aligned sequences were calculated by a customized version of CODEML (Alexandrov *et al.* 2009).

Syntenic blocks containing 12 or more gene pairs were assigned to an evolutionary event, whether speciation (orthologous) or whole genome duplication (homeologous), based on a unified synonymous substitution rate (Ks) for genes contained within the block. This unified synonymous substitution rate is defined as the average synonymous substitution rate among gene pairs contained within the syntenic block after discarding the most diverged two thirds of genes contained within the syntenic block. The calculation of synonymous substitution rates is very sensitive to errors in gene model annotation or sequence alignment and examining only the lowest one-third of Ks values provides sufficient data-set to differentiate sequence blocks while eliminating any distortion from the very high substitution rates calculated between incorrectly aligned coding sequences. Grass genomes also include a class of high 3rd base pair position GC content genes which generate unreliable synonymous substitution rate calculations (Alexandrov *et al.* 2009).

These calculations produced two fully distinct peaks for synonymous substitution rates of syntenic gene blocks for interspecies comparisons: one corresponding to orthologous syntenic blocks created by speciation, and the other to homeologous syntenic blocks resulting from the pre-grass tetraploidy. Intraspecies comparisons identified a single fully distinct peak of homeologous syntenic blocks resulting from the the pre-grass duplication in sorghum, rice, and Brachypodium, and the more recent maize lineage specific tetraploidy within maize (Figure S4).

**Joining pairs into orthologous blocks and identifying lost orthologs**
Homeologous and orthologous pairs of genes defined by inter and intra species comparison were merged using in-house python scripts to produce lists of pan-grass syntenic genes. When no ortholog of a syntenic group of genes was identified in a species, a predicted orthologous location was identified using the first orthologously conserved genes within that genome up and downstream of the missing gene. If these conversed genes were separated by more than 1 MB or were located on different chromosomes the group of genes was considered to have no syntenic coverage in the missing species.

When a predicted orthologous region was identified, a three step process was used to confirm the absence of a syntenic ortholog. First, all annotated genes within the predicted orthologous region were compared using LASTZ (Harris 2007) to all members of the group of syntenically conserved genes in other species. Any gene with sequence similarity to the existing group of conserved syntenic genes was considered a conserved ortholog and added to the syntenic group. If no gene within the predicted region was hit, the sequence of the entire predicted region was extracted and compared to the existing group of conserved syntenic genes using LASTZ with default settings. Any hit with a score of 3000 or greater within the region was considered an unannotated conserved gene

or gene fragment. Gaps with no syntenic matches to either annotated genes or unannotated sequences were further subdivided between those where a gap of 50 or more Ns were present at the predicted location and those were there were no annotated gaps within the predicted location.

If the same gene was included in multiple syntenic groupings, the group with fewer identified orthologous and homeologous genes was removed from our comparison. Syntenic groups were three or more genes not classified as local duplicates of each other were all identified as orthologs within the same species were also removed from the dataset. These predominately consisted of sequences that were treated as separate genes in some species but merged into single gene in others.

Putative homeologous gene pairs identified only in a single species where neither copy of the gene was sorted with evidence of syntenic orthologs in any other grass species were omitted from our analysis.

Local duplicate genes were defined as a series of homologous genes interrupted by now less than twenty intervening genes (forty genes in maize, given the greater gene density of the maize working gene set). Homology was defined using the same parameters used by SynMap.

**Assignment of regions as over/under-fractionated**

17 pairs of large contiguous homeologous regions were manually defined using a rice-rice syntenic dotplot. Regions with distinct elbows, as seen in the comparison of rice chromosomes 8 and 9 (Fig. 5), were split into multiple segments. For each homeologous pair of regions, the number of pan grass syntenic genes present in one region without any evidence of conserved homeologs in the other was extracted. In three pairs of regions, including the recombination prone end of rice chromosomes 11 and 12 the difference in pan-grass homeologs retained at syntenic locations was less than 10%. These regions were excluded from further analyses. In the remained 14 cases it was possible to assign one region to the over fractionated pan-grass subgenome and the other to the under fractionated pan-grass subgenome. As the mechanism of fractionation has previously been shown to be almost entirely single gene deletions (Woodhouse *et al.* 2010), p-values were calculated using a binomial approach with a null hypothesis that gene deletion was equally likely in both homeologous regions.

**Region loss methods**

The sorghum genome was scanned for cases where forty or more sequential genes lacked identified sytenic orthologs from the same maize subgenome. Cases where overlapping gaps in the coverage of both maize subgenomes were discarded as these likely represent regions where sorghum specific insertions and rearrangements have made it impossible to detect synteny. The remaining 16 apparent deletions were classified based on the average

number of maize genes each sorghum gene within the region served as the best hit for. Based on 1000 permutations of random sets of sorghum genes, we determined a median region averages 1.996 best hits of maize genes to each sorghum gene within the region, with 95% confidence bounds between 1.175 – 4.025 best hits of maize genes to each sorghum gene in the region (Figure S9). Three putative deletions fell outside of this confidence interval and were manually investigated using the CoGe toolkit.

**Visualizing fractionation bias**

In Figure 5 and Figure S8, biased gene content between duplicate regions is computed using the number of pan-grass genes located between neighboring homeologous gene pairs in two syntenic regions. The number of intervening gene pairs is averaged across a sliding window of thirty homeologous gene pairs. Homeologous pairs separated by >= eight pan-grass genes are omitted from this analysis as previous work has shown that these likely represent small translocations (Woodhouse *et al.* 2010).

**Results:**

**Identification of syntenic gene sets and lost genes**

Syntenic gene sets were generated using SynMap (Lyons *et al.* 2008b), and both inter- and intra- species comparisons between all sequenced grasses (see methods for details). Our primary dataset consisted of 16,923 orthologous gene groups where genes or predicted locations could be identified in the three grass species which have remained diploid since the radiation of the major grass lineages (Supplemental dataset S1). Figure 2 shows orthologous regions of the sorghum and brachypodium genomes and homeologous regions of the sorghum, rice, and brachypodium genomes aligned to the twelve chromosomes of the modern rice genome. Similar displays are possible using either the sorghum or brachypodium genomes as reference genomes (Figures S1,S2).

Our dataset included a significant number of predicted locations for orthologous genes where no orthologous gene was identified. These “missing” data points could be divided into three categories.

1) Recent pseudogenes or missed gene annotations: Cases where no annotated gene model matched the genes conserved in other grass species, but sequence homologous to the genes found in other species was identified at the predicted orthologous location (Figure 3A).
2) Gaps in sequence: Cases where no sequence similar to the missing gene was identified, but the predicted orthologous location included a gap in the pseudomolecule assembly, raising the possibility that the region containing the missing gene was not sequenced or assembled (Figure 3B).
3) True deletions: Cases where no gaps or unannotated homologous sequence were present (Figure 3C).

The higher frequency of gaps in the sorghum genome assembly meant only a small number of high confidence gene loss events were identified in this lineage (Figure 4A). However, because sorghum diverged before the split of the rice and brachypodium lineages, it is possible these genes were inserted into their present location since the rice/brachypodium lineage diverged from sorghum. Brachypodium contains more of both class #1 missing genes -- no gene model, but syntenic homologs sequence – ($p < .0001$, chi-square test, df=2) and class #3 missing genes -- high confidence gene losses – ($p < .0001$, chi-square test, df=2) (Figure 4A).

The rice orthologs of deleted brachypodium genes are not significantly enriched in any of the rice GOSlim annotations (Ouyang *et al.* 2007) relative to other syntenically conserved rice genes (Table S1). Lost genes were compared to the population of syntenically retained genes rather than all rice genes because we found that genes retained syntenically in rice and at least one other species were enriched in 72 of 94 terms in the rice GOSlim vocabulary. The mobile fraction of grass transcriptomes is largely uncharacterized (Schnable and Freeling 2011), and in rice this is reflected by the fact that 64% of rice genes with syntenic orthologs in other species have at least one GO-slim annotation while only 24% of nonsyntenic rice genes do.

**No evidence of segmental deletions in maize**

To identify large post-tetraploidy deletions from maize segments, the subgenomes of maize (i.e. maize1 and maize2) were aligned to the orthologous regions of the sorghum genome (Figure S6). After discarded sorghum regions absent from both maize subgenomes -- presumably representing clusters of gene insertions into sorghum or regions without sufficient conservation of synteny to identify orthology -- sixteen regions of forty or more sorghum genes were identified that were orthologous to a only one syntenic region in maize (Table S5).

To test whether these sixteen regions were indeed single copy – as opposed to one syntenic region simply not being detectable using our approach -- all annotated genes in maize were compared to sorghum genes within the candidate regions (Figure S9). For regions that were, in fact, deleted from the maize genome, sorghum genes within the region should be the “best” match to fewer maize genes, while regions without detectable synteny, as a result of rearrangement or misassembly, should show no difference in this metric. The average sorghum gene was found to be the best BLAST hit of 1.996 maize genes from the B73_refgen2 working gene set. Using random permutations of sorghum genes, it was determined that in intervals of at least 40 genes the average number of best BLAST hits from maize genes per sorghum gene was between 1.175 and 4.025 genes 95% of the time (see Methods). Thirteen of the sixteen sorghum regions with putative segmental maize deletions were within these bounds; the remaining three regions had an average number of maize best blast hits below the lower bound of this confidence interval (Table S5). These three regions were manually checked (Figure S7). Two were found to

have an additional syntenic region that was missed by computational approaches (Table S5), leaving only one potential segmental deletion spanning 56 genes in sorghum. Of these genes, 32 had syntenic matches in rice for which GOSlim annotations were available (Table S6). While there was no significant enrichment in GOSlim annotations, we note that only the most extreme enrichments will be significant with such small datasets.

Relative to the other grasses, maize has experienced a much higher rate of gene loss. This is expected given that maize underwent a second, more recent, paleopolyploidy and is experiencing ongoing fractionation of duplicate gene pairs (Woodhouse *et al.* 2010; Schnable *et al.* 2011b). Given that current assemblies of the maize genome exhibits high levels of presence absence variation in gene content (Springer *et al.* 2009; Swanson-Wagner *et al.* 2010) and current versions of the maize genome omit at least 300 genes found in the reference inbred B73 (Lai *et al.* 2010), we omitted maize from our subsequent analyses of gene loss following the pre-grass WGD.

**Fractionation bias between homeologs and subgenome reconstruction**

Given the recent reports that biased fractionation was a property of whole genomes in maize and *Arabidopsis suecica*, it might be possible to use fractionation bias as a marker to reconstruct ancestral genomes in ancient polyploid species such as the grasses. However, the lack of a suitable outgroup for the grasses creates new issues for quantifying fractionation bias. Between one and three quarters of the genes in arabidopsis have transposed to new locations since the divergence of the arabidopsis and papaya lineages ~70 million years ago (Freeling *et al.* 2008). As the pre-grass tetraploidy is estimated to also be approximately the same age (Paterson *et al.* 2004), any study of fractionation bias must first account for the mobile portion of grass genomes.

To compensate for recently inserted genes, we considered only genes orthologously conserved in sorghum and either rice or brachypodium to represent fractionated genes conserved in their ancestral locations. As sorghum and the rice-brachypodium lineage diverged ~50 million years ago (The International Brachypodium Initiative 2010), this comparison allows us to filter out genes inserted during 70% of the length of time since the pre-grass duplication. Excluding one duplicated region on rice chromosomes 11 and 12 that shows evidence of concerted evolution in multiple grass lineages (Figure S8B) (Wang *et al.* 2011), 14 of 16 homeologous regions showed at least a 10% bias in the pan-grass retained genes without homeologs (Supplementary table S2). Biased retention of genes was consistent across all of rice chromosomes 1 and 5 (Figure 5A) which are homeologous across their entire length (Salse *et al.* 2009) (Figure 5C) and are representative of most homeologous regions within the rice genome. The second largest homeologous region (shared by rice chromosomes 2 and 4) displayed a similar pattern (Figure S8A).

Bias in the number pan-grass genes with no homeolog between duplicate syntenic regions was used as a marker to assign duplicated regions to one of two subgenomes. Region which included more ancient syntenic genes without duplicates in the homeologous grass genomic region were assigned to the subgenome Grass A (under-fractionated subgenome) while the homeologous region with fewer ancient syntenic genes remaining after homeologous duplicates were excluded was assigned the the subgenome Grass B (over-fractionated subgenome) (Figure 1 and S3).

**Identification of an ancient homeologous recombination event**

The rice homeologous regions were scanned for locations where the direction of biased gene retention switched between homeologs in order to identify ancient recombination events. One such switch was identified between rice chromosomes 8 and 9 (Figure 5B,5D). Both the proximal and distal ends of the homeologous region contain more pan-grass syntenically retained genes on chromosome 9, however in the central portion of the homeologous region, more pan-grass syntenically retained genes are found on chromosome 8. The changes in content are only visible when comparing homeologous regions and not when comparing orthologous regions between species (Fig. S5). This indicates the change, likely an ancient homeologous recombination event, occurred prior to divergence of the sorghum and rice lineages. Interestingly, one of the two boundaries between the central and flanking portions of the region subsequently served as an inversion breakpoint in sorghum (Figure S5B).

**Ongoing gene loss from homeologous gene pairs**

Some homeologous duplicate genes are retained in only some of the grass species examined (Figure 4B). As with the total number of high confidence gene losses, the brachypodium genome includes the greatest number of these lost homeologous duplicates. Genes located on Grass B (under-fractionated regions) are significantly more likely to be lost from the genome of brachypodium than duplicate copies of the same set of genes located on Grass A (over-fractionated regions) ($p=.0062$, binomial test). The small bias in the same direction observed for homeologous genes lost from the rice genome is not statistically significant ($p=.2757$, binomial test). Only eight high confidence losses of homeologous genes were observed in sorghum. This likely is a result of the number of gaps in the sorghum pseudomolecules (Figure 4A) and not due to a lower overall rate of gene loss in this lineage.

**Reciprocal homeologous gene loss**

By including interspecies comparisons of the grasses, it was possible to identify reciprocally lost homeologous genes between rice, sorghum and brachypodium. For this analysis, gene sets were excluded if they contained missing genes that fall into class #2 predicted locations which include gaps in the pseudomolecules or contained genes not located in the Grass A (under-fractionated) or Grass B (over-fractionated) subgenomes.

The remaining dataset contained 1345 genes groups represented by retained duplicate genes from the pre-grass whole genome duplication. In 1111 cases -- 82.6% of the total -- both duplicate gene copies were retained in all three of sorghum, rice, and brachypodium. In another 222 cases – 16.5% -- one gene copy was retained in all three species while the other copy had been lost from the genomes of either one or two species. Genes copies located on the Grass B subgenome were marginally more likely to be the copy lost in one or more lineages – 121 cases – however this bias was not statistically significant. These lost genes were not found to be significantly enriched in any annotation using rice GOSlim terms.

In the remaining twelve cases, each copy of the gene was deleted in at least one lineage. However, in seven of these cases both copies of the gene were lost from the same species, suggesting these genes function in some non-essential role, making them unlikely candidates to drive hybrid incompatibility (Table S2). The final 5 cases (0.4% of all retained duplicated genes; 0.12% of single copy ancestral genes located within these duplicate regions) represent the only credible candidates for reproductive barriers resulting from reciprocal gene loss following whole genome duplication in the grasses and are summarized in Table 1.

**Discussion:**

**Ancient subgenomes and hidden evolutionary events**

Bias in gene loss between homeologous regions has been studied and confirmed for a wide range of species (Thomas *et al.* 2006; Sankoff *et al.* 2010; Woodhouse *et al.* 2010). However it only recently has been demonstrated that this bias is likely a property of the whole parental genomes of a tetraploid rather than of individual duplicated segments (Chang *et al.* 2010; Schnable *et al.* 2011b). As such, biased gene loss represents a powerful mark for reconstructing paleogenomes in ancient tetraploid species, even, or especially, in the absence of useful outgroups. In this study, we assigned nearly all duplicated regions in grass genomes derived from an ancient tetraploidy into low gene loss and high gene loss subgenomes, Grass A and Grass B respectively. In rice, over and under fractionated regions are often co-localized on the same chromosomes (Figure S3), meaning modern chromosomes are a chimera of subgenomes. Since reconstructions of paleochromosomes usually assume homeologous regions located on the same modern chromosome derive from the same ancestral chromosome, published reconstructions of grass ancestral protochromosomes (Salse *et al.* 2009) should be reexamined.

We identified a case in rice (chromosomes 8 and 9) and sorghum (chromosomes 2 and 7) where over- and under-fractionated regions are co-localized on the same chromosomes (Figure S4). Interestingly, this unique event is only apparent when comparing homeologous syntenic regions within a species and not orthologous syntenic regions between species. Such a pattern may occur by one of two processes: 1) fractionation

bias is not constant along a chromosome or 2) homeologous regions were exchanged between chromosomes through homeologous recombination. It has been previously reported that biased gene loss is consistent across entire ancestral chromosomes in maize and entire parental genomes in *Arabidopsis suecic*a (Chang *et al.* 2010; Schnable *et al.* 2011b) providing evidence that fractionation bias does not change across a chromosome. Additionally, one end of this apparently exchanged region later served as an inversion breakpoint on sorghum chromosome 7, which is consistent with current models regarding the reuse of chromosome breakpoints (Larkin *et al.* 2009).

Biased fractionation is likely a result of genome dominance (Schnable *et al.* 2011b), a phenomena observed in numerous allotetraploid species where genes from one parental genome tend to show higher expression in wide hybrids or allopolyploids than homeologous genes from originating from the other parental species (Buggs *et al.* 2010; Chang *et al.* 2010; Flagel and Wendel 2010). Given that genome dominance appears to be linked to qualitative differences between parental genomes rather than mode of inheritance (paternal vs maternal) (Flagel and Wendel 2010), the bias we observed may be evidence that the pregrass duplication resulted from allopolyploidy.

**Incomplete coverage of the pre-grass tetraploidy**

Only 65.7% of the rice genome has an identified homeologous region from the pre-grass tetraploidy (Yu *et al.* 2005). Deletion of large genomic regions has been observed in newly synthesized polyploids (Gaeta *et al.* 2007) so it might be argued that our analyses, which exclude all genes without identified homeologous regions, exclude a major category of fractionating gene loss. However, in an analysis of the several million year old maize tetraploidy, almost no evidence was found for large segmental deletions from either subgenome. The largest gaps in the syntenic coverage of the sorghum genome by maize (Figure S6) were shared by both maize subgenomes and particularly centered around centromeres. This finding is consistent with a previously report that there was no evidence of large deletions (>= 4 sequential genes) during fractionation in maize (Woodhouse *et al.* 2010). Therefore, the incomplete coverage of the sorghum, rice, and brachypodium genomes by duplicated segments from the pre-grass whole genome duplication likely results from duplications where the syntenic signal has sunk below the limits of detectability as the result of ongoing fractionation, gene insertion, chromosomal rearrangements, and genome assembly errors.

An unduplicated outgroup sequence will aid in the identification of these highly fractionated and rearranged regions for all grasses. While large deletions are common in the early generations of a newly tetraploid species, large scale deletions will almost always include one or more dose-sensitive genes and are expected to be selected against in subsequent generations, allowing paleopolyploids to retain near complete subgenomes at the level of whole regions, even as individual genes are lost by fractionation (Xiong *et al.* 2011).

**Ancient and Ongoing Gene Loss**

To enable the study of biased gene loss following whole genome duplication in the grasses, it was necessary to develop accurate methods of identifying genes which truly have been deleted from their ancestral location. We found that the rate of gene loss in the rice and brachypodium lineages has been significantly different. The rate of syntenic gene loss in the brachypodium lineage has been 75-115% higher than in rice since the divergence of those two lineages. The direction of this difference, although not the absolute rates of gene loss, is consistent with a study of genomic regions in sequenced grass orthologous to nine sequenced contigs from *Aegilops tauschii* (Massa *et al.* 2011). If the increased rate of gene loss observed in brachypodium is explained by the same evolutionary pressures for a small genome size that resulted in the brachypodium genome being only half the size of the rice genome, the fact that genes located on Grass B were significantly more likely to be lost in brachypodium than their homeologous duplicates on Grass A suggests that even after tens of millions of years, Grass B genes remain the more expendable member of a gene pair. The increased levels of unannotated syntenic blast hits in brachypodium may represent gene fragments generated by the on going deletion of genes via the same short deletion mechanism shown to remove genes in maize (Woodhouse *et al.* 2010). An alternative explanation is that these syntenic blast hits represent real genes missed during the annotation of the brachypodium genome. However, even if these genes are not counted as losses, Grass subgenome B has lost more genes in brachypodium (Figure 4B).

The rate of ongoing fractionation in the grasses may be higher than we are measuring. Grass B gene copies are more prone to fractionation overall. A study in yeast reported that in the later stages of fractionation the same copy of individual gene pair tends to be lost independently in multiple lineages (Scannell *et al.* 2006). Both of these pieces of data would tend to suggest that a significant number of duplicate pairs may have independently been lost in multiple lineages following the major grass lineage radiation. While independent deletions of the same gene copy would not create reproductive barriers, it is important to consider their existence when measuring the rate of fractionation in the grasses.

Based on data from teleosts and yeast the Wolfe laboratory has presented the hypothesis that genome duplications may sometimes drive speciation by increasing the speed at which reproductive barriers form. Even a small number of reciprocally lost loci between separate populations could result in hybrid offspring being unlikely to possess a full complement of essential genes (Lynch and Force 2000). The grasses, a diverse and highly successful clade whose origin is associated with genome duplication seemed an likely candidate for reciprocal gene loss driven speciation. However the frequency of reciprocally lost genes we observed was strikingly lower than that found in studies of whole genome duplication in other lineages. In polyploid yeast, 4-7% of ancestral loci

examined were identified as homeologs which had been reciprocally lost between different species (Scannell *et al.* 2006). A study in the ray-finned fishes (teleosts) reported that 8% of single copy genes between zebrafish and Tetraodon where in fact reciprocally lost homeologs (Sémon and Wolfe 2007a). Our own identification of only 5 putative reciprocally lost homeologs in the grasses out of thousands of gene pairs and single copy syntenic genes examined is strikingly different. One possible explanation for the difference we observe is that the teleosts and and yeast WGDs represent autopolyploidies, and, in the absence of genome dominance differentiated between two parental subgenomes, the early fractionation of gene pairs was more stochastic in these lineages, resulting in greater numbers of RGL events. This agrees with the observation in yeast that early gene losses were equally likely to remove either copy of a duplicate gene pair (Scannell *et al.* 2006). In plants, while the majority of polyploidy events are predicted to be autopolyploidies (Ramsey and Schemske 1998), the majority of named polyploid species arise through allopolyploidy (Mallet 2007). The impact of various forms on polyploidy on speciation and evolutionary success has been well reviewed (Rieseberg and Willis 2007; Soltis and Soltis 2009).

It may be tempting dismiss these finding as a result of the young age of the pre-grass tetraploidy relative to the yeast and teostate duplications. However, the hypothesis that reciprocal loss of duplicated genes enables increased rates of speciation requires that these gene deletions occur contemporaneously with speciation, and this was, in fact, found to be the case in yeast (Scannell *et al.* 2006). A small number of grass species diverged prior to the split between the most recent common ancestor of the maize-sorghum and rice-brachypodium lineages (Grass Phylogeny Working Group 2001) and these lineages may hold more examples of reciprocal gene losses. However the vast majority of grass species diverged contemporaneous with or following the maize-sorghum rice-brachypodium split (Grass Phylogeny Working Group 2001). Given the lack of evidence for significant levels of reciprocal gene loss from this point onwards, we conclude that reciprocal gene loss of duplicate genes resulting from whole genome duplication was probably not responsible for the radiation of the primary grass lineages. This contrasts individual reports that the reciprocal loss of duplicates genes resulting from individual dispersed duplications create hybrid incompatibility in arabidopsis and rice (Bikard *et al.* 2009; Mizuta *et al.* 2010).

**Concluding remarks**

Having multiple whole genome sequences for several clades of organisms provides a rich dataset for studying the evolution of genomes. Angiosperm genomes, in general, are remarkable for having repeated whole genome duplication events that permeate their lineages. In particular, the grass lineage combines these two facets: several grass genomes are currently available with several more arriving soon, and a whole genome duplication event occurred prior to their radiation. We show that by classifying the evolutionary history of sets of genes and identifying the subgenomes comprising modern

grass genomes provides an opportunity to understand the evolution of individual genomes and the grass lineage as a whole. Importantly, the ongoing process of fractionation remains biased in the grasses preferentially and consistently targeting one subgenome for gene loss, and that unlike previously studies in yeast and teleosts, reciprocal gene loss of duplicated genes is not likely to be the driving force of the grass radiation.

**Figures**

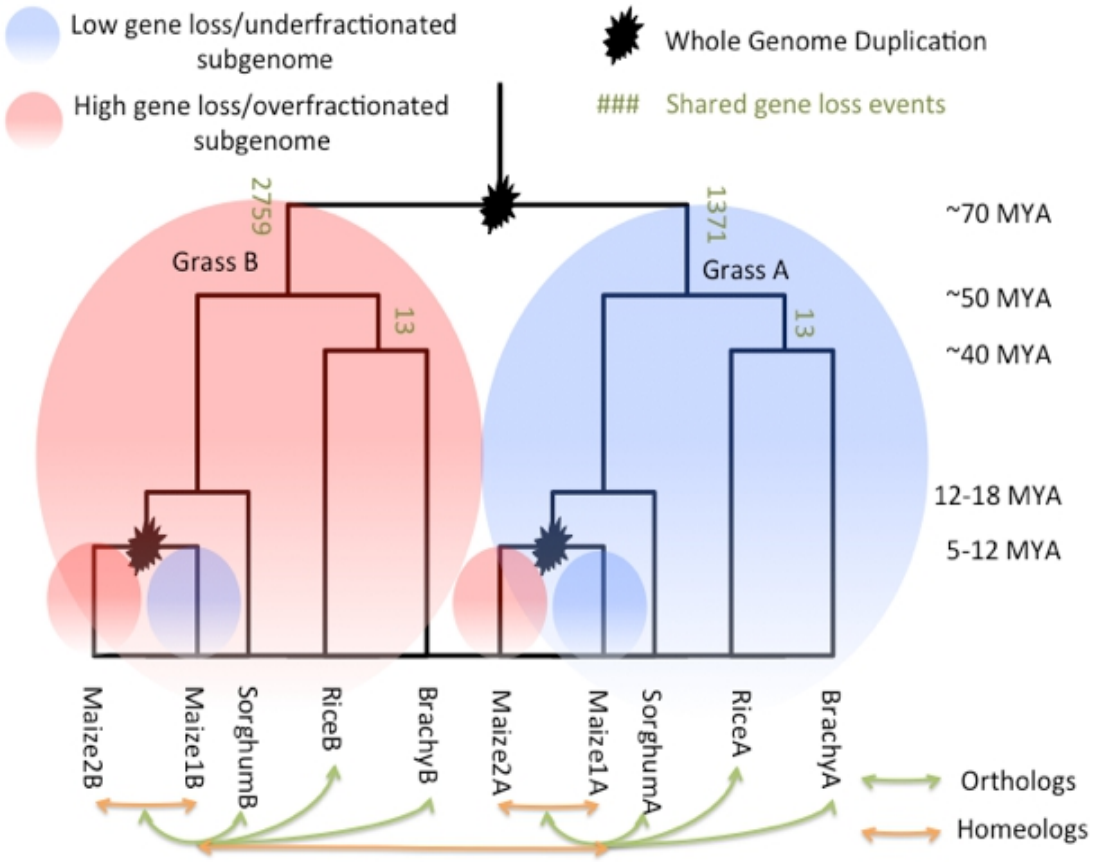



Figure 1
Relationships of a genomic location in the grasses, taking into account both the ancient whole genome duplication in the ancestor of all sequenced grass species and the more recent genome duplication in the maize lineage. Each duplication creates separate homeologous low gene loss (under fractionated) and high gene loss (over fractionated) subgenomes. Two loci are orthologous if the branch point where they diverged represents a speciation event (no mark), or homeologous if the branch point where they diverged is a whole genome duplication (marked with a starburst). Branch lengths not to scale.

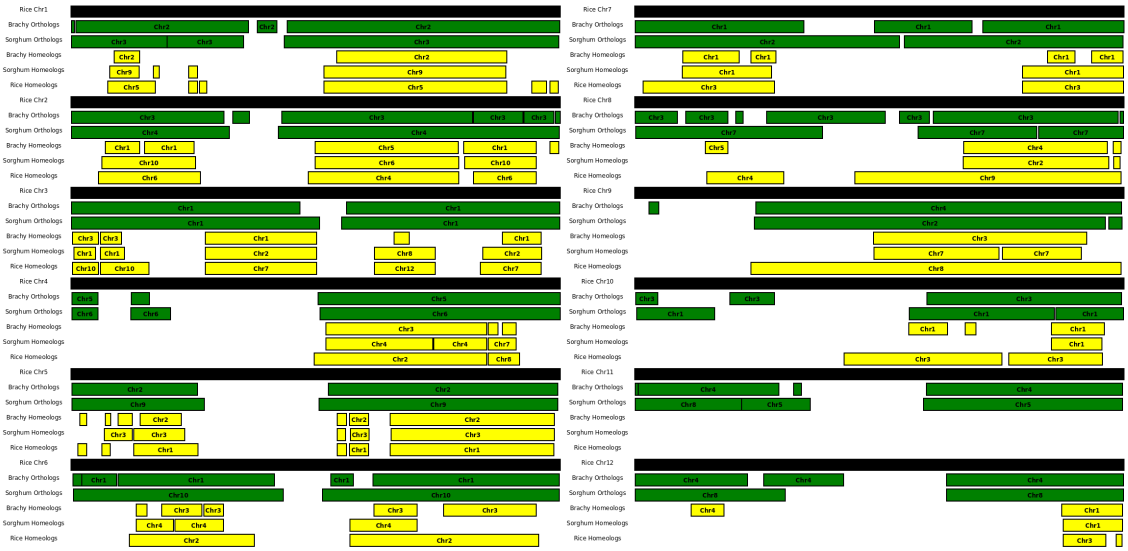


Figure 2
Orthologous and homeologous coverage of the rice genome by syntenic regions in the sorghum, brachypodium and rice genomes. Orthologous syntenic regions are marked in green and homeologous ones are marked in yellow. Coverage is scaled by gene counts, not nucleotides, which will tend to accentuate the gene rich chromosome arms and deemphasize the gene poor pericentromeric regions.

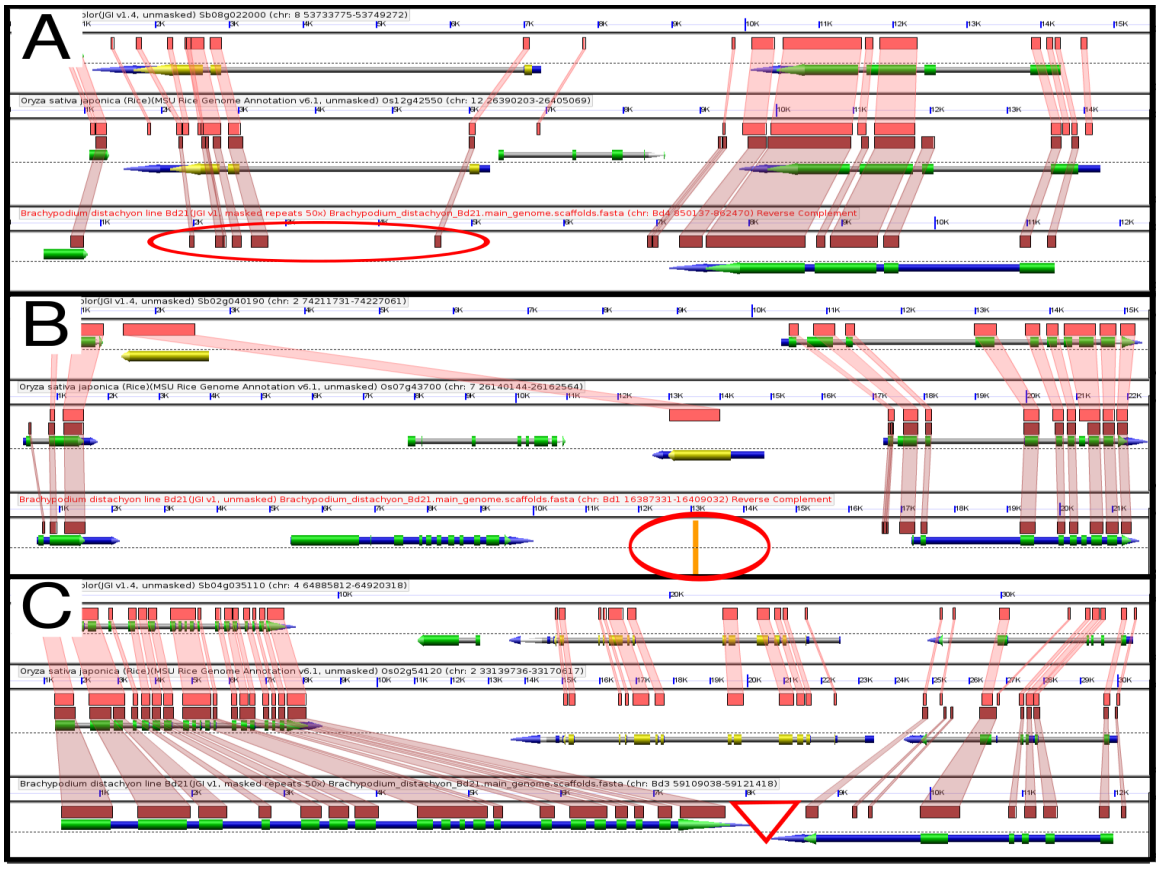



Figure 3
Three subcategories of potential gene loss identified by syntenic analysis. In each case a gene conserved in sorghum (top panel) and rice (middle panel) is shown along with the syntenically predicted orthologous location in the brachypodium genome (bottom panel). Each panel represents a genomic region with the dashed line separating the top and bottom strands of DNA. Gene models are composite arrows with gray representing the extent of the gene, blue the mRNA, and green/yellow protein coding sequence. A) No gene model corresponding to the conserved gene in rice (Os12g42550) and sorghum (Sb08g022000) was annotated in brachypodium, however unannotated sequence present at the predicted orthologous location in brachypodium (marked with a red circle) is similar to the coding sequence of the annotated rice and sorghum genes. B) Neither an annotated gene nor unannotated sequence in brachypodium corresponds to the syntenically conserved gene in rice (Os07g43700) and sorghum (Sb02g040190). However, a gap in the brachypodium genome assembly (orange bar marked with the red circle) raises the possibility that the brachypodium ortholog of these genes was simply not captured during the whole genome shotgun sequencing of the brachypodium genome, or not correctly assembled into the pseudomolecule. C) A high confidence gene deletion.

The example gene Sb04g035110/Os02g54120 has a predicted orthologous location (red triangle) which does not contain an orthologous gene, unannotated homologous sequence, or a gap in the pseudomolecule assembly.

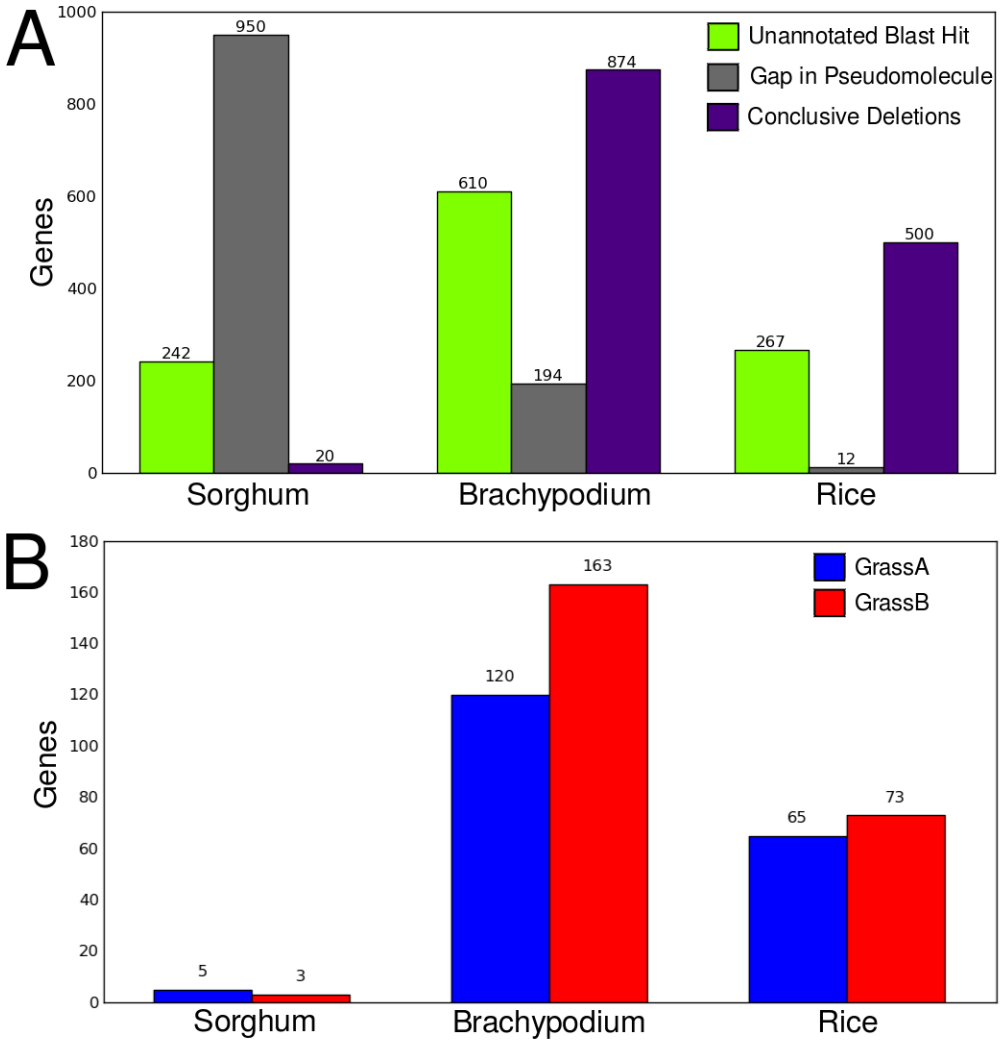



Figure 4
Rates of gene loss between species and subgenomes. A) Genome-wide counts of the three types of gene loss described in Figure 3 for the sorghum, brachypodium, and rice genomes. B) Counts of only conclusive deletions located in regions assigned to the GrassA or GrassB subgenomes. Only gene deletions where the homeologous duplicate is still retained by the species were counted in this analysis.

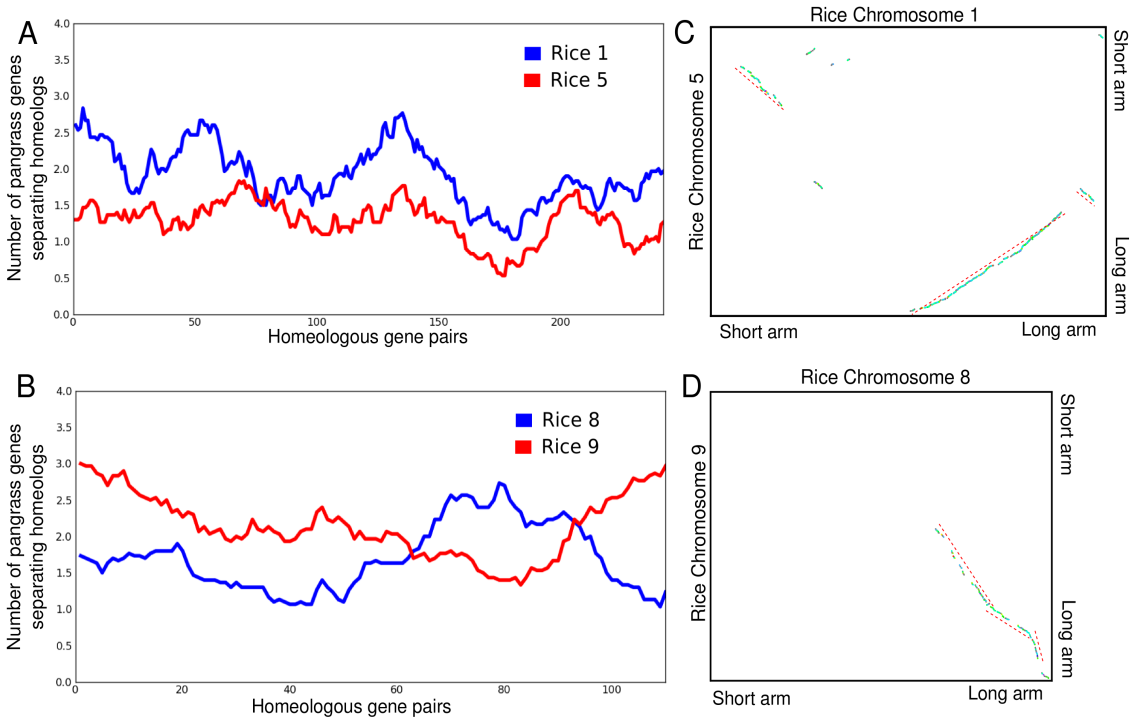



Figure 5

Bias in gene content between homeologous rice chromosomes. A) Running average of pan-grass genes between the homeologous regions of rice chromsomes 1 and 5. B) Running average of pan-grass genes between the homeologous regions of rice chromsomes 8 and 9. C and D are dotplots showing syntenic regions identified between pairs of rice chromosomes. These dotplots are scaled using gene content rather than total nucleotides so the slope of syntenic diagonals represents a crude measure of fractionation bias. C) Comparison of rice chromosomes 1 and 5. D) Comparison of rice chromosomes 8 and 9.

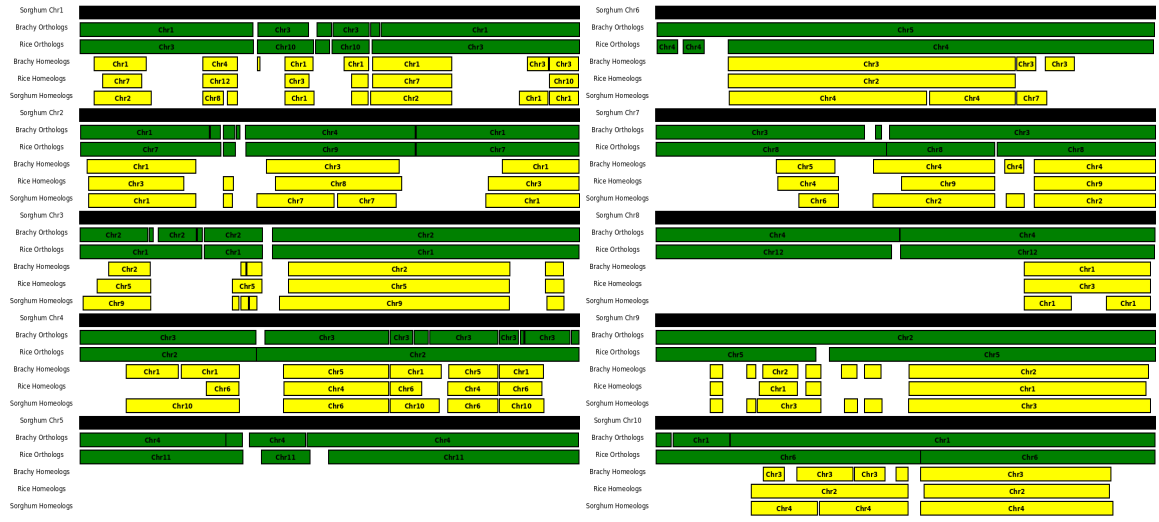



Figure S1:
Version of Figure 2 using the sorghum genome as a reference.

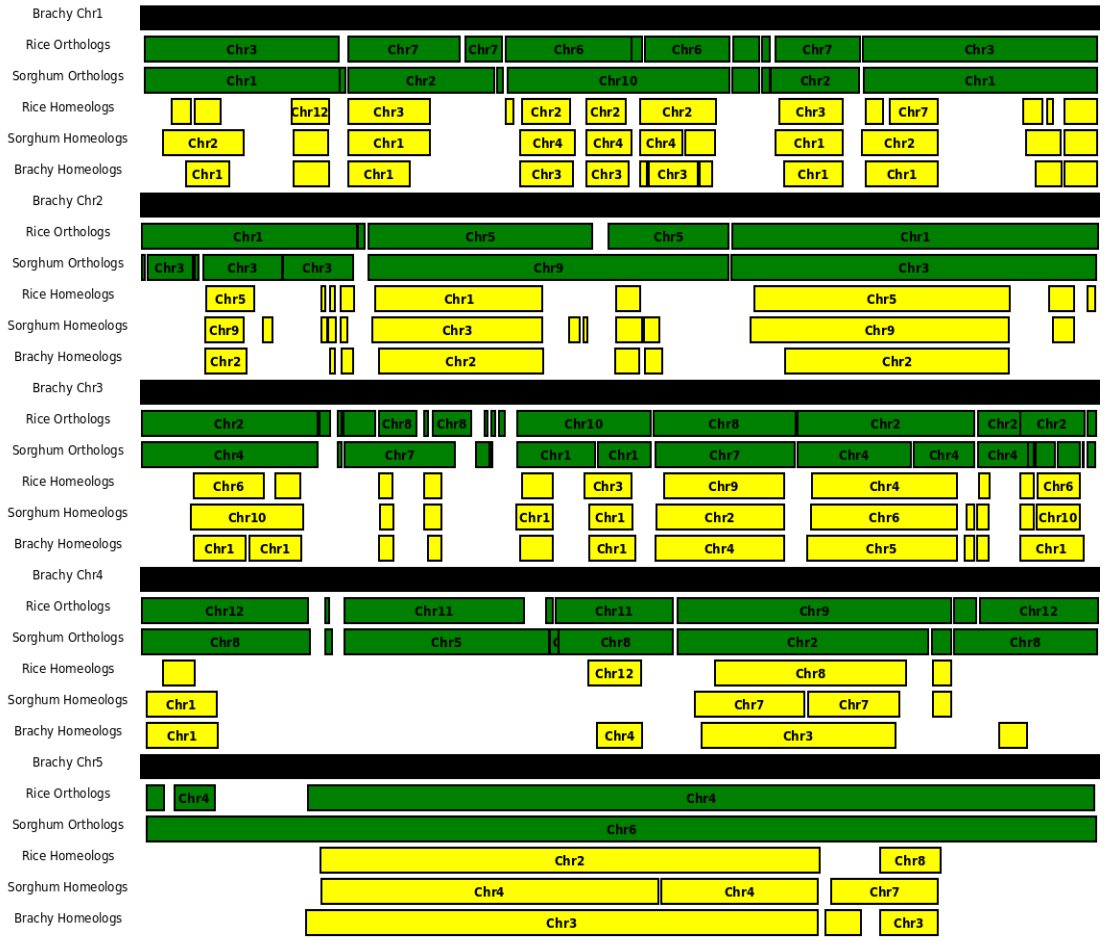



Figure S2:
Version of Figure 2 using the brachypodium genome as a reference

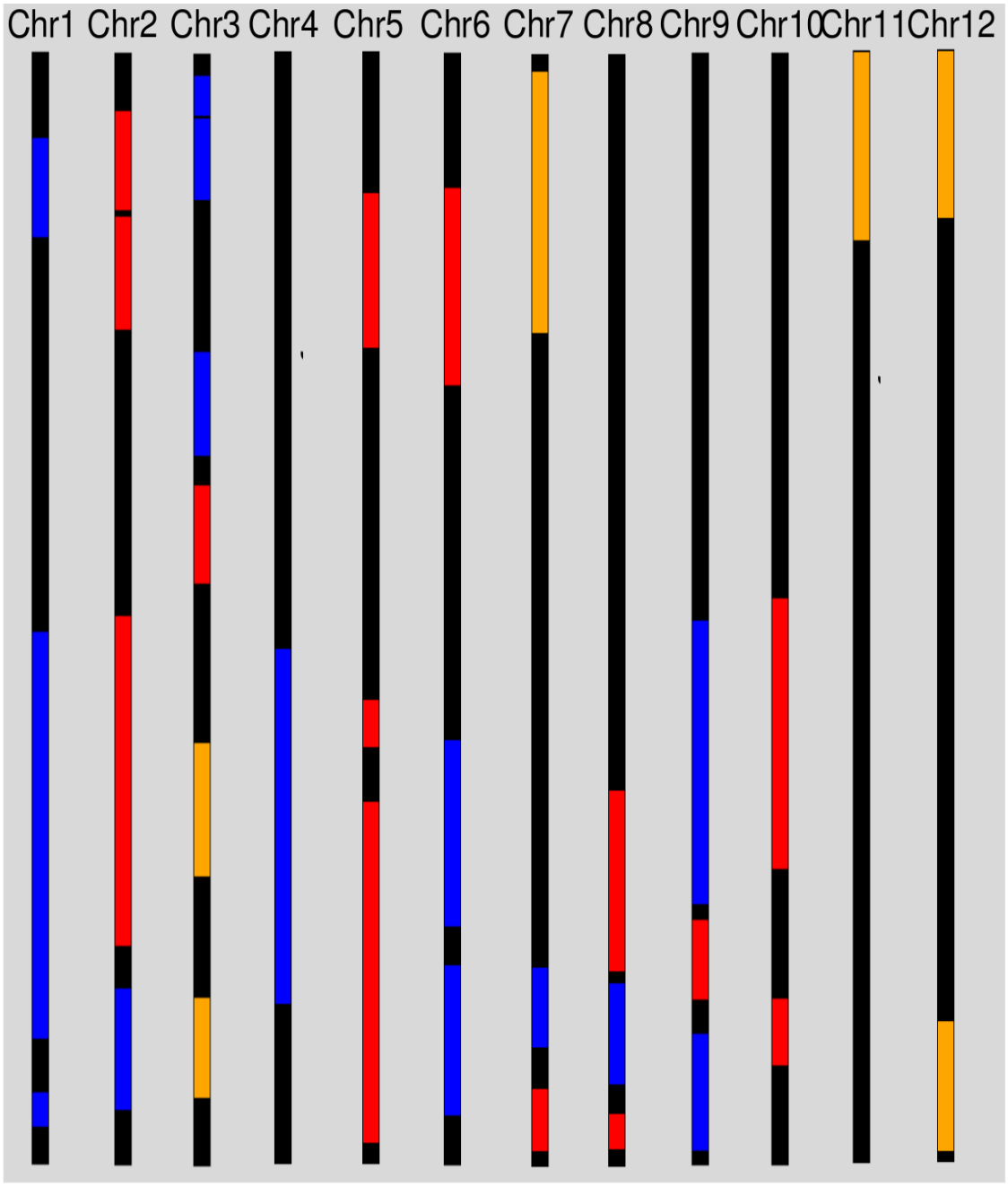



Figure S3
Division of the rice genome into GrassA (low gene loss, under fractionated) in blue and Grass B (high gene loss, over fractionated) subgenomes in red. Orange regions mark homeologous portions of the genome without sufficient bias in gene loss to be assigned to either subgenome. Black regions are those without a detectable syntenic homeologous duplicate.

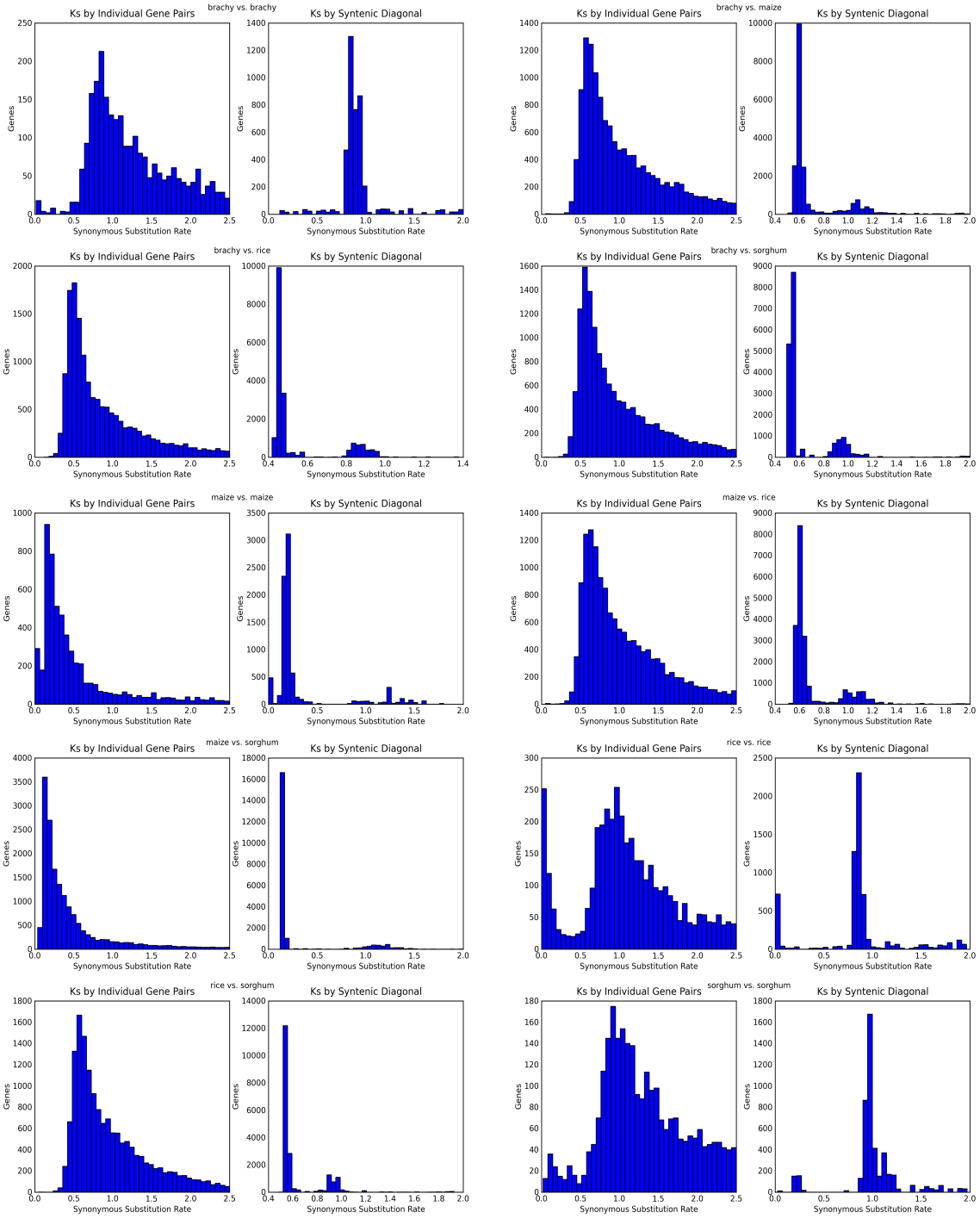



Figure S4
Comparison of the distributions of synonymous substitution rates for individual gene pairs in syntenic diagonals and the distribution when each gene pair is assigned a

synonymous substitution rate based on the aggregate divergence of all gene pairs located in the same syntenic block.

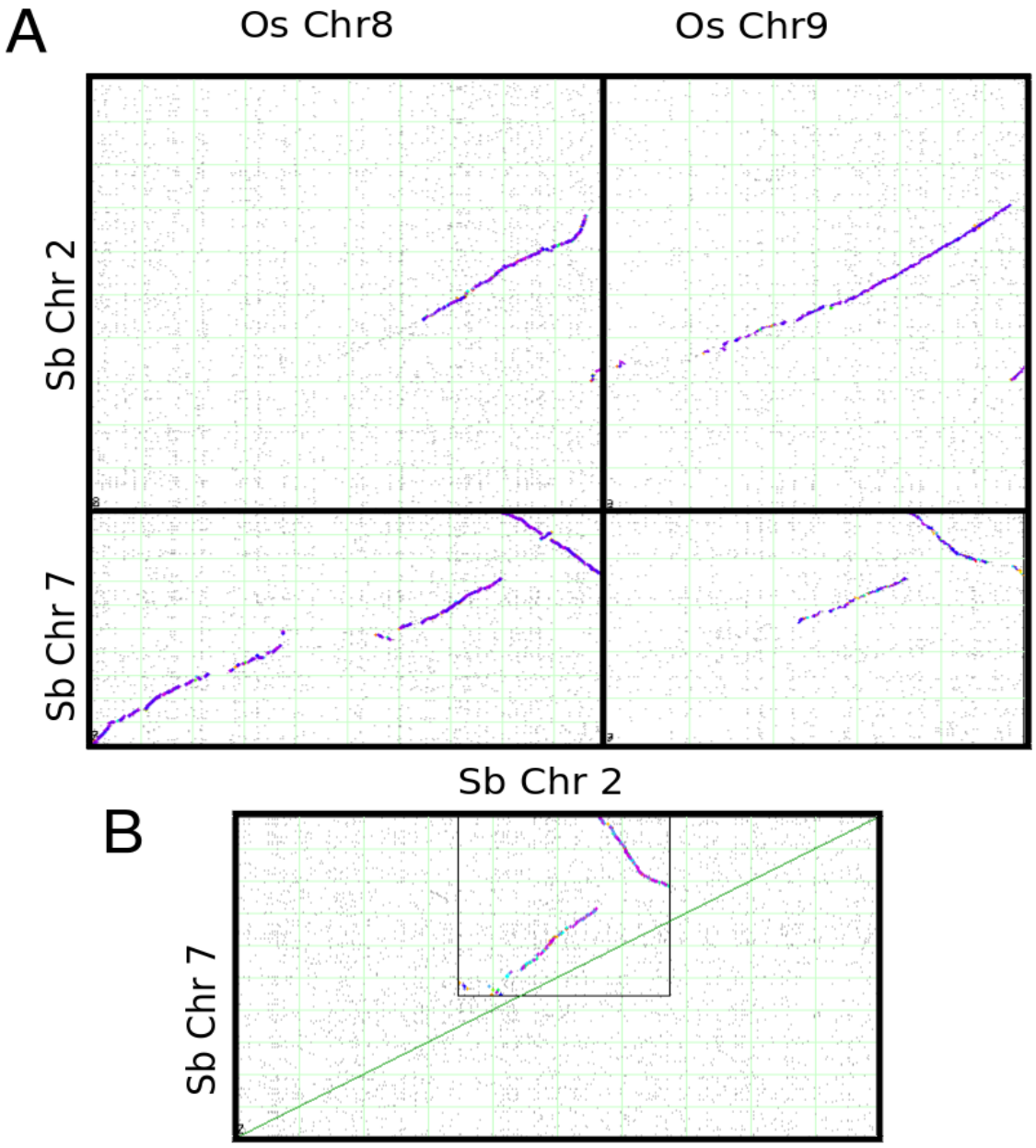



Figure S5
Further investigation of the homeologously exchanged region of rice chromosomes 8/9 using the orthologous regions of sorghum chromosomes 2/7 as an outgroup. A: The change in gene density is visible in the homeologous comparisons (Sb2/Os8 and Sb7/Os9) but not in orthologous comparisons (Sb2/Os9 and Sb7/Os8), indicating the exchange occurred prior to the divergence of the rice and sorghum lineages. B: One end of the exchanged region served as an inversion breakpoint on sorghum chromosome 7 after the rice sorghum divergence.

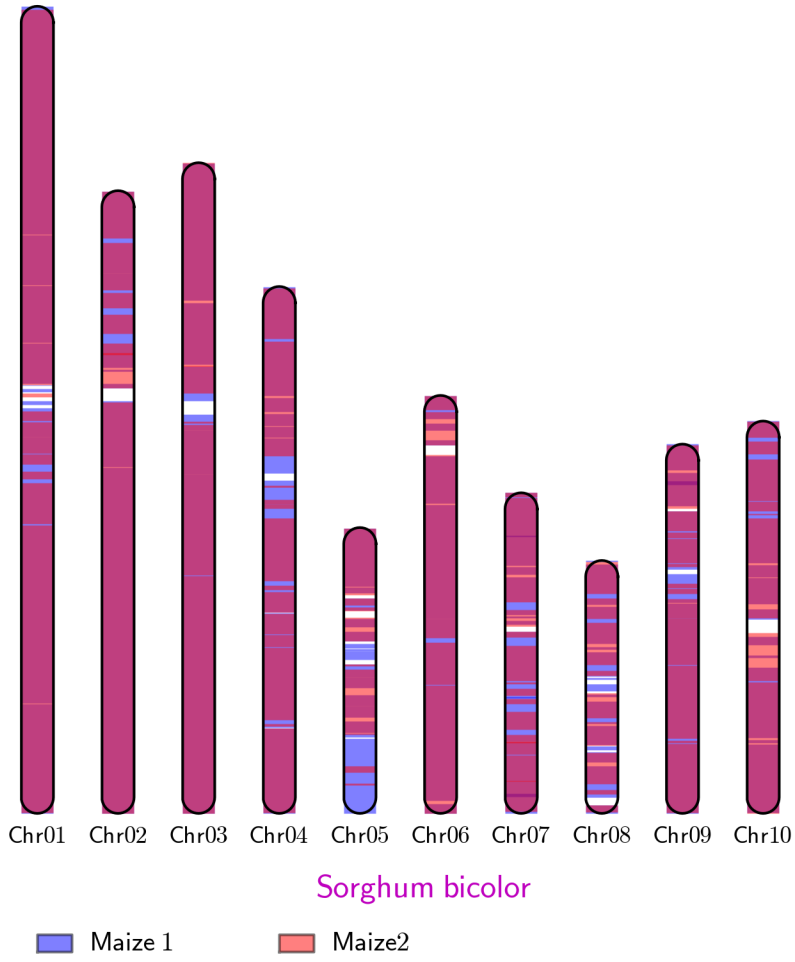



Figure S6
Orthologous coverage of the sorghum genome by regions of the maize region belonging to the maize1 and maize2 subgenomes. Areas of the sorghum genome marked in purple are represented orthologously in both the maize1 and maize2 subgenomes. Areas marked in blue are represented orthologously in only the maize1 subgenome. Areas marked in red are represented orthologously only in the maize2 subgenome. Areas left white had no detectable syntenic orthologous relationship to maize. These regions are concentrated in centromeric and pericentromeric regions where the syntenic signal of orthologous regions is expected to be less detectable.

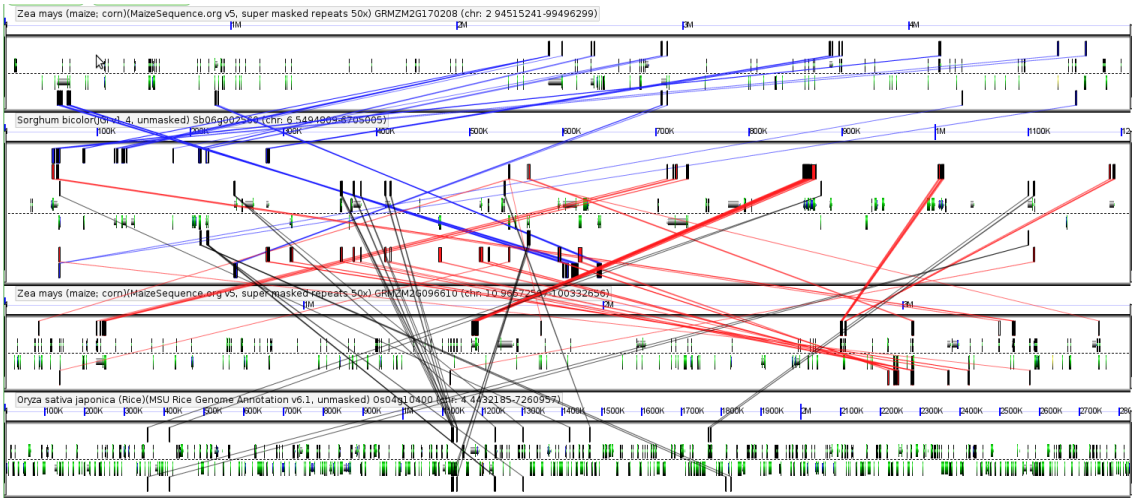



Figure S7

Example of a manually disqualified candidate region for a large scale (>= 40 genes) segmental deletion from maize. Panel legend same as Fig 3. This image spans a 54 gene region of sorghum chromosome 6 computationally identified as deleted from maize subgenome1. Red boxes/lines mark blast (lastz) hits to the orthologous maize2 region on chromosomes which was identified automatically. Black boxes mark blast hits to the orthologous region of the rice genome. Blue boxes mark blast hits to a manually identified syntenic segment on maize chromosome 2 spanning more than half of the putative deletion, disqualifying it. Compared regions use only annotated protein coding sequences (all other regions were masked from the analysis). Results may be regenerated at http://genomevolution.org/r/2vch .

## Tables

### Table 1

| Sorghum Gene(s) | Rice Gene(s) | Brachy Gene(s) | Annotation | Link |
|---|---|---|---|---|
| Sb04g033870 | | Bradi3g58300 | | |
| Sb10g007450 | Os06g11410 | | Cyclin | http://genomevolution.or |
| Sb02g011380 | | Bradi4g38330/40 | Calcineurin-like | |
| Sb07g024380 | Os08g44300 | | phosphoesterase | http://genomevolution.or |
| | | | Regulator of chromosome | |
| Sb06g020480/90 | | Bradi5g13690 | condensation | http://genomevolution.o |
| Sb04g024990 | Os02g38350 | | domain containing | rg/r/2qwr |
| | | | OsArgos: Arabidopsis | |
| Sb06g017750 | Os04g36670 | | ortholog regulates | |
| Sb04g023130 | | unannotated sequence | organ size | http://genomevolution.or |

Table S1

| Rice GOSlim Category | Ratio in Genes Lost in Brachy | Ratio in All Syntenic Rice Ge | Ratio in All Rice Genes |
|---|---|---|---|
| endoplasmic reticulum | 0.037338 (23/616 genes) | 0.030209 (322/10659 genes) | 0.013355 (619/46349 genes) |
| signal transduction | 0.055195 (34/616 genes) | 0.056384 (601/10659 genes) | 0.032255 (1495/46349 genes) |
| photosynthesis | 0.001623 (1/616 genes) | 0.002252 (24/10659 genes) | 0.000302 (14/46349 genes) |
| reproduction | 0.004870 (3/616 genes) | 0.006473 (69/10659 genes) | 0.002438 (113/46349 genes) |
| vacuole | 0.009740 (6/616 genes) | 0.012853 (137/10659 genes) | 0.004013 (186/46349 genes) |
| membrane | 0.094156 (58/616 genes) | 0.089689 (956/10659 genes) | 0.037649 (1745/46349 genes) |
| peroxisome | 0.001623 (1/616 genes) | 0.003471 (37/10659 genes) | 0.000647 (30/46349 genes) |
| regulation of gene expression, | 0.004870 (3/616 genes) | 0.000751 (8/10659 genes) | 0.000237 (11/46349 genes) |
| nucleotide binding | 0.079545 (49/616 genes) | 0.062858 (670/10659 genes) | 0.035535 (1647/46349 genes) |
| transferase activity | 0.055195 (34/616 genes) | 0.061357 (654/10659 genes) | 0.028264 (1310/46349 genes) |
| respiratory electron transport c | 0.000000 (0/616 genes) | 0.000563 (6/10659 genes) | 0.000367 (17/46349 genes) |
| cellular homeostasis | 0.006494 (4/616 genes) | 0.003753 (40/10659 genes) | 0.001338 (62/46349 genes) |
| Golgi apparatus | 0.008117 (5/616 genes) | 0.023079 (246/10659 genes) | 0.003495 (162/46349 genes) |
| oxygen binding | 0.009740 (6/616 genes) | 0.003190 (34/10659 genes) | 0.006753 (313/46349 genes) |
| response to biotic stimulus | 0.042208 (26/616 genes) | 0.044845 (478/10659 genes) | 0.024725 (1146/46349 genes) |
| tropism | 0.000000 (0/616 genes) | 0.000094 (1/10659 genes) | 0.000000 (0/46349 genes) |
| response to external stimulus | 0.021104 (13/616 genes) | 0.023736 (253/10659 genes) | 0.013139 (609/46349 genes) |
| protein metabolic process | 0.038961 (24/616 genes) | 0.046158 (492/10659 genes) | 0.018080 (838/46349 genes) |
| cytosol | 0.011364 (7/616 genes) | 0.010414 (111/10659 genes) | 0.005113 (237/46349 genes) |
| molecular_function | 0.021104 (13/616 genes) | 0.032179 (343/10659 genes) | 0.012212 (566/46349 genes) |
| binding | 0.071429 (44/616 genes) | 0.067173 (716/10659 genes) | 0.021834 (1012/46349 genes) |
| nucleic acid binding | 0.025974 (16/616 genes) | 0.026738 (285/10659 genes) | 0.005761 (267/46349 genes) |
| DNA binding | 0.063312 (39/616 genes) | 0.067361 (718/10659 genes) | 0.018533 (859/46349 genes) |
| thylakoid | 0.025974 (16/616 genes) | 0.032555 (347/10659 genes) | 0.006235 (289/46349 genes) |
| cell growth | 0.000000 (0/616 genes) | 0.004503 (48/10659 genes) | 0.001661 (77/46349 genes) |
| cellular component organizatic | 0.032468 (20/616 genes) | 0.032179 (343/10659 genes) | 0.009256 (429/46349 genes) |
| nucleoplasm | 0.003247 (2/616 genes) | 0.001501 (16/10659 genes) | 0.000345 (16/46349 genes) |
| motor activity | 0.003247 (2/616 genes) | 0.004691 (50/10659 genes) | 0.000949 (44/46349 genes) |
| hydrolase activity | 0.073052 (45/616 genes) | 0.101229 (1079/10659 genes) | 0.033507 (1553/46349 genes) |
| kinase activity | 0.053571 (33/616 genes) | 0.047003 (501/10659 genes) | 0.035837 (1661/46349 genes) |
| lipid metabolic process | 0.022727 (14/616 genes) | 0.017825 (190/10659 genes) | 0.008048 (373/46349 genes) |
| transport | 0.048701 (30/616 genes) | 0.049629 (529/10659 genes) | 0.014412 (668/46349 genes) |
| catalytic activity | 0.086039 (53/616 genes) | 0.092785 (989/10659 genes) | 0.038081 (1765/46349 genes) |
| response to stress | 0.110390 (68/616 genes) | 0.094756 (1010/10659 genes) | 0.062418 (2893/46349 genes) |
| biological_process | 0.016234 (10/616 genes) | 0.028990 (309/10659 genes) | 0.012018 (557/46349 genes) |
| metabolic process | 0.025974 (16/616 genes) | 0.027301 (291/10659 genes) | 0.011457 (531/46349 genes) |
| ribosome | 0.009740 (6/616 genes) | 0.009100 (97/10659 genes) | 0.005178 (240/46349 genes) |
| transcription | 0.064935 (40/616 genes) | 0.071395 (761/10659 genes) | 0.018382 (852/46349 genes) |
| multicellular organismal devel | 0.025974 (16/616 genes) | 0.018951 (202/10659 genes) | 0.006602 (306/46349 genes) |
| cellular amino acid metabolic | 0.040584 (25/616 genes) | 0.021297 (227/10659 genes) | 0.016225 (752/46349 genes) |
| cell differentiation | 0.009740 (6/616 genes) | 0.006661 (71/10659 genes) | 0.005502 (255/46349 genes) |
| post-embryonic development | 0.009740 (6/616 genes) | 0.009945 (106/10659 genes) | 0.002567 (119/46349 genes) |
| embryo development | 0.001623 (1/616 genes) | 0.002533 (27/10659 genes) | 0.000583 (27/46349 genes) |
| anatomical structure morphog | 0.004870 (3/616 genes) | 0.005160 (55/10659 genes) | 0.001446 (67/46349 genes) |
| cell cycle | 0.006494 (4/616 genes) | 0.004316 (46/10659 genes) | 0.001640 (76/46349 genes) |
| response to endogenous stimi | 0.047078 (29/616 genes) | 0.056478 (602/10659 genes) | 0.036268 (1681/46349 genes) |
| nucleobase, nucleoside, nuc | 0.012987 (8/616 genes) | 0.018482 (197/10659 genes) | 0.004164 (193/46349 genes) |
| nucleolus | 0.016234 (10/616 genes) | 0.013979 (149/10659 genes) | 0.004229 (196/46349 genes) |
| nucleus | 0.073052 (45/616 genes) | 0.091378 (974/10659 genes) | 0.026948 (1249/46349 genes) |
| nuclear envelope | 0.001623 (1/616 genes) | 0.001783 (19/10659 genes) | 0.001402 (65/46349 genes) |
| mitochondrion | 0.084416 (52/616 genes) | 0.081058 (864/10659 genes) | 0.038232 (1772/46349 genes) |
| transporter activity | 0.040584 (25/616 genes) | 0.052256 (557/10659 genes) | 0.015772 (731/46349 genes) |
| translation factor activity, nuc | 0.009740 (6/616 genes) | 0.004972 (53/10659 genes) | 0.001467 (68/46349 genes) |
| lipid binding | 0.006494 (4/616 genes) | 0.007318 (78/10659 genes) | 0.002589 (120/46349 genes) |
| No GO terms | 0.405844 (250/616 genes) | 0.359602 (3833/10659 genes) | 0.760189 (35234/46349 genes) |
| ripening | 0.001623 (1/616 genes) | 0.001501 (16/10659 genes) | 0.000669 (31/46349 genes) |
| nuclease activity | 0.004870 (3/616 genes) | 0.004785 (51/10659 genes) | 0.001618 (75/46349 genes) |
| cell | 0.011364 (7/616 genes) | 0.006098 (65/10659 genes) | 0.001187 (55/46349 genes) |
| intracellular | 0.019481 (12/616 genes) | 0.022985 (245/10659 genes) | 0.005006 (232/46349 genes) |
| RNA binding | 0.016234 (10/616 genes) | 0.021390 (228/10659 genes) | 0.004854 (225/46349 genes) |
| protein binding | 0.076299 (47/616 genes) | 0.073178 (780/10659 genes) | 0.046171 (2140/46349 genes) |
| receptor binding | 0.000000 (0/616 genes) | 0.000000 (0/10659 genes) | 0.000237 (11/46349 genes) |
| chromatin binding | 0.003247 (2/616 genes) | 0.002064 (22/10659 genes) | 0.000129 (6/46349 genes) |
| response to extracellular stimi | 0.006494 (4/616 genes) | 0.003940 (42/10659 genes) | 0.001855 (86/46349 genes) |
| carbohydrate binding | 0.003247 (2/616 genes) | 0.002721 (29/10659 genes) | 0.010162 (471/46349 genes) |
| receptor activity | 0.008117 (5/616 genes) | 0.005160 (55/10659 genes) | 0.009148 (424/46349 genes) |
| signal transducer activity | 0.006494 (4/616 genes) | 0.012102 (129/10659 genes) | 0.003215 (149/46349 genes) |
| cell death | 0.001623 (1/616 genes) | 0.003753 (40/10659 genes) | 0.002050 (95/46349 genes) |
| generation of precursor metab | 0.006494 (4/616 genes) | 0.003940 (42/10659 genes) | 0.000669 (31/46349 genes) |
| lysosome | 0.000000 (0/616 genes) | 0.000188 (2/10659 genes) | 0.000065 (3/46349 genes) |
| plastid | 0.048701 (30/616 genes) | 0.059762 (637/10659 genes) | 0.013053 (605/46349 genes) |
| growth | 0.001623 (1/616 genes) | 0.004034 (43/10659 genes) | 0.001726 (80/46349 genes) |
| cell wall | 0.032468 (20/616 genes) | 0.031992 (341/10659 genes) | 0.020842 (966/46349 genes) |
| transcription regulator activity | 0.014610 (9/616 genes) | 0.017450 (186/10659 genes) | 0.005308 (246/46349 genes) |
| structural molecule activity | 0.012987 (8/616 genes) | 0.012384 (132/10659 genes) | 0.005998 (278/46349 genes) |
| flower development | 0.011364 (7/616 genes) | 0.006567 (70/10659 genes) | 0.003279 (152/46349 genes) |
| cytoskeleton | 0.001623 (1/616 genes) | 0.006849 (73/10659 genes) | 0.001467 (68/46349 genes) |
| cellular process | 0.068182 (42/616 genes) | 0.080964 (863/10659 genes) | 0.030206 (1400/46349 genes) |
| enzyme regulator activity | 0.012987 (8/616 genes) | 0.006380 (68/10659 genes) | 0.001748 (81/46349 genes) |
| secondary metabolic process | 0.030844 (19/616 genes) | 0.013697 (146/10659 genes) | 0.016031 (743/46349 genes) |
| protein modification process | 0.038961 (24/616 genes) | 0.035088 (374/10659 genes) | 0.030400 (1409/46349 genes) |
| biosynthetic process | 0.063312 (39/616 genes) | 0.040904 (436/10659 genes) | 0.023927 (1109/46349 genes) |
| plasma membrane | 0.048701 (30/616 genes) | 0.047659 (508/10659 genes) | 0.044769 (2075/46349 genes) |
| cell communication | 0.003247 (2/616 genes) | 0.002064 (22/10659 genes) | 0.001122 (52/46349 genes) |
| catabolic process | 0.016234 (10/616 genes) | 0.025049 (267/10659 genes) | 0.006559 (304/46349 genes) |
| response to abiotic stimulus | 0.053571 (33/616 genes) | 0.050192 (535/10659 genes) | 0.021424 (993/46349 genes) |
| carbohydrate metabolic proce | 0.017857 (11/616 genes) | 0.016137 (172/10659 genes) | 0.003495 (162/46349 genes) |
| sequence-specific DNA bindir | 0.063312 (39/616 genes) | 0.075054 (800/10659 genes) | 0.019375 (898/46349 genes) |
| cytoplasm | 0.047078 (29/616 genes) | 0.056384 (601/10659 genes) | 0.018253 (846/46349 genes) |
| cellular_component | 0.001623 (1/616 genes) | 0.001783 (19/10659 genes) | 0.001100 (51/46349 genes) |
| extracellular region | 0.006494 (4/616 genes) | 0.004034 (43/10659 genes) | 0.004337 (201/46349 genes) |
| pollination | 0.001623 (1/616 genes) | 0.002439 (26/10659 genes) | 0.000475 (22/46349 genes) |
| DNA metabolic process | 0.006494 (4/616 genes) | 0.009757 (104/10659 genes) | 0.001812 (84/46349 genes) |
| endosome | 0.001623 (1/616 genes) | 0.000938 (10/10659 genes) | 0.000086 (4/46349 genes) |
| translation | 0.019481 (12/616 genes) | 0.016043 (171/10659 genes) | 0.007573 (351/46349 genes) |

# Table S2

| Sorghum (Grass1) | Rice (Grass1) | Brachy (Grass1) | Sorghum (Grass2) | Rice (Grass2) | Brachy (Grass2) | Annotation | GEvo Link |
|---|---|---|---|---|---|---|---|
| Sb03g034110 | chr1:30898252-30918678:1:gone | Bradi2g49340 | Sb09g026120 | chr5:25992012-26032694:1:gone | Bradi2g19360 | Profilin-3 like | http://genomev |
| Sb07g028130\|\|Sb07g028090\|\|Sb07g028110 | Os08g39290 | chr3:41977023-41983612:1:gone | Sb02g028130 | Os09g31040\|\|Os09g30490\|\|Os09g3( | chr4:39164642-39206468:1:gone | EF Hand famil | http://genomev |
| Sb03g037990 | Os01g60120 | chr2:52457051-52463816:1:gone | Sb09g023740 | Os05g40700 | chr2:19460832-19465859:0:gone | Transmembrar | http://genomev |
| Sb02g012201\|\|Sb02g012300 | chr9:22701946-22708110:1:gone | Bradi4g38610 | Sb07g024090 | chr8:27998952-28004402:1:gone | Bradi3g42930 | Chlorophyllase | http://genomev |
| Sb06g015025 | Os04g32960 | chr5:11296036-11303670:1:gone | Sb04g021280 | Os02g32350 | chr3:46508631-46511473:1:gone | TUDOR proteir | http://genomev |
| Sb04g028220 | chr2:31107358-31116444:0:gone | chr3:58678061-58687781:1:gone | Sb10g008350 | Os06g12690 | chr1:43142481-43158448:0:gone | DCN1-like | http://genomev |
| chr6:50718632-50721420:1:gone | Os04g41980 | chr5:18031624-18042651:1:gone | Sb04g025620 | Os02g39620 | chr3:49856952-49866342:1:gone | ATOZI1 | http://genomev |

# Table S3

| Species 1 | Species 2 | Expected Hits Per Sequence | Average Gene Distance | Max Gene Distance | Minimum # of Aligned Pairs | Merge Distance | Orthologous Ks Lower Limit | Orthologous Ks Upper Limit | Homeologous Ks Lower Limit | Homeologous Ks Lower Limit | CoGe Link |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Brachypodium distachyon | Brachypodium distachyon | 4 | 10 | 20 | 5 | 100 | NA | NA | 0.75 | 1.05 | http://genomevolution.org/r/1ojl |
| Brachypodium distachyon | Oryza sativa (MSU 5) | 4 | 15 | 30 | 5 | 100 | 0.4 | 0.6 | 0.7 | 1 | http://genomevolution.org/r/1ojw |
| Brachypodium distachyon | Sorghum bicolor | 4 | 10 | 20 | 5 | 100 | 0.45 | 0.65 | 0.75 | 1.2 | http://genomevolution.org/r/1ok1 |
| Brachypodium distachyon | Zea mays (Working Set) | 8 | 15 | 30 | 5 | 200 | 0.5 | 0.7 | 0.85 | 1.25 | http://genomevolution.org/r/1ojq |
| Oryza sativa (MSU 5) | Oryza sativa (MSU 5) | 4 | 15 | 30 | 5 | 200 | NA | NA | 0.7 | 1 | http://genomevolution.org/r/1okb |
| Oryza sativa (MSU 5) | Sorghum bicolor | 4 | 10 | 20 | 5 | 200 | 0.5 | 0.7 | 0.85 | 1.1 | http://genomevolution.org/r/1oko |
| Oryza sativa (MSU 5) | Zea mays (Working Set) | 8 | 15 | 30 | 5 | 300 | 0.5 | 0.7 | 0.9 | 1.3 | http://genomevolution.org/r/1okj |
| Sorghum bicolor | Sorghum bicolor | 4 | 10 | 20 | 5 | 100 | NA | NA | 0.8 | 1.3 | http://genomevolution.org/r/1okr |
| Sorghum bicolor | Zea mays (Working Set) | 8 | 15 | 30 | 5 | 200 | 0.1 | 0.25 | 0.85 | 1.25 | http://genomevolution.org/r/1oku |
| Zea mays (Working Set) | Zea mays (Working Set) | 8 | 15 | 30 | 8 | 300 | 0.05 | 0.45 | NA | NA | http://genomevolution.org/r/1okv |

## Table S4

| Chr1 | Start1 | Stop1 | Chr2 | Start2 | Stop2 | Len1 | Len2 | Syntenic genes without homeologs1 | Syntenic genes without homeologs2 | Ratio |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 41169000 | 42560000 | 5 | 17485000 | 18789000 | 1391000 | 1304000 | 52 | 26 | 2 |
| 1 | 22931000 | 39073000 | 5 | 29464000 | 20232000 | 16142000 | 9232000 | 593 | 208 | 2.8509615385 |
| 1 | 3398000 | 7342000 | 5 | 3818000 | 8019000 | 3944000 | 4201000 | 150 | 46 | 3.2608695652 |
| 2 | 5126000 | 1872000 | 6 | 25640000 | 29886000 | 3254000 | 4246000 | 90 | 160 | 0.5625 |
| 2 | 18221000 | 28942000 | 4 | 18932000 | 30205000 | 10721000 | 11273000 | 283 | 454 | 0.6233480176 |
| 2 | 8987000 | 5298000 | 6 | 19320000 | 24581000 | 3689000 | 5261000 | 83 | 96 | 0.8645833333 |
| 2 | 34255000 | 30292000 | 6 | 3800000 | 9371000 | 3262000 | 3364000 | 169 | 118 | 1.4322033898 |
| 3 | 17408000 | 14167000 | 7 | 24632000 | 26818000 | 3241000 | 2186000 | 60 | 77 | 0.7792207792 |
| 3 | 30982000 | 34302000 | 7 | 469000 | 7548000 | 3320000 | 7079000 | 142 | 133 | 1.0676691729 |
| 3 | 22621000 | 27011000 | 12 | 24020000 | 27265000 | 4390000 | 3245000 | 88 | 81 | 1.0864197531 |
| 3 | 711000 | 2038000 | 10 | 19569000 | 20968000 | 1327000 | 1399000 | 74 | 40 | 1.85 |
| 3 | 4829000 | 2121000 | 10 | 11277000 | 16900000 | 2708000 | 5623000 | 137 | 48 | 2.8541666667 |
| 3 | 13205000 | 9773000 | 7 | 27924000 | 29604000 | 3432000 | 1680000 | 168 | 56 | 3 |
| 8 | 27169000 | 28082000 | 9 | 20367000 | 22822000 | 913000 | 2455000 | 28 | 152 | 0.1842105263 |
| 8 | 18880000 | 23533000 | 9 | 11795000 | 17705000 | 4653000 | 5910000 | 85 | 148 | 0.5743243243 |
| 8 | 23801000 | 26429000 | 9 | 18012000 | 19680000 | 3113000 | 1988000 | 91 | 49 | 1.8571428571 |
| 11 | 42000 | 4889000 | 12 | 42000 | 4189000 | 4847000 | 4147000 | 119 | 123 | 0.9674796748 |

## Table S5

**Average # of maize genes whose best hit is a given sorghum gene:** 1.996
**95% confidence interval:** 1.175 maize genes/sorghum gene – 4.025 maize genes per sorghum gene

| Maize Subgenome of Gap | Sorghum Chr | Length of Gap (Sorghum Genes) | First Sorghum Gene In Gap | Last Sorghum Gene In Gap | Maize best hit | Significant? | GEvo Link to |
|---|---|---|---|---|---|---|---|
| 1 | Chr02 | 68 | Sb02g011390 | Sb02g012620 | 1.8529411765 | No | |
| 1 | Chr05 | 40 | Sb05g019510 | Sb05g019950 | 1.725 | No | |
| 1 | Chr06 | 54 | Sb06g002240 | Sb06g003190 | 1.0740740741 | Yes | http://genomev |
| 1 | Chr10 | 60 | Sb10g021590 | Sb10g022110 | 1.45 | No | |
| 1 | Chr10 | 55 | Sb10g022270 | Sb10g022870 | 2.2181818182 | No | |
| 2 | Chr01 | 41 | Sb01g030730 | Sb01g031140 | 1.1951219512 | No | |
| 2 | Chr02 | 56 | Sb02g008970 | Sb02g009530 | 1.0535714286 | Yes | real? |
| 2 | Chr03 | 42 | Sb03g022950 | Sb03g024570 | 2.0714285714 | No | |
| 2 | Chr04 | 70 | Sb04g018290 | Sb04g020210 | 1.5857142857 | No | |
| 2 | Chr04 | 55 | Sb04g021140 | Sb04g021690 | 1.6 | No | |
| 2 | Chr05 | 163 | Sb05g023210 | Sb05g025250 | 1.5214723926 | No | |
| 2 | Chr05 | 65 | Sb05g025610 | Sb05g026240 | 1.8615384615 | No | |
| 2 | Chr07 | 46 | Sb07g006400 | Sb07g007450 | 0.7173913043 | Yes | http://genomev |
| 2 | Chr07 | 48 | Sb07g015450 | Sb07g019260 | 2.2916666667 | No | |
| 2 | Chr07 | 43 | Sb07g023010 | Sb07g023430 | 2.2790697674 | No | |
| 2 | Chr09 | 56 | Sb09g014000 | Sb09g016880 | 2.6607142857 | No | |

**Table S6**

| Rice Orthologs of Sorghum Genes In Deleted Region | GO Terms |
|---|---|
| Os07g17370 | |
| Os07g17400 | GO:0005488;G |
| Os07g17460 | |
| Os07g17490 | |
| Os07g17520 | |
| Os07g17680 | |
| Os07g18070 | |
| Os07g18120 | GO:0003824 |
| Os07g18230 | GO:0000166;G |
| Os07g18240 | GO:0004872;G |
| Os07g18510 | |
| Os07g18600 | |
| Os07g18710 | |
| Os07g18720 | GO:0009536;G |
| Os07g18750 | GO:0006810;G |
| Os07g18874 | GO:0005215;G |
| Os07g18990 | |
| Os07g19000 | |
| Os07g19030 | GO:0005739;G |
| Os07g19040 | GO:0016787 |
| Os07g19160 | |
| Os07g19210 | GO:0003824;G |
| Os07g19390 | |
| Os07g19400 | GO:0005739 |
| Os07g19444 | GO:0009536;G |
| Os07g19460 | GO:0005215;G |
| Os07g19470 | |
| Os07g19494 | |
| Os07g19530 | |
| Os07g20270 | |
| Os07g20290 | GO:0016020 |
| Os07g20340 | GO:0005488;G |

## Chapter 5: Escape from preferential retention following repeated whole genome duplication in plants.

The following chapter (excluding the preface) has been published as a peer reviewed article in the journal Frontiers in Plant Science (Frontiers in Plant Genetics and Genomics section):

Schnable JC, Wang X, Pires JC, Freeling M. (2012) "Escape from preferential retention following repeated whole genome duplication in plants." Frontiers in Plant Science doi: 10.3389/fpls.2012.00094



Contributions:

J. Chris Pires and Xiaowu Wang provided pre-publication access to the genome of *Brassica rapa.*

**Preface:**

In Chapters 2 & 3 I had outlined and tested a model for explaining how difference in expression between parental subgenomes in a polyploid could provide an explanation for why gene loss was unequal between duplicated regions of the genome – and indeed between whole parental subgenomes. However, in all the ancient whole genome duplications we examined in plants between one and three thousand gene pairs remain retained as duplicate copies. This sample included events 70-100 million years old (the pre-grass duplication in rice and sorghum and the core eudicot hexaploidy in grape vine). Previous work from four different research groups had shown that certain classes of genes were significantly enriched among these retained gene pairs including transcription factors, ribosomal subunits and other genes encoding subunits of multi-protein complexes like proteosomal subunits. This bias was explained by the Gene Dosage Hypothesis which predicted that genes encoding products which interacted in ways sensitive to stoichiometry would be resistant to independent duplication but resistant to loss of duplicate copies when all genes were duplicated simultaneously as occurs during a whole genome duplication.

This model did a good job of explaining the data on the types of genes retained following individual whole genome duplications. However it also raises another question. It was now clear that many flowering plants were the result of many sequential rounds of whole genome duplication. The arabidopsis genome has duplicated at least a cumulative 48-fold relative to the common ancestor of all vascular plants and the maize lineage has experienced as 64-fold duplication over the same time frame (see Figure 1 of this

chapter). If certain classes of genes can never be reduced by to a single copy after being duplicated following whole genome duplications, why haven't these dose sensitive, deletion resistant genes taken over the entire genome?

In the following chapter I propose an answer to this question using the key insight from Chapter 2 that duplicated genes are not functionally equivalent. I used two parallel systems:

1) The crucifers with the alpha tetraploidy as the ancient whole genome duplication and *Brassica rapa* as the more recent “reporter” polyploid.
2) The grasses with the pregrass tetraploidy as the ancient whole genome duplication and maize as the more recent “reporter” polyploid.

I first showed that, as previously reported, the tendency to retain duplicate copies following whole genome duplications is heritable. Genes retained as duplicates in one whole genome duplication are more likely to also be retained as duplicates in subsequent whole genome duplications, which was completely consistent with the Gene Dosage Hypothesis's predictions. However, by incorporating expression data in multiple species, along with comparative datasets such as strength of purifying selection (as measured by the ratio of synonymous to nonsynoymous substitution between orthologous genes) it was possible to show that for most gene pairs there was clearly a dominant and non-dominant gene copy. The non-dominant gene copy was less expressed and under less purifying selection, and was significantly more likely to experience the deletion of one (or more) duplicate copies following a subsequent whole genome duplication. This paper showed that genome dominance provided an explanation for how dose sensitive genes could escape the ratchet of repeated retention of duplicates and as a result hadn't overwhelmed flowering plant genomes even in species like *Brassica rapa* which had experienced a cumulative 144-fold genome duplication.

**Introduction:**

Plants have been colorfully labeled the “big kahona of polyploidization” (Sémon and Wolfe 2007b). The lineages leading to the two preeminent models plant genetics -- arabidopsis (a eudicot) and maize (a monocot) -- each show evidence of multiple independent whole genome duplications (Fig 1) since monocots and eudicots diverged approximately 120 million years ago (Soltis *et al.* 2009). Recent evidence suggests at least two additional, shared, whole genome duplications prior to the monocot/eudicot split (Jiao *et al.* 2011). The cumulative ploidy numbers relative to a pre-seed plant ancestor are listed in parentheses in Figure 1. Whole genome duplication creates duplicate, potentially redundant, copies of all the genes within a genome. The loss of these duplicate copies from the genomes of ancient polyploid species is known as fractionation (Langham *et al.* 2004) and -- over evolutionary time scales -- the majority of genes duplicated by polyploidy will be reduced back to a single copy. If fractionation did not occur, an ancestral genome of 10,000 genes would grow to an unrealistically large 640,000 genes in maize, and 1.44 million genes in *Brassica rapa*.

Some classes of genes, particularly those encoding organelle, preferentially revert to single copy status following whole genome duplications (Duarte *et al.* 2010). However, other classes of genes -- such as subunits of large multiprotein complexes, transcription factors, and signal transduction machinery tend to resist fractionation following whole genome duplication (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Maere *et al.* 2005). This observation has been explained by the Gene Dosage Hypothesis (Birchler and Veitia 2007) which predicts that fractionation of genes encoding proteins involved in dose sensitive interactions will be selected against, as the loss of either gene copy is expected to throw the dosage of that gene pair's product out of balance with its interaction partners, partners that also tend to remain duplicated. The topic of the influence of gene dosage constraints on post-tetraploidy genome evolution has been well-reviewed (Sémon and Wolfe 2007b; Freeling 2009; Edger and Pires 2009; Birchler and Veitia 2010). A previous study of multiple sequential tetraploidies the arabidopsis lineage found a general tendency for genes retained following one tetraploidy to also be retained following a second one (Seoighe and Gehring 2004).

Since the divergence of the arabidopsis and grape lineages, arabidopsis has experienced two additional rounds of whole genome duplication. The rate of duplicate gene retention for transcription factors after single polyploidies have been observed to be approximately 25% (Blanc and Wolfe 2004; Seoighe and Gehring 2004). If no mitigation of gene dosage occurred, our expectation after two rounds of whole genome duplication is that arabidopsis should contain approximately 156% as many transcription factor encoding genes as grape. However, a detailed annotation of transcription factors using conserved protein domains found the number of transcription factors in the arabidopsis genome is only 25.4% greater than the number found in grape (Lang *et al.* 2010). The fitness cost of

changes in relative gene dosage must, to some extent, be mitigated over multiple whole genome duplications or the genomes of plants would long ago have become overburdened with genes encoding life's most complicated machines.

This paper provides evidence that duplicate genes do not equally maintain their progenitor's preference for duplicate gene retention. Duplicate genes produced by whole genome duplication are not equivalent. Parental genomes originating from different species within a polyploid almost immediately differentiate into into dominant and nondominant subgenomes (Chang *et al.* 2010), and these expression differences are preserved for millions of years (Flagel and Wendel 2010; Schnable *et al.* 2011b). Bias in gene loss between duplicate regions (fractionation bias) has been observed in arabidopsis (Thomas *et al.* 2006) and maize (Woodhouse *et al.* 2010) and seems to be a general rule for whole genome duplications ranging from paramecium to fish (Sankoff *et al.* 2010). Bias in fractionation and genome dominance are linked because it is expected that genes on the underexpressed, nondominant subgenome simply matter less to purifying selection and dosage-constraints (Schnable *et al.* 2011b). In maize, genes with known mutant phenotypes are indeed preferentially found on the dominant subgenome (Schnable and Freeling 2011). As bias in expression predicts which subgenome will experience more fractionation following polyploidy, either subgenome identity or the expression patterns of individual gene pairs may also predict which copy of a duplicate gene pair will be more prone to duplicate gene retention in future polyploidies.

We addressed the issue of mitigation of gene dosage constraints with two experimental systems, the grasses and the crucifers. Both clades have roughly parallel histories of polyploidy among species with sequenced genomes (Fig. 1). Both grasses and crucifers contain a more ancient whole genome duplication which is shared by all sequenced species in the clade (Bowers *et al.* 2003; Paterson *et al.* 2004) and in both clades one well studied species with a sequenced genome has experienced a second subsequent whole genome duplication – maize in the grasses (Gaut and Doebley 1997) and *Brassica rapa* in the crucifers (Lysak *et al.* 2005). In both cases any duplicate genes retained from the older clade-wide polyploidy did not retain additional duplicate copies in the subsequent lineage specific polyploidy. Therefore we were able to carry out parallel experiments to identify characteristics associated with preferential retention. It was possible to control, to some extent for the effect of protein function, by focusing on pairs of duplicate genes retained in the clade-wide polyploidy which had different fates in the subsequent lineage-specific polyploidy. A model is proposed to explain how the duplicate copies of dose sensitive genes escape preferential retention in later polyploidies.

**Methods:**

**Data sources**

The genome assemblies and annotation used in this study were TAIR 10 (*Arabidopsis*

*thaliana*) , *Arabidopsis lyrata* v1.0 (Hu *et al.* 2011), the initial release of the *Brassica rapa* genome (The Brassica rapa Genome Sequencing Project Consortium 2011), MSU 6 (*Oryza sativa*) (Goff *et al.* 2002), *Sorghum bicolor* 1.4 (Paterson *et al.* 2009), and B73_refgen1 (*Zea mays*) (Schnable *et al.* 2009).

**Gene pair identification**
Orthologous genes between *Arabidopsis thaliana* and *Arabidopsis lyrata* were identified using SynMap (Lyons *et al.* 2008b) with QuotaAlign settings of 1:1 (Tang *et al.* 2011). Arabidopsis-Brassica orthologous relationships were taken from (Tang *et al.* 2012). All orthologous and homeologous relationships between grass species are those published in (Schnable *et al.* 2012a).

**Expression calculations**
Gene expression levels were calculated using previously published RNA-seq data from wild type seedlings of *Arabidopsis thaliana* (SRX019140: 44.7 million reads (Deng *et al.* 2010)) and rice (SRX020118: 8.9 million reads (Zemach *et al.* 2010)). These datasets were selected because, at the time these analysis were originally conducted they represented the RNA-seq experiments with the most sequencing depth for these two species deposited in the sequence read archive. Reads were aligned to reference genomes using Bowtie (Langmead *et al.* 2009) and gene expression levels were quantified using Cufflinks (Trapnell *et al.* 2010). Bowtie does not perform spliced alignments, which means some reads from regions of mRNA molecules which span exon junctions were not recovered in our analysis. However, given that homeologous genes will in almost all cases posses the same intron-exon structure, any bias introduced by this approach will be equivalent between gene copies.

**Measuring Purifying Selection**
Synonymous and nonsynonymous substitution rates were calculated using the synonymous_calculation package included with bio-pipeline (https://github.com/tanghaibao/bio-pipeline/) using the Nei-Gojobori method (Nei and Gojobori 1986) All other settings remained as default.

**Identification of Rice CNSs**
Rice CNSs were identified using version 3 of the CNS Discovery pipeline (https://github.com/gturco/find_cns) (Schnable *et al.* 2011a).

**Statistics**
P-values for the the difference in retention frequencies between singleton genes and homeologously paired genes were calculated using Fisher's Exact Test. In the crucifers, arabidopsis genes with two or three retained co-orthologs in *Brassica rapa* were grouped together as “retained.”

**Results:**

Genes syntenically conserved through the crucifiers or grasses were categorized as 1) those without a homeologous duplicate from the older polyploidy in each lineage 2) those with a retained homeolog from the older polyploidy in each lineage. In the crucifer lineage, the older tetraploidy is arabidopsis lineage alpha (23-40 MYA); in the Poales, the earlier tetraploidy was 'pre-grass'' (about 70 MYA) (Fig, 1). In crucifers, these genes are classified by the number of co-orthologs conserved in Brassica rapa after the hexaploidy shared by all *Brassica* species (Fig. 2A). In grasses, genes were classified by whether maize retained only one or both co-orthologs following the more recent tetraploidy of the Zea/Tripsacum lineage (Fig 2B). Retention in older polyploidies does predict retention in future polyploidies ($p < 2.2 *10^{-16}$ for both crucifers and grasses), as previously showing in arabidopsis (Seoighe and Gehring 2004). However in both experiments approximately half of genes previously retained as a duplicate pair in the older whole genome duplication -- and therefore presumed to be sensitive to changes in gene dosage – fractionated to a single copy in the more recent whole genome duplication.

The crucifer dataset consisted of 817 arabidopsis gene pairs where one copy was orthologous to only a single gene in *Brassica rapa* and the other possessed either two or three co-orthologs (Supplemental data S1). The grass dataset consisted of 407 gene pairs conserved in both rice and sorghum where one copy was orthologous to only a single gene in maize, its duplicate having been fractionated and the other represented by two co-orthologs in maize (Supplemental data S2). Gene pairs result from more ancient whole genome duplications were identified and removed, as these tend to introduce confounding factors. Members of gene pairs were assigned to under and over fractionated subgenomes using differences in the number of genes syntenically retained in multiple species between homeologous regions of the rice and arabidopsis genomes (Schnable *et al.* 2011b, 2012a). In both datasets, the analysis of the relative levels of RNA encoded by duplicate genes pairs -- measured by RNA-seq – was carried out in an outgroup lineage which shared only the older clade-wide polyploidy. In the grasses we used the expression of syntenic orthologs in rice and in the crucifers syntenic orthologs in Arabidopsis thaliana (see Methods). The relative levels of purifying selection acting on each members of a gene pair were also compared using the ratio of nonsynonymous substitutions to synonymous substitutions between orthologous genes in *Arabidopsis thaliana* and *Arabidopsis lyrata* (for the crucifers) and between rice and sorghum (for the grasses) (see Methods). Promoter complexity, as measured by number of conserved noncoding sequences, has previously shown to influence the odds a gene will be retained as a duplicate pair following polyploidy in the grasses (Schnable *et al.* 2011a) – so gene pairs were also sorted based on number of conserved noncoding sequences, in the grasses, and total quantity of upstream non-transposon sequence in arabidopsis, this length being a crude proxy for promoter complexity having previously been shown to correlate with complexity of gene expression patterns (Sun *et al.* 2010).

All four potential markers examined showed significant power to predict which copy of a homeologous gene pair would be more resistant to fractionation in subsequent whole genome duplications (Figure 3). In general the gene copy retained in duplicate tended to also be the higher expressed copy, show evidence of greater purifying selection and to be associated with greater amounts of noncoding regulatory sequence. These genes also tended to be located on the dominant subgenome.

**Discussion:**

Following polyploidy, a genome possesses two or more homeologous genes, each with the same coding sequence and regulatory elements. Yet these gene copies can immediately show very different patterns of expression (Flagel *et al.* 2008; Buggs *et al.* 2011). It has been proposed that the deletion of less expressed copy of a gene following polyploidy is more likely to be be selectively neutral (Schnable and Freeling 2011; Schnable *et al.* 2011b). When combined with the observation that expression levels are unequal between parental subgenomes in allotetraploids (Chang *et al.* 2010; Flagel and Wendel 2010; Schnable *et al.* 2011b), this model may explain the bias fractionation bias which has been found in ancient polyploids species (Schnable *et al.* 2011b).

Here we have shown that that the dominant gene copy -- more expressed, under higher purifying selection, associated with more regulatory sequence – of a homeologous gene pair is more likely to retain the ancestral characteristic of preferential retention of duplicate copies in subsequent polyploidies. A number of explanations could be proposed for the link between expression and future resistance to fractionation. We propose a model based on the same link between expression and which predicts fractionation bias between parental subgenomes. If all the co-orthologs of a single ancestral gene contribute to a single pool of gene product, the loss of less expressed gene copies would result in the smallest change in total gene product dosage. If the total expression of a group of homeologous genes is constrained in either relative or absolute terms (Bekaert *et al.* 2011) smaller changes in total gene product dosage -- created by the loss of a less expressed gene copy -- are predicted to be more often selectively neutral, and therefore more common (Figure 4). This model also predicts that, for gene pairs in *Arabidopsis thaliana* where only one copy possesses any orthologous genes in *Brassica rapa*, it should more often be the more expressed copy; as is indeed the case (Table S1).

When combined with previous results linking genome dominance with biased fractionation (Chang *et al.* 2010; Schnable *et al.* 2011b), our results suggest he Gene Dosage Hypothesis could perhaps be better thought of as the Gene-Product Dosage Hypothesis in that it can generally be considered to act on the concentration of the proteins encoded by duplicate genes, not gene copy number itself. Even when both copies of a gene are retained following whole genome duplication, the less expressed copy will

often be lost in subsequent whole genome duplications. Furthermore, the greater the number of duplicate copies of a gene are found within a genome the less each individual copy contributes to total expression and the more likely it becomes that the loss of individual copies can be tolerated. In other words, the protection against fractionation provided by selection for gene dosage – either absolute or relative -- becomes less powerful the less a give gene copy contributes to total expression, and the more total gene copies are present within the genome. This explains, at least in part, why despite being the “big kahuna” of whole genome duplications plant genomes are not over-burdened with fractionation resistant gene families.

## Figures

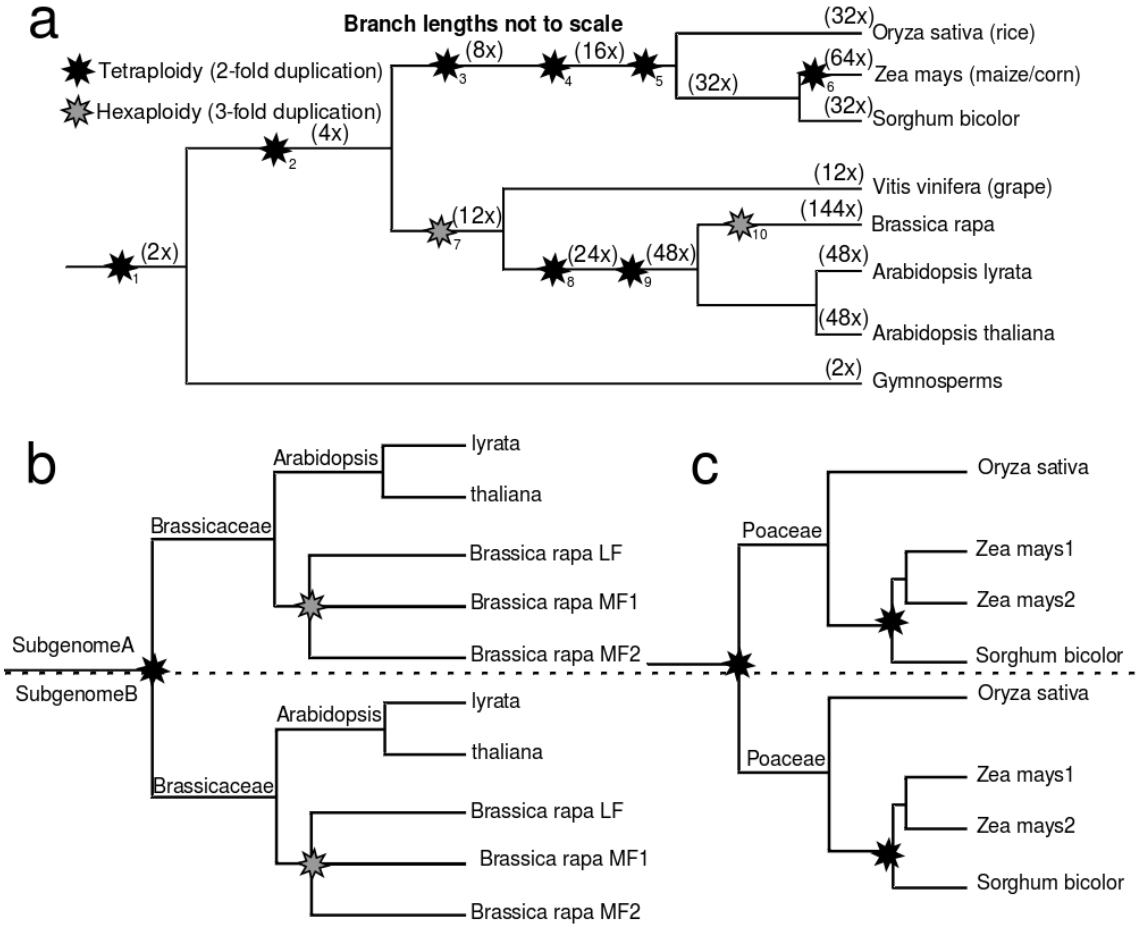



Figure 1
Phylogenetic trees showing the distribution of whole genome duplications throughout the flowering plants. A) Relationships of all species and whole genome duplications referenced in this paper. Ploidy levels relative to the common ancestor of seed plants are noted along individual branches in parentheses. Footnotes mark individual ancient whole genome duplications, described in more detail in Table 1. B) The relationship between the multiple subgenomes created by whole genome duplications within the crucifers. The LF, MF1, MF2 terminology for brassica subgenomes comes from (The Brassica rapa Genome Sequencing Project Consortium 2011). C) The relationship between the multiple subgenomes created by whole genome duplication within the grasses. The Maize1, Maize2 terminology for maize subgenomes comes from (Schnable *et al.* 2011b).

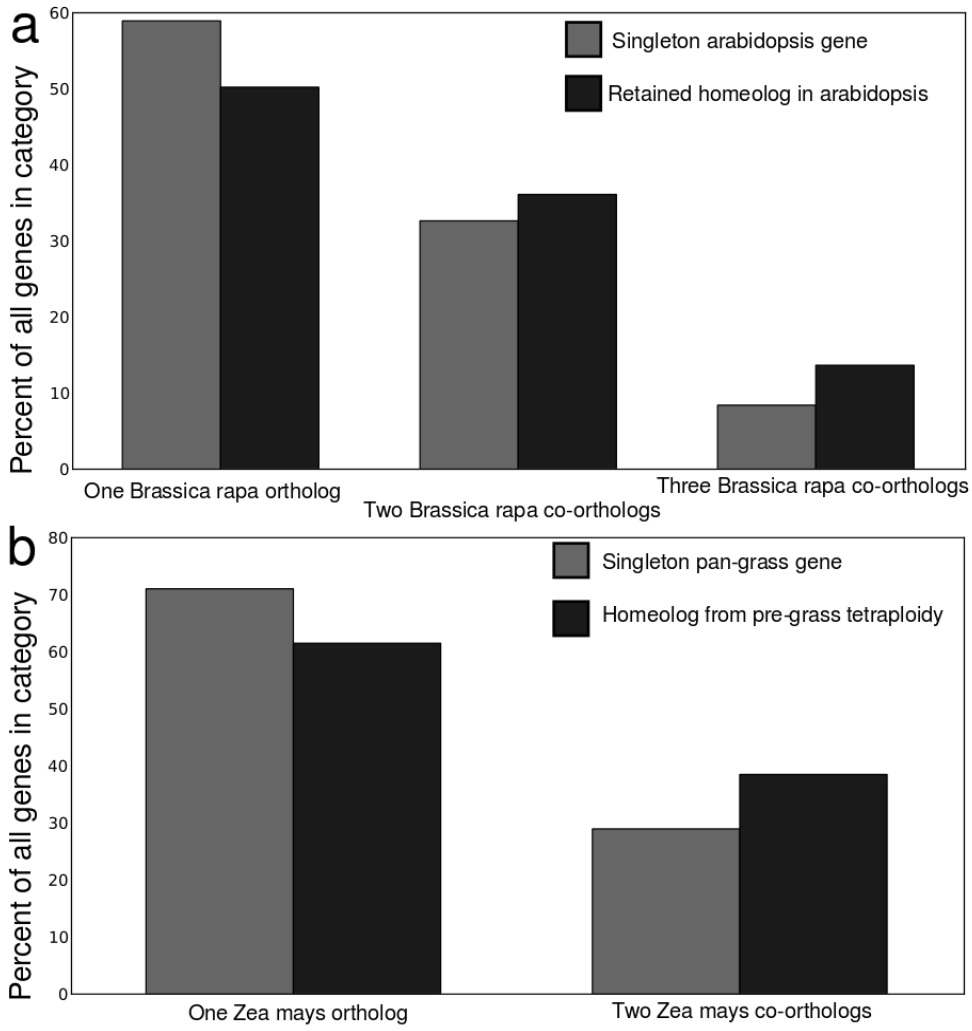



Figure 2
Impact of retention in a previous whole genome duplication on retention in future whole genome duplications. A) Proportion of arabidopsis genes with one, two, or three co-orthologs in *Brassica rapa.* B) Proportion of genes syntenically retained in sorghum and rice with one, or two co-orthologs in maize.

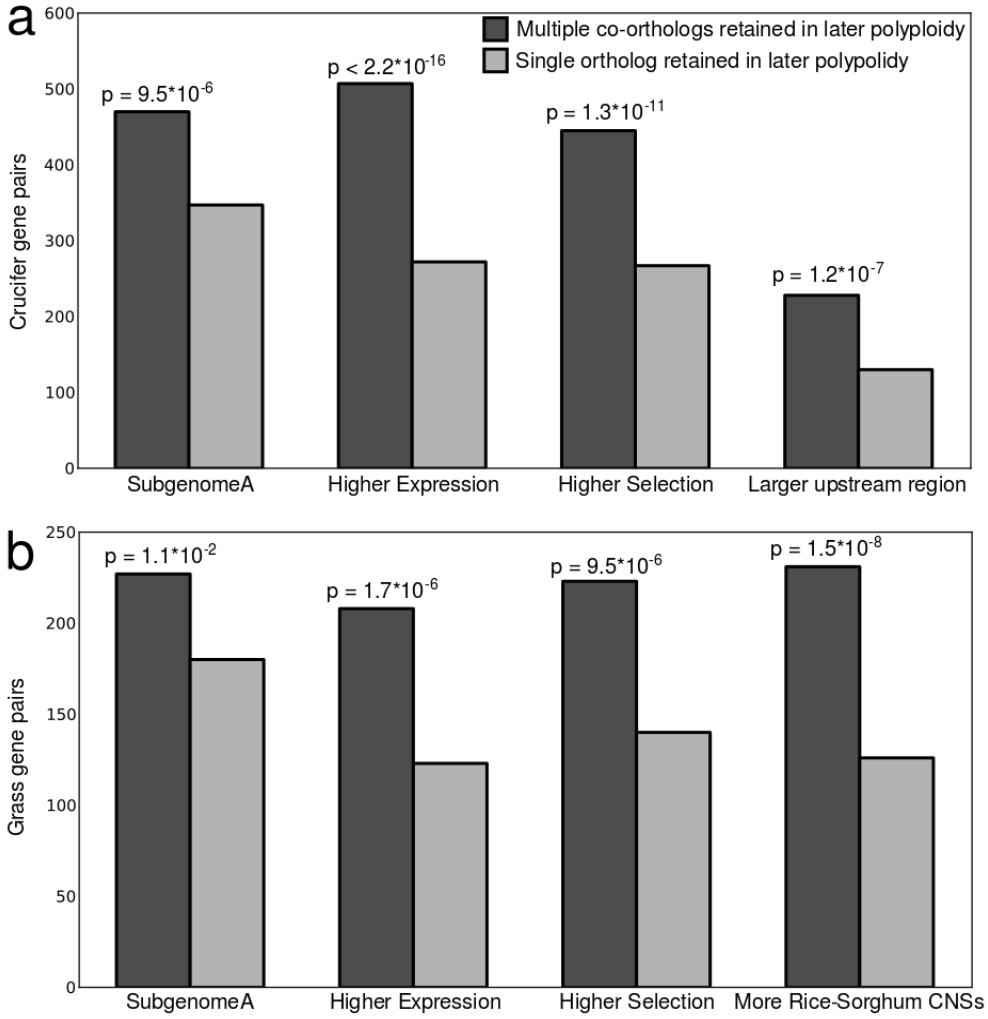



Figure 3
Correlation between subsequent duplicate gene retention and a number of predicting factors including gene expression, ratio of non-synonymous to synonymous substitutions, and subgenome identity for a) crucifer and b) grass gene pairs. P-values relative to a 50/50 binomial distribution.

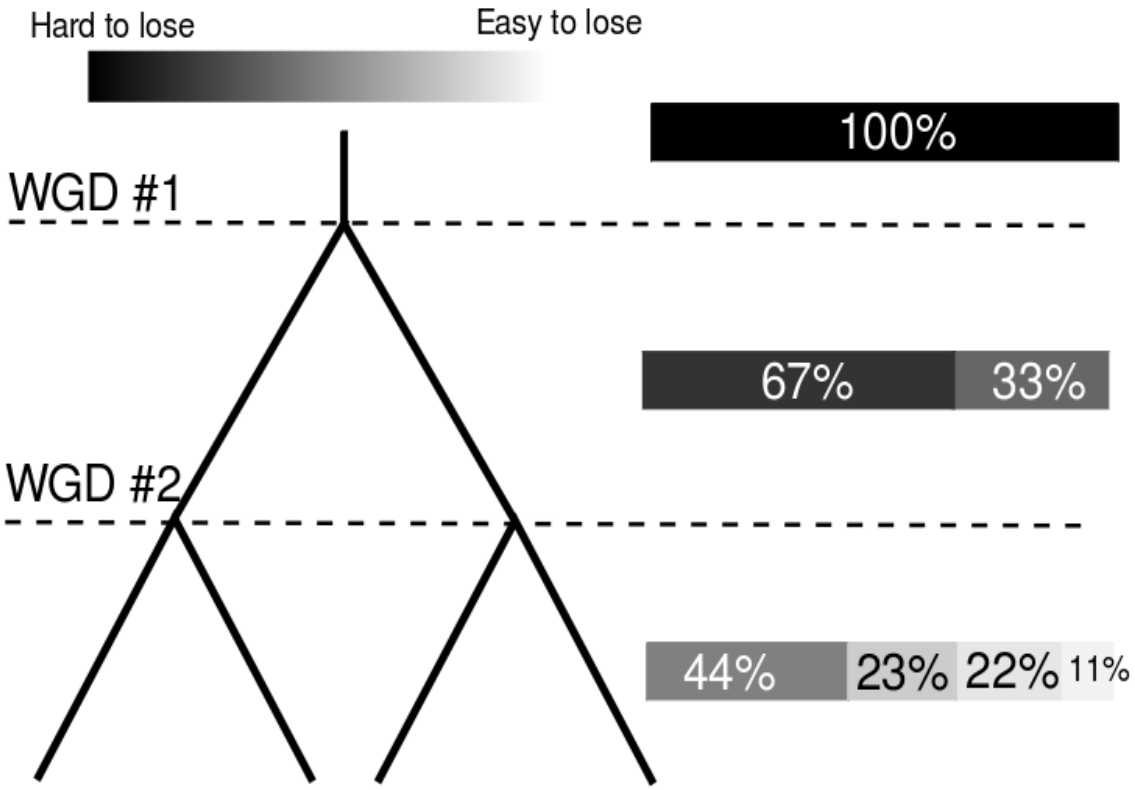



Figure 4
Model for the intrinsic mitigation of gene dosage following multiple rounds of whole genome duplication. On the left, the phylogenetic tree of a perfectly retained gene after two rounds of whole genome duplication. On the right a model of how total expression is partitioned among increasing numbers of gene copies assuming genome dominance. Darkness of individual bars indicates how large an effect the loss of individual gene copies will have on total expression, and, presumably, on fitness.

**Tables**

Table 1: Whole Genome Duplications

| Footnote ID from Figure 1 | One name (often of many) | One citation (often of many) |
|---|---|---|
| 1 | Pre-seed plant | (Jiao *et al.* 2011) |
| 2 | Pre-flowering plant | (Jiao *et al.* 2011) |
| 3 | Sigma1 | (Tang *et al.* 2010) |
| 4 | Sigma2 | (Tang et al., 2010) |
| 5 | Pre-grass/Rho | (Paterson *et al.* 2004) |
| 6 | Maize Lineage WGD | (Gaut and Doebley 1997) |
| 7 | Gamma/Pre-eudicot hexaploidy | (Jaillon *et al.* 2007) |
| 8 | Beta | (Bowers *et al.* 2003) |
| 9 | Alpha | (Bowers et al., 2003) |
| 10 | Brassica hexaploidy | |

Table 2 Completely Deleted Gene Copies in Brassica rapa

| | Less Expressed Copy Lost in Brassica rapa | More Expressed Copy Lost in Brassica rapa | P-value |
|---|---|---|---|
| All alpha pairs where one copy has been completely lost in Brassica rapa | 428 gene pairs | 217 gene pairs | p = 3.60*10^-17 |
| Alpha pairs where there are multiple co-orthologs in Brassica rapa of the retained copy | 271 gene pairs | 98 gene pairs | p = 3.48*10^-20 |
| Both copies expressed above 5 FPKM in Arabidopsis thaliana | 191 gene pairs | 128 gene pairs | p = 2.49*10^-4 |

## Chapter 6: Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses.

The following chapter (excluding the preface) has been published as a peer reviewed article in the journal Frontiers in Plant Science (Frontiers in Plant Genetics and Genomics section):

Schnable JC, Pedersen BS, Subramaniam S, Freeling M. (2011) "Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses." Frontiers in Plant Science DOI: 10.3389/fpls.2011.00002



Contributions:

The set of rice-sorghum conserved noncoding sequences using in this analysis were generated by Brent Pedersen using the CNS Discovery Pipeline 1.0 software package he developed. Sabarinath Subramaniam conducted the analysis of global correlation between CNS count and odds of possessing a retained homeolog displayed in Figure 3.

**Preface:**

When gene expression was previously discussed in this document (mostly in Chapters 2 and 5), the focus was on the global levels of expression. Data often came from large aggregates of tissues and organs such as “Whole Seedlings.” However gene expression obviously is a more complex trait than can be captured by a single number. Regulatory sequences which do not themselves code for protein determine in which cells and in what quantities the mRNA coded for by specific proteins will be produced.

Regulatory sequences can often be identified as regions near homologous protein coding genes which show an unexpectedly high level of sequence similarity, because these noncoding regulatory sequences, like protein coding exons, are functionally constrained. These conserved noncoding sequences are distributed unevenly with some genes – often transcription factors and genes annotated as responding to environmental stresses -- associated with many large conserved noncoding sequences and other genes, particularly those involved in housekeeping functions with constant global levels of expression, showing little or no sign of conserved promoter sequences.

This chapter addresses the link between genes associated with large numbers of conserved noncoding sequences and greater retention of duplicate copies following whole genome duplication. There are many parallels between the approach used here to test the effect of conserved noncoding sequence richness on retention and the method employed

in Chapter 5 to examine the effects of expression level and strength of purifying selection, however only the grasses were used in this analysis as this research was conducted before the publication of the *Brassica rapa* genome so the crucifers were not yet a viable parallel research system.

This Chapter is also the transition between studying genes on a whole plant level and attempting to understand the regulation of gene expression levels in individual organs, tissues, and cell types. The conserved noncoding sequences introduced here will form the foundation for the research approach described in Chapter 8.

**Introduction**

It was almost half a century ago that Susumo Ohno first proposed a role for whole genome duplications in the evolution of vertebrates (Ohno 1970) just as E.B. Lewis did for duplications of individual genes two decades before Ohno (Lewis 1951). While the most recent tetraploidy in the lineage leading to humans is estimated to be half a billion years old (Kasahara 2007), both modern and ancient whole genome duplications are abundant in flowering plants. An estimated 35% of flowering plants are polyploid relative to the baseline level for their genera (Wood *et al.* 2009). Arabidopsis thaliana – a species selected for its small genome – contains readily detectable evidence of two rounds of whole genome duplication within its order and a more ancient hexaploidy, all estimated to have occurred within the last 120 million years (Bowers *et al.* 2003; Maere *et al.* 2005; Paterson *et al.* 2010).

Whole genome duplications create two copies of every gene and all associated regulatory sequences. These duplicate genes and chromosomal segments are referred to has homeologs and homeologous throughout this paper. However they are known variously throughout the literature as ohnologs, homoeologs, or syntenic paralogs. In most cases, one of the two homeologs, each now potentially redundant, is lost by fractionation. In maize the mechanism of fractionation was shown to involve short deletions by non-homologous recombination (Woodhouse *et al.* 2010). Although duplicated regions are initially identical or near-identical, gene loss data from all studied tetraploidies show clear bias between duplicate chromosomal segments with one region sustaining the majority of gene copy deletion (Thomas *et al.* 2006; Sankoff *et al.* 2010; Woodhouse *et al.* 2010). This bias remains consistent across each pair of paleochromosomes in maize and is paralleled by differences in expression levels of duplicate genes located on homeologous paleochromosomes (Schnable *et al.* 2011b).

While duplicate copies of many genes are lost following whole genome duplication, in some cases both copies of a gene are retained. It was initially thought that these cases were consequences of sub- or neofunctionalization. However, most researchers now embrace an entirely different explanation: duplicate genes are retained following whole-genome duplication in cases where loss generates imbalance in dosage sensitive interactions of the products of those genes with other proteins encoding by duplicated genes. This explanation, a corollary of the Gene Dosage Hypothesis (Birchler *et al.* 2005; Veitia *et al.* 2008), is a powerful tool for explaining many observations regarding genes retained as duplicate copies following whole genome duplication (reviews: (Birchler *et al.* 2007; Sémon and Wolfe 2007b; Freeling 2009)). Genes involved in forming multi-protein complexes – such as the proteasome core, ribosome components and, molecular motors – are some of the most enriched in retained duplicate copies following whole genome duplication, and any gene annotated with the molecular function GO0003700, “transcription factor activity” is particularly likely to have been retained after the most recent tetraploidy in arabidopsis (review: (Freeling 2009)). An inverse

relationship has been found between genes that form local duplicates, a process that disrupts gene dosage, and genes that are retained following tetraploidy (Cannon *et al.* 2004; Freeling 2009). Subfunctionalization cannot explain this result as both forms of duplication represent sources of potentially subfunctionizable genes, however the result is consistent with selection to maintain the relative dosage between many genes.

Genes encoding transcription factors are not typical genes. The gene dosage hypothesis is generally discussed as applying to interactions between or among gene products. There is no reason why protein-DNA interactions, such as those between a transcription factor and its binding site, might not also be subject to dosage constraints. Known transcription factor binding sites tend to be short and are represented at many sites throughout the genome. Only a small fraction of these are biologically relevant (as reviewed (Wray *et al.* 2003)); even in prokaryotes, finding functional motifs computational is extraordinarily challenging (Salama and Stekel 2010). Rather than attempt to predict which binding sites are functionally relevant *ab initio*, it is possible to use comparative genomics to discover which noncoding sequences surrounding a gene are likely to function. Functional regions are expected show lower base pair substitution rates than functionless sequences. Data in animals (Miller *et al.* 2004) and plants (Freeling and Subramaniam 2009) support this. By comparing the noncoding sequence surrounding orthologous or homeologous plant genes, we can identify conserved regions termed conserved noncoding sequences (CNSs) a procedure sometimes referred to as "phylogenetic footprinting." Previous studies comparing orthologous genes between maize and rice (Inada *et al.* 2003) and homeologous duplicated genes in arabidopsis (Thomas *et al.* 2006) found that genes with many associated CNSs tend to encode transcription factors, particularly those expressed in response to external stimuli. Very CNS-rich genes have been called "bigfoot genes" (Thomas *et al.* 2006).

Identification of CNSs requires comparing pairs of orthologous – diverged by speciation – or homeologous – diverged by whole genome duplication – genes within a critical window of sequence divergence. Noncoding sequences surrounding recently diverged genes will show sequence conservation even in the absence of purifying selection for function, while functional noncoding sequences will sometimes fall below the limits of detectability, especially if the divergence times are too great. No species with a sequenced genome is a suitable evolutionary distance from arabidopsis for CNS detection. Therefore, CNSs in arabidopsis were identified by comparing the noncoding sequences surrounding retained homeologous genes (Freeling *et al.* 2007). As a result, all arabidopsis genes with associated CNSs, by definition, were retained as a homeologous pair following the most recent whole genome duplication in the arabidopsis lineage and obviously do not represent a useful system for studying any possible correlation between CNS content and retainability.

The grasses provide a model system in which to test our question: Does CNSs-richness

correlate with an increased tendency to have both duplicate copies retained following a whole genome duplication? In other words, are some genes retained following tetraploidy, not because their protein products are involved in dosage sensitive interactions, but because their own cis-regulatory sequences (promoters, enhancers, locus control regions, insulators, etc) are the target of dosage sensitive transcription factors? The genomes of all grass species studied to date contain a core gene set that is maintained in a well-conserved syntenic order (Bennetzen and Freeling 1993; Moore *et al.* 1995) making the identification of true orthologs and homeologs, as well as the predicted locations of deleted genes, possible. The pre-grass lineage experienced a whole genome duplication an estimated 50-70 million years ago (Vandepoele *et al.* 2003; Paterson *et al.* 2004; Yu *et al.* 2005). The grasses have since radiated into a few deep tribal lineages, three of which are represented by at least one species with a published genome sequence (Figure 1). The first plant CNSs described were identified by comparing orthologous rice and maize genes (Kaplinsky *et al.* 2002; Inada *et al.* 2003; Guo and Moose 2003). Sorghum and rice share the same divergence as rice and maize and are ideally spaced for the discovery of CNSs between orthologous genes. As neither species has experienced a whole genome duplication, the CNS richness of individual genes can be quantified while independently quantifying that gene's history of retention or loss following the whole genome duplication preceding the grass radiation.

The Andropogoneae, a tribe of the grasses, contain two species with sequenced genomes: sorghum and maize. The maize lineage experienced a second whole genome duplication (Gaut and Doebley 1997) contemporaneous with its divergence from the sorghum lineage, while the sorghum lineage has remained unduplicated since the the pre-grass tetraploidy (Swigoňová *et al.* 2004). Ongoing fractionation in the maize genome provides a second dataset to test predictions about dosage-sensitivity made using comparisons of rice-sorghum orthologs and homeologs (Woodhouse *et al.* 2010; Schnable *et al.* 2011b). The phylogenetic relationships of genome segments between rice, sorghum, and maize are summarized in Figure 1. The availability of these grass genome sequences and their relationships allow us to evaluate the role CNSs – and the regulatory sequences they mark – play in gene retention following tetraploidy and, presumably, in dose-sensitivity.

**Results**

**Sorghum-rice CNSs obtained in automated fashion and sorted to their nearest gene**

An automated pipeline compared the genomes of japonica rice and sorghum for orthologous genes (Woodhouse *et al.* 2010). These published methods also include methods for the automated discovery of CNSs. Using these orthologous genes as syntenic anchors, CNSs conserved within, upstream and downstream of orthologous rice and sorghum genes were identified (see Methods and Supplemental Dataset S1) The single most CNS rich gene in the sorghum genome is the *myb* transcription factor gene *Sb01g037110* (Fig. 2). This gene’s noncoding space covers about 30 kb in sorghum, and 70 kb in the longest of the maize homeologs. The GEvo comparison panel shown in

Figure 2 – derived from the CoGe software suite -- is an example of how we check the results of our automated pipeline while tuning the parameters for optimum CNS discovery between different pairs of species. Every pair of rice-sorghum orthologous genes has an associated GEvo link included in Supplemental Dataset S1, allowing any researcher to visually proof the accuracy of our automated CNS identification pipeline.

**CNS counts and retention from the pregrass tetraploidy**

We first asked if genes with greater numbers of associated CNSs were more likely to possess a retained homeologous copy from the pregrass whole genome duplication than genes with fewer or no associated CNSs. Figure 3 reports the percent of genes with a retained pre-grass homeolog in rice, binned by number of associated CNSs. Genes not retained at syntenic locations between rice and sorghum are excluded from the analysis as it is not possible to annotate CNSs for these genes. The data show a rise in the percent of genes with a retained homeologous gene from the pregrass whole genome duplication as the number of associated conserved noncoding sequences increases. This trend is continuous over a range from zero to 15 CNSs. The smallest bin in Figure 3 contains 230 genes (>15 CNSs and 33% retention). Six of the 15 rice-sorghum gene pairs with >28 CNSs possess a retained homeolog (40% retention) and 25 of the 56 gene pairs with 22-28 CNSs possess a retained homeolog (45% retention). There is an obvious positive correlation between CNS-richness and retention of duplicate gene copies post-tetraploidy

There are many gene categories, especially those encoding ancient components like ribosomal proteins or motor proteins—that are significantly over-retained and are conspicuously low in CNSs (Thomas *et al.* 2006). Dose sensitive product-product binding into large heterogenous complexes is certainly adequate to explain many categories of over-retained genes. The large collection of genes encoding transcription factors are, on average, both CNS-rich and over-retained (Freeling 2009). So, not only is our positive correlation of CNS-richness with retention not universal to all gene groups, it is also possible that it is a mere reflection of the fact that transcription factors are both CNS rich and highly retained following tetraploidy and not an effect of CNSs themselves. We attempted a crude experiment to test this trivial explanation.

We asked: For individual transcription factor gene families- each acting in complexes we assume to be of equivalent molecular complexity/connectivity—were CNS-rich genes retained from the pre-grass tetraploidy at a frequency significantly higher than the frequency for homologous CNS-poor genes? From the 1923 entries in the Database of Rice (*Japonica*) Transcription Factors in 2009 (http://drtf.cbi.pku.edu.cn/) we identified families with ≥ 6 members in rice (discounting tandem duplicates and genes not conserved as syntenic orthologs in sorghum). The orthologously paired members of each family were ranked by number of CNSs. If the bin had the minimum number of genes, 6-10, the one most CNS-rich and the one least CNS-rich gene were evaluated for whether or not they had a pre-grass homeolog (i.e. were retained). For families with greater than

the minimum number of genes, the total orthologous pair gene count was divided by 10, and that number was sampled from the most-CNS-rich and the most CNS-poor ends of the distribution. In this way, each transcription factor family data point was weighted by its total sorghum-rice orthologous pair count.

168 CNS-rich TF genes were paired with 168 CNS-poor genes from the same family. Overall 60% of these genes possessed a retained homeolog from the pre-grass tetraploidy. CNS-rich transcription factor genes possessed a retained duplicate copy in 75% of cases while only 45% of the CNS-poor members of the same families possessed retained duplicate copies. This distribution is significantly different from our null hypothesis of 60% retention in both groups of genes with a p-value of .006 (Chi-square test df=1). However, the tenuous nature of our assumption that transcription factors of the same family should, on average, engage in complexes of equivalent complexity precludes any clean conclusion.

**Differential retention of pre-grass homeologs in the subsequent maize tetraploidy**

The addition of the maize genome to the collection of grasses with sequenced genomes, and the second whole genome duplication found in that lineage (Fig. 1), permits a more controlled experiment. An organism possesses two copies of every gene at the moment of whole genome duplication. Even if the whole genome duplication is the result of a wide cross (allotetraploidy) each duplicate copy possesses near-identical regulatory sequence, and encodes a protein with near-identical function that participates in a near-identical set of potentially dose-sensitive interactions within the cell. Specific regulatory sequences may be deleted from the promoters of either gene copy over evolutionary time – likely by the same short deletion mechanism observed to remove duplicate gene copies following the most recent tetraploidy in maize (Woodhouse *et al.* 2010). The expectation is that homeologous gene pairs from the pre-grass duplication will often possess unequal numbers of associated CNSs (Fig. 4). This expectation was met.

Homeologous genes resulting from whole genome duplication start out possessing the same functions and interaction partners; this provided a more precise control for gene function than simply belonging to the same gene family. The behavior of these genes in the subsequent maize whole genome duplication – whether one of the two new duplicates is lost or both are retained – provides a read-out of differences in dose sensitivity which accumulated since the two genes diverged following the pre-grass tetraploidy. Using a dataset of 497 homeologous pairs of genes conserved in both rice and sorghum where the most CNS-rich rice-sorghum gene pair possessed at least 5 CNSs (Supplemental Information S2), we tested whether or not duplicated genes were retained at different rates in a subsequent tetraploidy (maize) when they possessed different numbers of CNSs. We identified the two syntenic orthologous locations in the reduplicated maize genome for each sorghum gene. We then classified each sorghum gene as 1) retained, with orthologous genes present at both orthologous location in the maize genome 2)

fractionated, with an orthologous gene present at one of the two orthologous location in the maize genome, but deleted from the second or 3) completely lost. Data for all 497 gene lineages are reported in Table 1. Genes with more associated CNSs are more likely to be retained as a homeologous pair in maize (282 cases, 56.7%) than their less CNS rich homeologs (217 cases, 43.7%). These numbers are significantly different from the 1:1 ratio (p=.0036 chi-square test, df=1) expected if CNS richness did not impact dose-sensitivity, and are in agreement with our hypothesis that CNS-richness *per se* confers a significantly greater chance of duplicate gene retention following tetraploidy.

**Discussion**

As documented in the Introduction, over-retention of genes (as post-tetraploidy gene pairs) encoding proteins of ribosomes, proteasomes, motors and cell walls certainly make sense in light of dose-sensitive protein-protein interactions. Transcription factor genes encode proteins that sometimes function in complex multiprotein units as well, so perhaps protein-protein interactions explain the over-retention of this very large category of genes. However this is not the only possible explanation. High-level or upstream transcription factors tend to be under tight regulatory control, and the anchor sequences that act in *cis* on such genes are often involved in complex interactions involving proteins and multi-protein complexes; an example of this in animals is the "enhancosome" complex (Levine 2010). We hypothesized that protein-DNA interactions should be sensitive to the concentration of all players including the protein binding sites located in the cis-regulatory regions of the gene encoding such an upstream transcription factor.

This report presents three primary results. 1) Grass genes associated with many CNSs tend to possess homeologous duplicates retained over the ~70 million years since the pre-grass tetraploidy (Fig. 3). 2) Within individual transcription factor gene families, the most CNS-rich members are significantly more likely to possess retained duplicate copies than the least CNS-rich members. 3) Looking at copies of the same genes from the pre-grass tetraploidy, the less CNS-rich copy is significantly less likely to have both duplicate copies retained in a second round of whole genome duplication in the maize lineage (Table 1).

The concentration of the DNA binding sites and the concentration of the proteins that bind them would tend to have evolutionarily preferred stoichiometries such that fractionation (deletion) of a copy of the gene would be selectively negative because this changes the relative concentration of binding sites and binding proteins. While our results are consistent with and support our hypothesis, *our explanation is not proved*. There is at least one alternative explanation for our data. It is possible that the deletion of the regulatory sequences identified by CNSs reduces the contexts – tissue/organ/cell types, developmental time points, responses to stimuli – in which a gene is expressed. If a gene only participates in dose-sensitive protein-protein interaction in some specific expression contexts, the loss of CNSs could conceivably reduce the opportunities for the resulting

protein to continue participating in dose-sensitive interactions and this could eliminate the selective cost associated with the loss of a duplicate gene copy. Without a detailed gene expression atlases for maize and its outgroup sorghum it is impossible to definitively rule out this alternative.

The supposition that the over-retention of transcription factor genes following whole genome duplications is the result of dose sensitive protein-protein interactions is an extrapolation from better-known CNS-poor gene categories such as genes encoding ribosomal proteins and is not directly supported for genes encoding transcription factors. Gene dosage effects are clearly the best single explanation for the changes that occur to gene content following whole genome duplication. However, the teoretical mechanisms explaining gene dosage should be broadened from its current focus on the concentration of protein products (Veitia 2010) to include, for transcription factors at least, the concentration of cis-acting protein binding sequences associated with genes themselves.

## Methods

### CNS Discovery

The evolutionary distance between the genomes of rice and sorghum places them within the interval for CNS discovery (as reviewed (Freeling and Subramaniam 2009)). Using the CNS Discovery Pipeline (Woodhouse 2010), 48744 total CNSs (all strictly syntenic) were identified near 16,013 pairs of rice TIGR5- -sorghum JGI1.4 orthologs. This list is called the Os-Sb genelist, v2. B. Pedersen Freeling Lab, 2009, and is included as Supplemental Information S1.

### Identification of orthologous and homeologous syntenic segments for use in these experiments

Inter- and intra- species blocks of collinear homologous genes were identified using the online tool SynMap (Lyons *et al.* 2008b) and enlarged using the merge function of the QuotaAlign algorithm enabled within SynMap (Tang et al., **Submitted**). Collinear blocks were classified as either homeologous or orthologous based on analysis of aggregate synonymous substitution rates between all homologous gene pairs within a block of collinear genes, as previously described (Schnable *et al.* 2011b).

### Classification of maize retention

For each orthologous rice-sorghum gene pair we identified two orthologous locations within the maize genome. An orthologous maize gene was considered to be present either if a gene present at the predicted orthologous location matched against the rice and sorghum orthologs, or if a LASTZ (Harris 2007) search of the region identified a putative unannotated gene or gene fragment similar to the rice and sorghum orthologs.

**Figures**

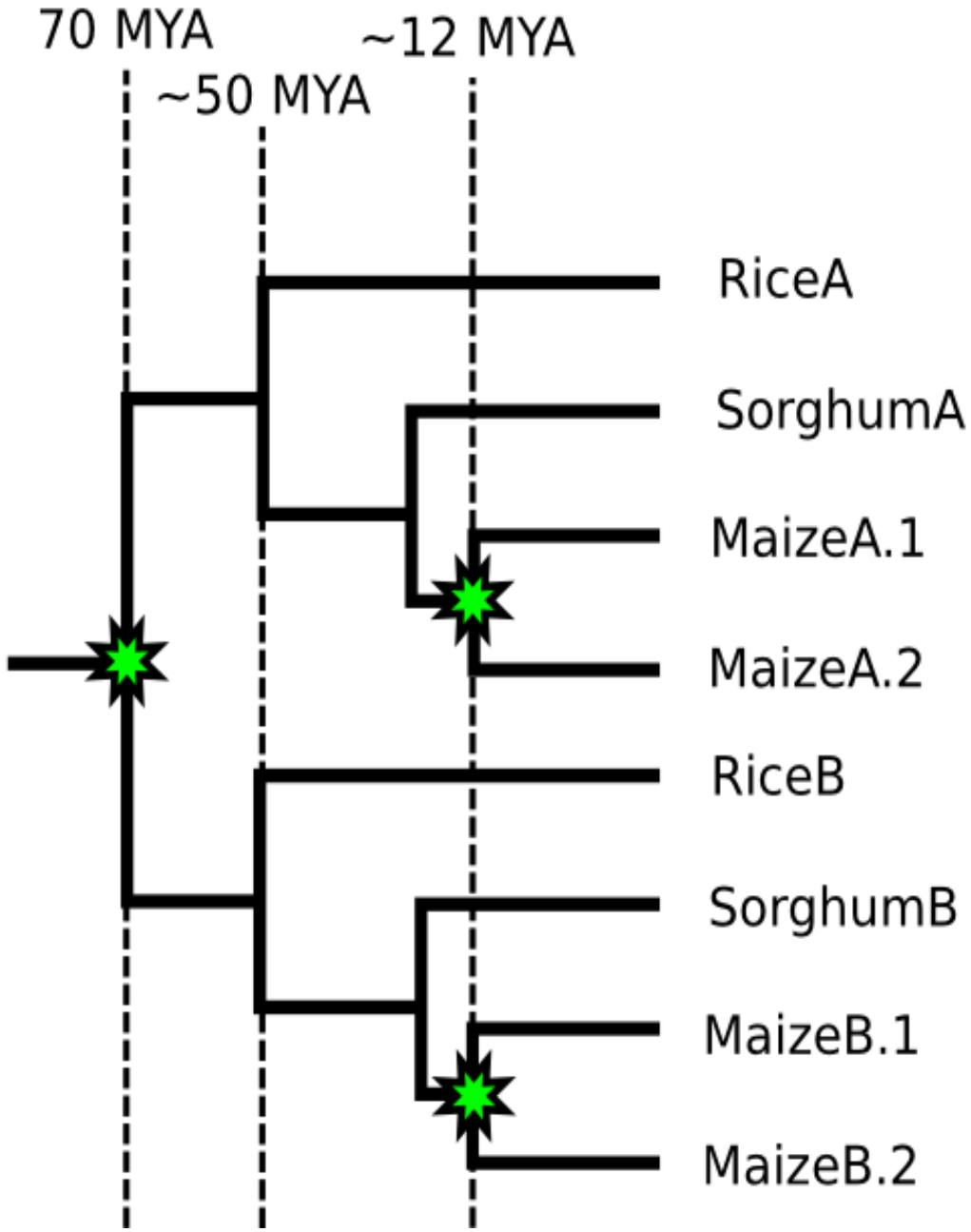



Figure 1
Genomic relationships between the species rice, sorghum and maize. Nodes marked with stars represent divergence by whole genome duplication. All other nodes represent divergence by speciation.

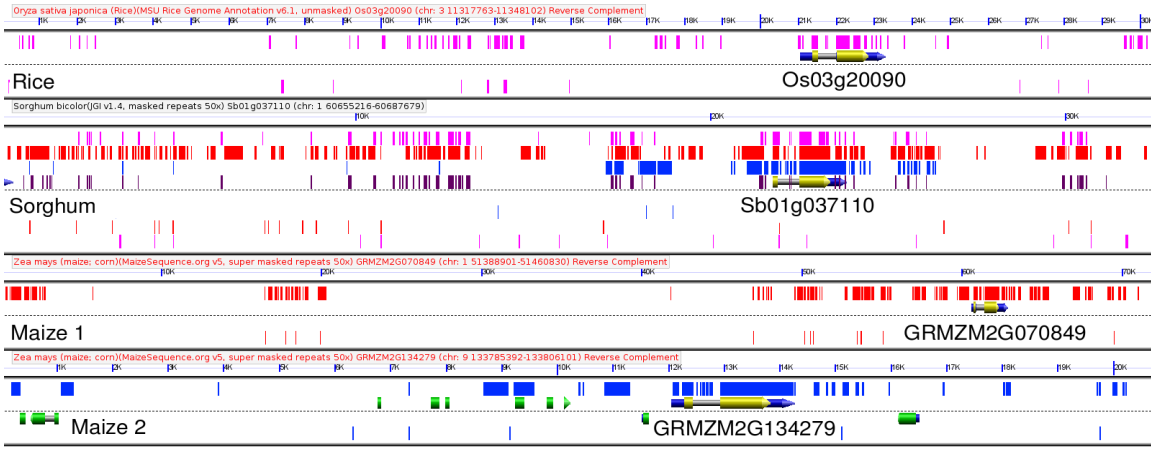



Figure 2
Relationship between myb transcription factor gene Sb01g037110, the single most CNS rich gene in sorghum, and its syntenically retained orthologs in rice and maize. Exons of the genes in the orthologous group containing this gene are marked in yellow, exons of all other genes are marked in green. Sequences identified as homologous by blastn between sorghum and rice are identified by purple rectangles. Sequences annotated as conserved noncoding sequences by the CNS-PIPELINE version 1 are marked in dark brown on the sorghum track, second from the top. Blastn hits between and maize1/maize2 are marked with red and blue rectangles respectively. This graphic was generated using GEvo, part of the CoGe toolkit (LYONS *et al.* 2008a) An interactive version of this experimental result can be regenerated by visiting the following link: http://genomevolution.org/r/2bgw

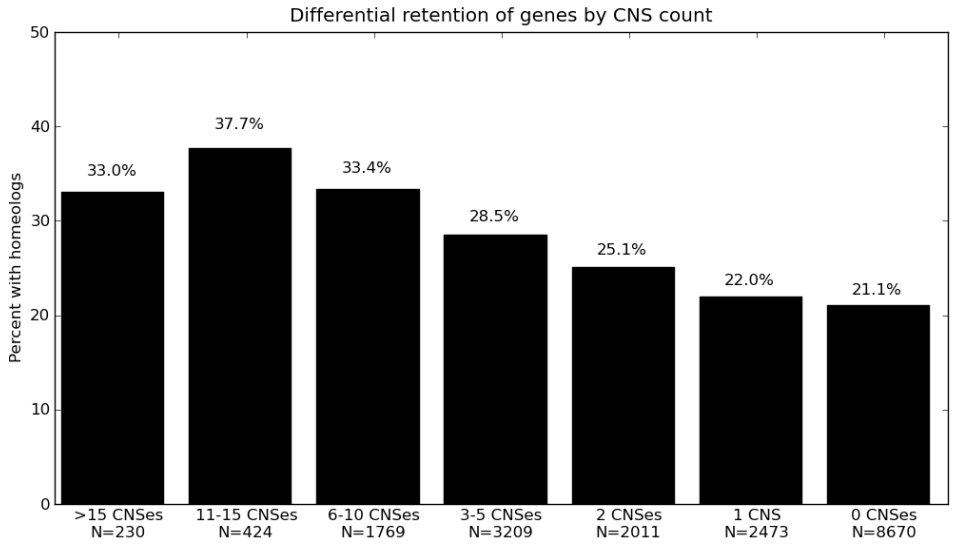



Figure 3
Odds of possessing a retained homeologous gene from the pregrass whole genome duplication for genes with different numbers of associated CNSs.

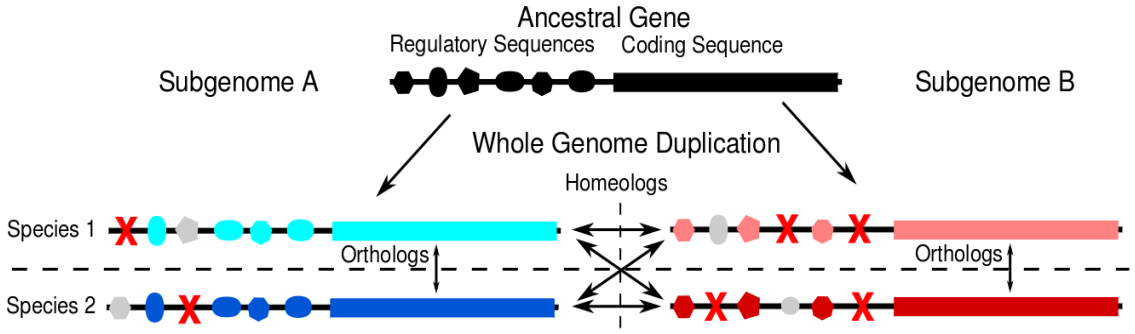



Figure 4
A hypothetical example of how regulatory sequences of duplicate genes might evolve following whole genome duplication. The original whole genome duplication creates two homeologous copies of an ancestral gene, both of which evolve separately in two species, rice and sorghum that arose from the original tetraploid species. Red X's mark deleted sequences. Gray shapes represent intact regulatory elements which will not be identified as CNSs by comparing orthologous genes between species 1 and 2 because they are no longer shared between the two species. In this example, the genes located in subgenome A has retained more regulatory elements in both species than have the homeologous genes in subgenome B. As a result the genes in subgenome A possesses four orthologous CNSs, while the gene in subgenome B possess only three.

**Tables**

**Table 1:**

| | Both copies retained in maize | Fractionated (only one copy retained) | Neither Copy Retained |
|---|---|---|---|
| Homeolog with more CNSs | 282 (56.7%) | 202 (40.6%) | 12 (2.4%) |
| Homeolog with less CNSs | 217 (43.7%) | 253 (50.9%) | 27 (5.4%) |

# Chapter 7: qTeller

The following chapter has not yet been published beyond this dissertation.

**Preface:**

Often – although not always – papers describing the generation of RNA-seq expression data would also include supplemental datasets listing the expression of each gene in each condition. However, while these datasets were useful for examining the expression of genes among the datasets generated by a single research group it was not possible to use these data to compare expression among different data from different research groups, because there are a diverse set of tools and “best practices” for the analysis of RNA-seq data. As a side effect of the research project described in chapter 2, I ended up with sets of data describing the expression of every gene in maize in different tissues and cell types from a number of different research groups all quantified using the same analytical pipeline.

Other maize researchers started asking me to look up the expression values for their favorite gene within these datasets. To automate this process I first developed a tool for retrieving expression data for all genes within an interval. Later, because the human brain is better at spotting patterns in graphic data than spreadsheets I added the ability to display the expression data for a single gene or pair of genes graphically. Both the continues publication of maize RNA-seq datasets and the occasion tweek of my analytical pipeline required that I run the analysis pipeline over again so I ultimately automated the analysis process itself with customizable python scripts. The end result was qTeller, a modular tool which can be deployed on a web server to let the public or a select group of users visualize and compare expression data from public or internally generated expression datasets in any species with a sequenced genome. Recent updates have added the ability to use a transcriptome assembly in place of a genome to quantify gene expression in nonmodel species.

One use of qTeller is to compare the the expression of homeologous gene pairs. These comparisons underlie the technique of fractionation mutagenesis which uses natural deletions within promoters to characterize the function of specific promoter elements. Fractionation mutagenesis will be described in more detail in Chapter 8.

## Introduction:

RNA-seq is rapidly displacing microarrays as the preferred technology for measuring genome-wide levels of gene expression. In contrast to the older microarray technology, RNA-seq is has a much higher dynamic range, reports absolute rather than relative expression and can be employed in any species without the necessity of developing custom microarray chips. The count-based character of RNA-seq expression measurements also results in simpler and more powerful statistics than were possible with the analog expression values reported by microarrays.

The most common use of RNA-seq data is the identification on differentially expressed genes between an experimental dataset – generally a mutant, specific cell type-tissue, or an organism exposed to a specific outside stimulus -- and a control dataset. This analysis requires the consideration of both technical and biological variation as well as stringent corrections for multiple testing. The diversity of tools and statistical packages developed to address the change of identifying differentially expressed genes using these datasets is too great to be discussed in depth here.

However single genome-wide approaches to the analysis of RNA-seq datasets leave a great deal of potentially informative biologicial data on the table. Geneticists and other researchers conducting investigations on a smaller than genome-wide scale are often interested in the expression of either a single gene or a group of genes – perhaps a gene family or the genes located within a specific interval on a chromosome to which a mutant or QTL has been mapped. Ideally these researchers would be able to gather data on the expression level of their gene(s) of interest across all published RNA-seq datasets to narrow down lists of candidate genes, or gain additional insight into gene function. In practice a number of issues stand in the way.

First, many research papers publish only lists of differentially expressed genes and not estimated expression levels for all genes in the dataset. Second, even when expression levels for all genes are published, comparisons between different datasets can be confounded by differences in the pipelines and assumptions used to quantify gene expression as well as differences between different versions of genome assemblies and gene model annotations used to by different research groups to quantify expression.

Fortunately, high caliber journals continue to require that authors deposit their raw sequence data in online repositories just as NCBI's SRA,  of gene expression between experiments  the alignment of short sequencing reads to a genome, Different analytical pipelines produce different measures of gene expression so it is necessary to return to the raw sequencing reads for each experiment. These data are presently available from the NCBI Sequence Read Archive in the US, EBI's European Nucleotide Archive, Japan's DDBJ Sequence Read Archive. This raw data can be reprocessed using a single analytical

pipeline to produce comparable measures of gene expression for diverse sets experiments conducted by multiple research groups – although obviously the usual caveats about comparing data from plants grown at different locations by difference researchers still apply.

## Results:

To facilitate this reprocessing of multiple raw-read datasets a python script was developed which automated the analysis of all steps of gene expression quantification including quality and adapter trimming using cutadapt (Martin 2011), short read alignment using GSNAP (Wu and Nacu 2010), and gene expression quantification using either Cufflinks (Trapnell *et al.* 2010) or eXpress (Robers & Pacher, in review) depending on the specific use case (Figure 1). GSNAP was selected because of its combination of tolerance of SNPs and InDels, ability to detect spliced alignments of reads which span exon-junctions, and reasonable run time. The various options available for spliced alignment of short reads to reference genomes have been recently reviewed (Grant *et al.* 2011).

A second python script automates the formatting of the resulting expression data files – along with other genome level data such as functional annotations and lists of orthologous genes in related species – into an SQLite database which serves as the modular data store for the qTeller web interface (Figure 2). The qTeller web interface allows researchers pull out detailed information on the predicted functions, orthologs, and expression patterns of all genes within an genomic interval (useful for identifying candidate genes while fine mapping mutants or QTLs) or to visualize expression for individual genes across multiple conditions and genotypes (Figure 3). This web interface can either be made publicly accessible – for published datasets – or protected by username/password authentication to allow research groups to share expression data internally.

The expression levels of a set of well studied “classical” maize mutants were visualized using the qTeller interface. Most genes showed the expression patterns expected based on the mutant phenotype and previously published individual investigations (Figure 3). However, on occasion qTeller would identify a previously unreported expression domain for a known maize mutant.

*Tasselseed1* (*ts1*) is a mutant whose history in the maize literature dates back to a description by R. A. Emerson, the founder of modern maize genetics in 1920 (Emerson 1920). In plants carrying mutant alleles of *ts1* the staminate (male) florets of the maize tassel are transformed into pistillate (female) florets. In 2009 the gene encoding *ts1* was cloned and show to encode a class 2 13-lipoxygenases involved in the biosynthesis of the plant hormone jasmonic acid (JA) (Acosta *et al.* 2009). Given the diverse roles of JA

signaling (KAZAN and MANNERS 2008) the fact that the phenotype of a JA biosynthenic mutant is confined to the florets of the maize tassel is notable (ACOSTA *et al.* 2009). The same research group also identified the homeolog of *ts1* which they refer to as *ts1b*. *Ts1b* showed a similar expression pattern to *ts1* but as significantly lower absolute levels of expression in all five tissues examined (roots, stems, leaves, tassels, and ears) (ACOSTA *et al.* 2009).

Using qTeller it was possible to revisit this question and quantify the relative expression of *ts1* and *ts1b* in a much wider range of tissues. RNA-seq data largely agreed with the published ratio of expression between the two homeologous genes, however the greater diversity of developmental data represented allowed the discovery that *ts1b* is expressed at equal or greater levels than *ts1* in developing anthers and mature embryos (Figure 4). This result suggests double mutants of *ts1* and *ts1b* may be required to determine the function of the the common ancestor of these duplicated genes played in the development of grass embryos.

## Discussion:

The qTeller analytical pipeline is not a substitute for a trained bioinformatician or computational biologist. It is not designed to produce lists of differentially expressed genes supported by statistical significance and proper controls for multiple tested and experimental noise. What it does effectively is provide information on the expression of individual genes or the genes in a short genomic interval comparable to what can be learned from conducting an rtPCR experiment. As such, it can aid in the selection of candidate genes when mapping QTL or mutants, and developing hypotheses about differences in gene regulation between duplicate genes. Hopefully this software will make the the broad insights into gene expression RNA-seq experiments make possible more readily accessible to all biologists, not only those with the training and computer resources to download and recapitulate computational analyses on gene expression on their own.

When comparing data from multiple experiments conducted in multiple institutions, it is important to keep in mind that there are countless possible sources of variation beyond the tissues, mutants, or environmental stimuli named in the dataset. These concerns can partially be addressed by data richness. For example, when all the data from a single paper shows a different pattern of expression, this indicates the difference is likely of some confounding factor (differences in growing conditions, sample collection technique, RNA-seq library preparation, etc). In contrast, when the same pattern is observed in multiple samples from different research groups, this actually provides increased confidence in the result since it has been observed in the most unlinked possible of biological replicates.

We have currently deployed qTeller instances for four major grass species with published genomes: brachypodium, rice, sorghum, and maize which are open to the community at http://qteller.com. qTeller's web interface allows other databases to link directly to the expression reports for individual genes, and a collaboration with MaizeGDB (Lawrence *et al.* 2008) has added direct links from MaizeGDB locus pages to qTeller expression reports allowing individual maize researchers with no expertise in RNA-seq analysis to quickly access the information one the expression of specific genes-of-interest in all published maize RNA-seq data. In the month of November 2012 these instances of qTeller recieved >500 unique visits (Supplemental Figure S2) with traffic coming primarily from the US, China, and Mexico.

The source code for the qTeller analysis pipeline and web interface are freely available for download ( https://github.com/jschnable/qTeller ) and it is hoped that the modular nature of these tools will allow them to be of use to researchers working a wide range of species both for establishing community resources and visualizing internally generated datasets in an intuitive fashion.

## Methods:

**Sources of RNA-seq Data:**
Sequence read archive. Downloaded using downloader.py (included in github repository). Maize expression data was taken from six papers (Wang *et al.* 2009; Jia *et al.* 2009; Li *et al.* 2010; Davidson *et al.* 2011; Waters *et al.* 2011; Bolduc *et al.* 2012) and two unpublished set of experiment (SRP006965). Sorghum data was taken from two papers (Dugas *et al.* 2011; Davidson *et al.* 2012). Rice data was taken from three papers (Oono *et al.* 2011; Zhang *et al.* 2012; Davidson *et al.* 2012). Currently the only brachypodium expression data we have access to is that published in (Davidson *et al.* 2012).

**Quantifying Gene Expression:**
RNA-seq data was aligned to the reference genomes of maize (B73_refgen2) (Schnable *et al.* 2009), rice (MSU 7 (Ouyang *et al.* 2007)), and sorghum (Sbi 1.4 (Paterson *et al.* 2009)) using GSNAP (2012-7-20 release), a splice, indel, and SNP tolerant short read align program (Wu and Nacu 2010). The resulting alignment files were reformatted from SAM to BAM format and sorted using SAMtools 0.1.17 (Li *et al.* 2009). Final gene expression values in FPKM -- (F)ragments (P)er (K)ilobase of exon per (M)illion aligned reads – were calculated using cufflinks 1.3.0 (Trapnell *et al.* 2010). All of these steps are automated within the auto_analyze_generic.py script (included in the github repository).

**Syntenic Ortholog Data:**
Syntenic orthologs and homeologs for the maize, sorghum, and rice instances of qTeller were generated using SynMap (Lyons *et al.* 2008b) with the QuotaAlign filter (Tang *et al.* 2011). A quota of 2:1 was used for maize vs other grasses and 1:1 for all other

comparisons. Equivalent datasets for other species can be generated using SynMap's web interface. The raw data format downloaded from SynMap was to an easily parseable and human readable format using the orthologs_from_synmap.py script (included in the github repository).

**Data On Gene Function:**
Automated functional gene data was taken from MSU's rice annotation project (Ouyang *et al.* 2007), phytozome's annotation of the sorghum genome (Paterson *et al.* 2009), and maizesequence.org's annotation of the maize genome (Schnable *et al.* 2009). In addition maize genes studied by geneticists (the classical genes of maize genetics) were loaded into the maize qTeller instance as a second functional annotation track (Schnable and Freeling 2011).

**Generating the qTeller Database:**
Data on the expression, function, orthologs, and physical location data of each gene in a given genome were stored within an SQLite database using the build_qt_db.py script (included in the github repository). The schema for this database is reported in Figure S1.

**Gene Expression Visualization:**
Custom python scripts included in the qTeller web installation retrieve expression data from the SQLite database and draw graphics using the matplotlib software package (http://matplotlib.org). Expression graphics can either be generated on the fly through the web interface at the request of users or pregenerated for all genes in the dataset for distribution or integration into other research tools.

**Identifying Orthologs Among Grass Species:**
Grass orthologs were identified using the method described in (Schnable *et al.* 2012a). Briefly, blocks of syntenic genes were identified in SynMap (Lyons *et al.* 2008b) and then the aggregate synonymous substitution rate of gene pairs within each block were used to distinguish orthologous regions from syntenic blocks resulting from the pregrass whole genome duplication.

**Figures**

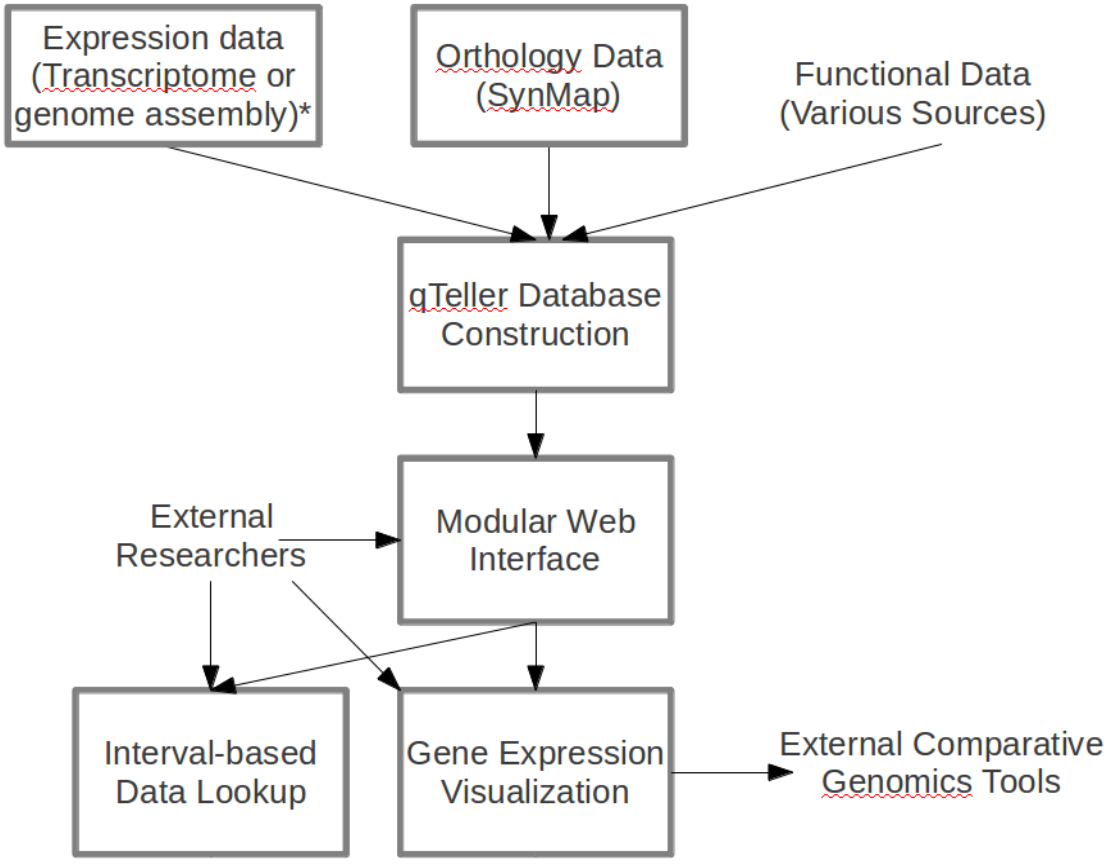



Figure 1: Overall workflow for the qTeller pipeline

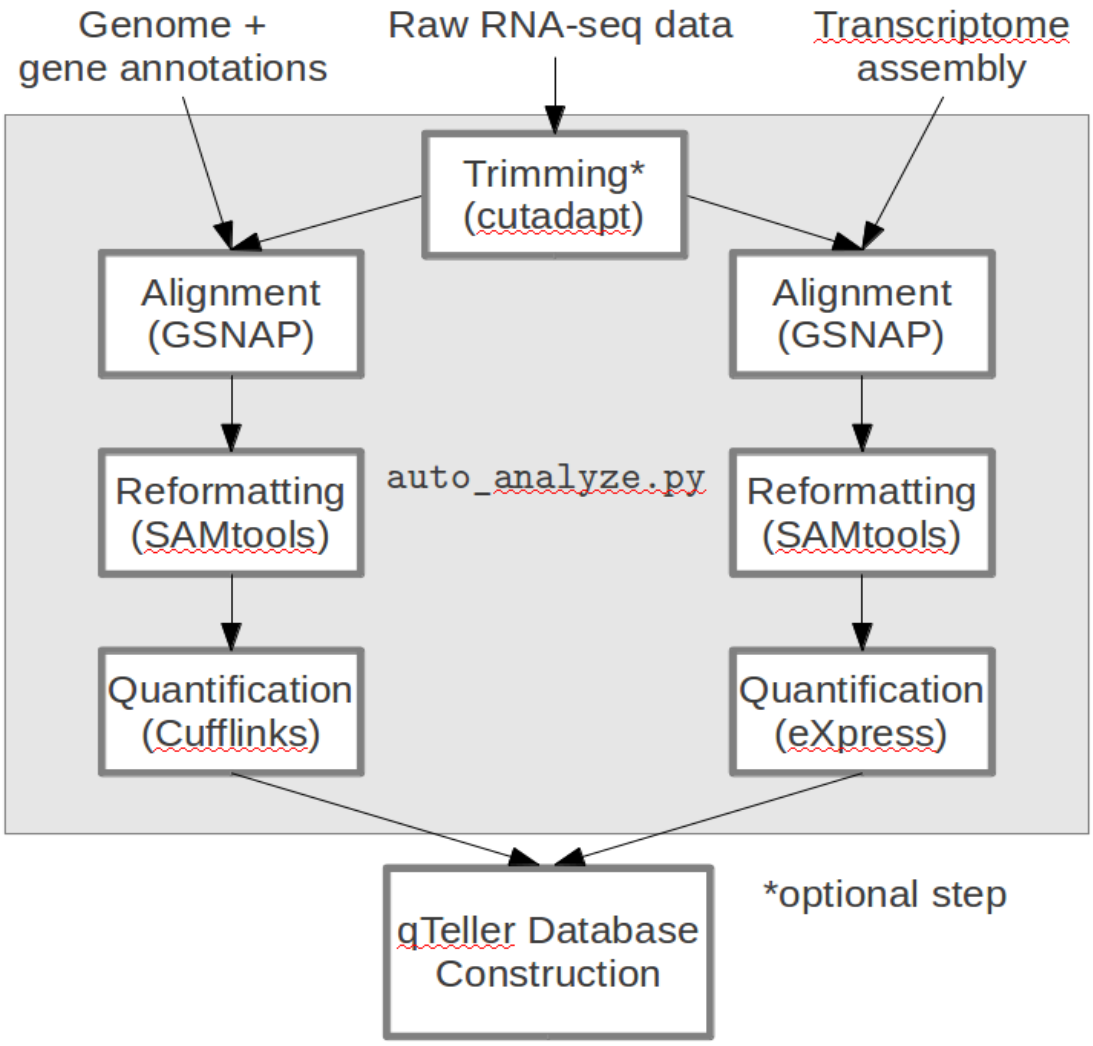



Figure 2: Detailed view of the RNA-seq analysis pipeline implemented within qTeller with alternative options for transcriptomic data.

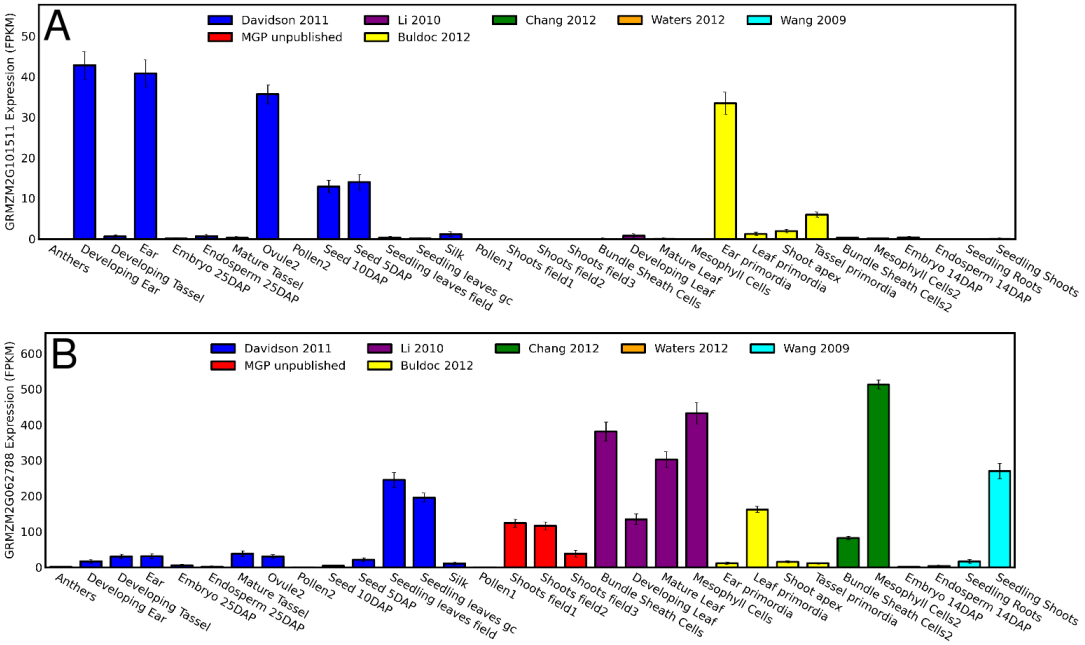



Figure 3: Examples of expression data profiles for two classical maize genes. A) Expression of tga1, a known domestication locus and a gene regulating the development of glumes around kernels. Consistent with its function, this gene shows high expression in ear, ovule, and seed datasets. B) Expression of bsd2 a gene whose mutants disrupt chloroplast development. Consistent with its known function, this gene shows the highest expression in photosynthetic vegetative tissues.

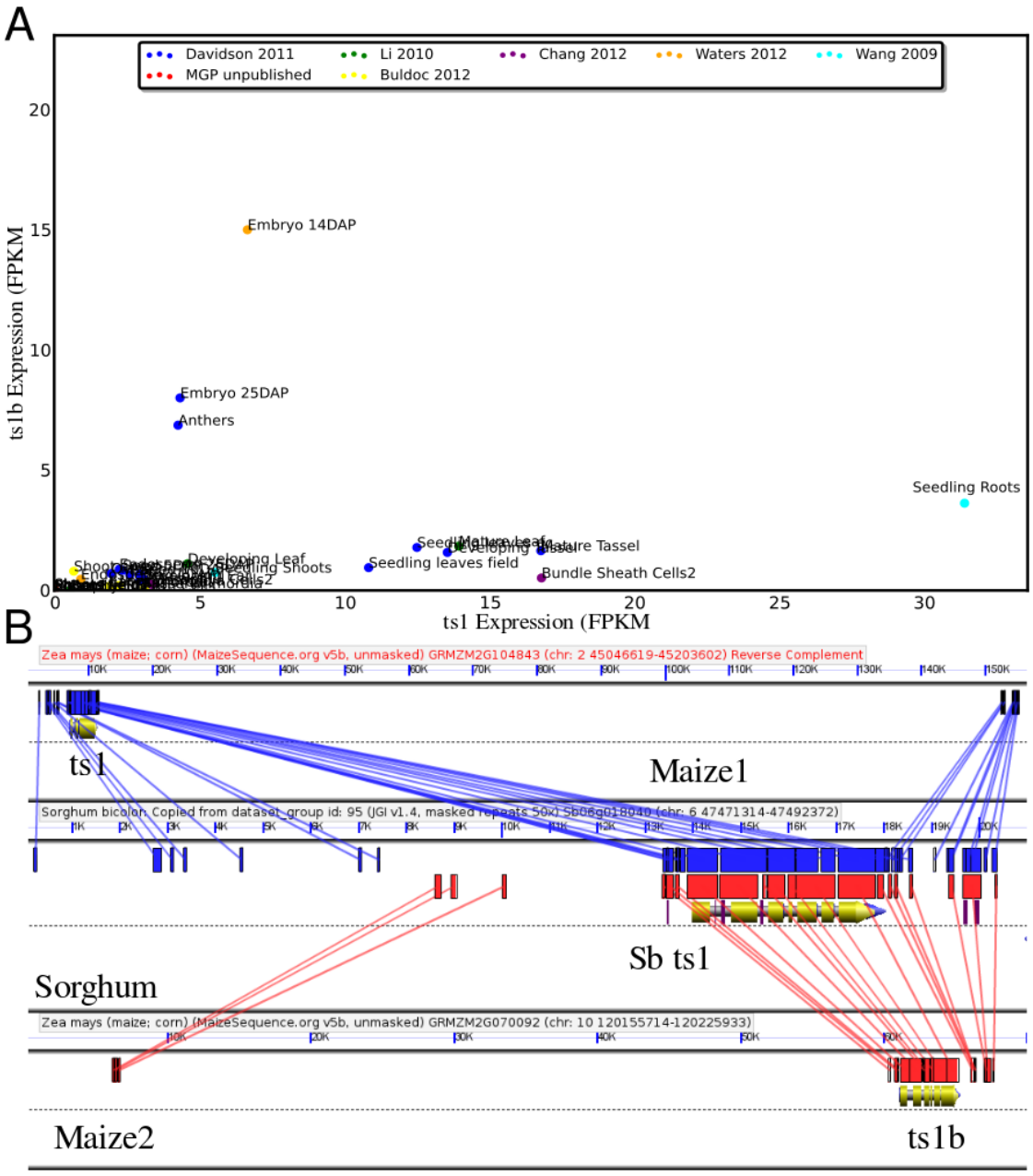



Figure 4: A) A comparison of the expression levels of ts1 (GRMZM2G104843) and ts1b (GRMZM2G070092) observed in published maize RNA-seq datasets. Regenerate this analysis at qTeller with the following link http://qteller.com/qteller3/scatter_plot.php?name1=GRMZM2G104843&name2=GRMZM2G070092&xmax=32&ymax=22&info= B) A GEvo comparison (Lyons *et al.* 2008a) of the genomic regions surrounding ts1 and

ts1b in maize to the sequence adjacent to their shared co-ortholog in sorghum (Sb06g018040). Exons are show in yellow. Regions of similar sequence, as identified by blastn are marked with blue boxes (ts1 and sorghum) or red boxes (ts1b and sorghum). Purple rectangles mark functionally constrained conserved noncoding sequences identified in a comparison of rice and sorghum (Turco et al, in prep). Regenerate this analysis with the following link: http://genomevolution.org/r/68qe

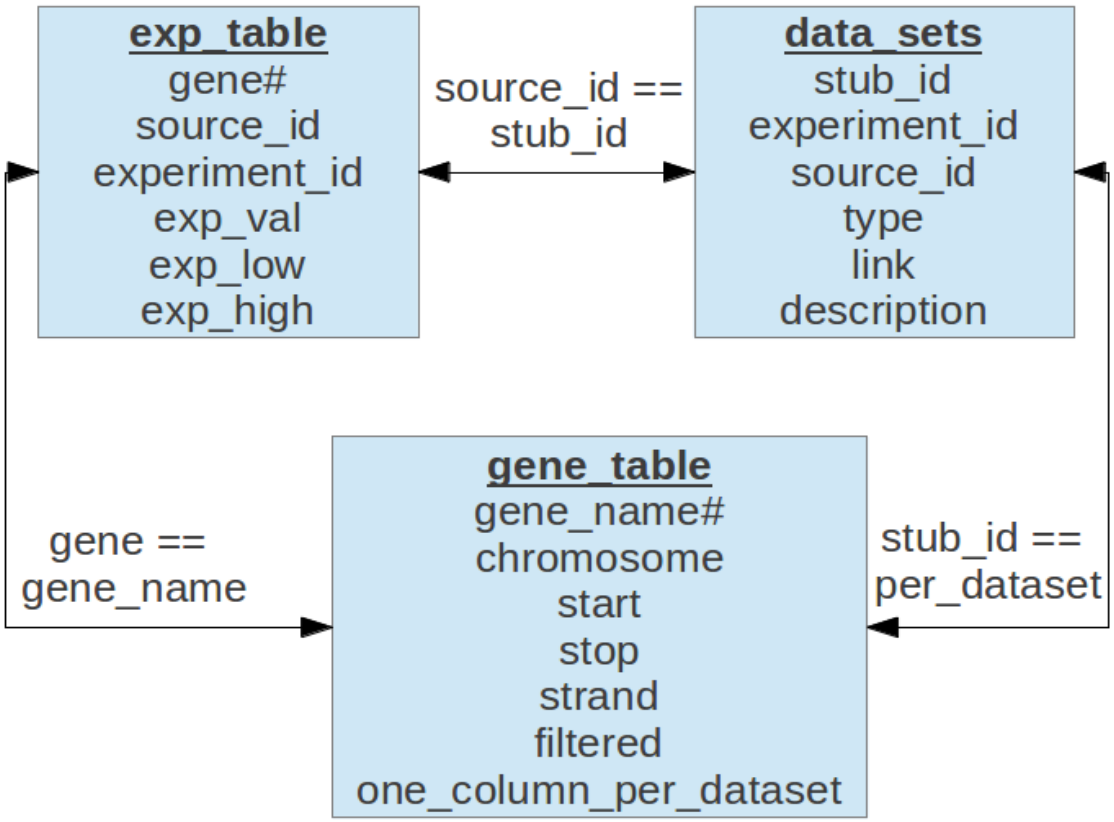



Figure S1: Schema for the qTeller database generated by the build_qt_db.py script.


Figure S2: Summary of growth in qTeller internet traffic since launch. recorded and visualized using google analytics.

## Chapter 8: Observing the function of conserved regulatory sequences using natural deletions in maize

The following chapter has not yet been published elsewhere.

Contributions:

Hugh Young conducted the test of the pollen/anther enhancer identified by fractionation mutagenesis reported in this chapter including the cloning of the promoter sequence, generation of the transgenetic Brachypodium distachyon lines and visualization of GUS expression (Figure 3 and Supplemental Figure S4).

### Preface:

In classical genetics a gene was defined as the basic unit of heredity. This definition includes both the segments of DNA specifying the amino acid sequence of a protein but also the regulatory sequences which determine where and when and in what quantities those protein coding regions will be transcribed. In contrast, most genome visualization tools and in most published genome annotations a gene is treated the sum of transcribed exons. In many cases even the transcribed but untranslated five prime and three prime UTRs are excluded. This shift of definition does not represent some nefarious plot on the part of genome annotators but instead reflects the simple fact that current methods of genome annotation have become reasonably accurate at identifying the protein coding regions of genes while the community lacks high throughput or automated methods of identifying specific regulatory sequences such as enhancers or repressors within genomes.

In this chapter a number of different types of analysis discussed previously are combined together to demonstrate a new method of identifying putative functions for conserved regulatory sequences *in silico*. Orthologous genes in grass species such as rice, sorghum, foxtail millet, and brachypodium are compared to each other to identify conserved noncoding sequences (Chapter 6). Pairs of maize homeologs co-orthologous to these sets of grass genes are then compared to identify deletions which have removed noncoding regions near the genes (Chapter 2) and the expression levels of both gene copies are compared diverse RNA-seq datasets gathered from published literature (Chapter 7). When the deletion of a specific conserved noncoding sequence from one gene copy coincides with loss of expression in certain datasets (for enhancers) or a gain of expression in certain datasets (for repressors) a putative function can be assigned to that particular sequence. The other gene copy serves as the wild-type control reporting on gene expression patterns in all datasets. It is important to observe the pattern of gene expression rather an absolute levels because the biased expression discussed in Chapter 2

appears to result in equal biases in expression across all datasets.

In this chapter it is also demonstrated that a candidate pollen/anther enhancer identified using this method in maize drives expression of GUS in a transgenic *Brachypodium distachyon* line.

## Introduction:

Computational tools for identifying protein coding genes within genomes have shown dramatic improvements in the past decade, as have the tools for predicting the functions of proteins encoding by novel genes. The sequence of regulatory sites, like the sequence of protein coding exons, are functionally constrained, allowing their identification as islands of conserved noncoding elements (CNEs) near the exons coding for orthologous proteins in multiple species (Gumucio *et al.* 1992, 1993). Unfortunately, the next step is usually "promoter bashing" of individual genes where all possible cis-acting sequence is assessed for cis-regulatory activity using transgenic assays. It would be useful to find a way to obtain functional hypotheses for individual cis-acting sequences without doing either wet-lab or field genetic research.

The earliest studies of conserved regulatory sequences were conducted by the sequencing and comparison of individual gene spaces from multiple mammalian species, where the function of each site could be determined using molecular biology techniques (Gumucio *et al.* 1988, 1992, 1993, 1996) However, as whole genome sequences became available from multiple species it became possible to conduct searches for conserved noncoding sequences on a genome-wide basis resulting in the identification of thousands of conserved noncoding sequences (Shin *et al.* 2005; Venkatesh *et al.* 2006; Thomas *et al.* 2007). In mammals, conserved noncoding sequences were shown to sometimes regulate multiple genes along a chromosome (Loots *et al.* 2000), and generally carry cis-regulatory information (Hardison 2000; Loots and Ovcharenko 2007). In vertebrates, the function of a number of elements have been assayed using transgenic reporter constructs and the majority of these functioned as regulators of gene expression in vivo (Shin *et al.* 2005; Ritter *et al.* 2010). A study of 25 conserved sequences associated with a four genes in both humans and zebrafish showed that individual elements drove expression in difference cell types, suggesting that many elements function as independent cis-regulatory modules which combine to create the total expression pattern observed by the gene (Woolfe *et al.* 2004). The gene regulatory function of a number of conserved noncoding sequences from plants have been demonstrated through mutant analysis (Freeling and Subramaniam 2009), or in the recent case of *Lateral Suppressor* transgenic reporter constructs (Raatz *et al.* 2011). While transgenic approaches to characterizing the function of conserved regulatory sequences have proven successful in the case of single target genes, the technique remains too time and resource intensive to characterize the function of regulatory elements on a genome-wide scale.

The adaptations and drift plant genomes has a different character than that observed in vertebrates, necessitating different comparative genomic approaches. Vertebrate conserved noncoding elements are still detectable in large numbers when comparing the genomes of human and elephant shark, a cartilaginous fish from a lineage that last shared a common ancestor with humans an estimated 530 million years ago (Venkatesh *et al.*

2006). Studies in flowering plants have determined that most CNS become non detectable within one hundred million years (Reineke *et al.* 2011), and only a few predate the divergence of monocots and dicots (D'Hont *et al.* 2012). This difference may be linked to the remarkable difference in the frequency of ancient whole genome duplications between vertebrates and flowering plants. A single tetraploidy in the lineage leading to fishes teolost fishes was associated with a significant decline in the number of CNEs detectable in these lineages (Venkatesh *et al.* 2007). All flowering plants sequenced to date carry the traces of multiple rounds of whole genome duplication within their genomes; for example, there have been at least 6 sequential tetraploidies in the lineage of a maize plant (Gaut and Doebley 1997; Paterson *et al.* 2004; Tang *et al.* 2010; Jiao *et al.* 2011).

A whole genome duplication creates two copies of every gene, each with a full ancestral complement of regulatory sequence. Over time many of these duplicate genes are lost from a genome, a process of gene loss that accompanies diploidization. This mutagenic consequence of polyploidy has been called fractionation (Langham *et al.* 2004) to distinguish it from diploidization, the rapid evolution of diploid inheritance and segregation during meiosis. While many duplicate copies of genes are rapidly lost following whole genome duplications, thousands of duplicate genes pairs are still retained in plant genomes, including that of grape which last experienced a whole genome duplication more than 100 million years ago (Jaillon *et al.* 2007). These retained duplicate genes tend to belong to certain functional categories, including the subunits of complicated molecular machines such as ribosomal (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Maere *et al.* 2005; Thomas *et al.* 2006). This observation is consistent with the maintenance of relative dosage between interacting proteins being a key driver of duplicate gene retention, the Gene Dosage Hypothesis (Birchler and Veitia 2007, 2010). Transcription factors and stimulus response genes are also more likely to show retention of duplicate copies following whole genome duplication. These genes tend to be under tight regulatory control and associated with large numbers of conserved noncoding sequences (Freeling *et al.* 2007). This mechanism for the retention of dose-sensitive genes functions by preserving ancestral balances of expression between interacting proteins, and requires no change in the function or regulation of duplicated genes.

More than a decade ago, another model was proposed as an explanation for the high numbers of duplicate gene pairs observed in eukaryotic genomes. The duplication/degeneration/complementation (DDC) model proposed that following gene duplication, independent loss of cis-regulatory sequences from each gene copy could create situations where both genes became necessary to maintain the ancestral pattern of gene expression (Force *et al.* 1999; Lynch and Force 2000). This subfunctionalization by promoter disruption could rapidly make the loss of either copy detrimental to fitness, locking in the requirement to retain both duplicate genes within the genome. However, it functions only after two independent mutational loss events. Retention by dosage

balances and retention by subfunctionalization are not mutually exclusive. For example, initial retention could be by preservation of dosage balances and this could be made permanent by subfunctionalization.

In plants, the same mechanism that removes duplicate genes following whole genome duplication – short to medium size deletions mediated by nonhomologous recombination – also removes conserved noncoding sequences from the promoters of duplicated genes following whole genome duplications (Subramanium et al, *in press*).

Deletion of conserved regulatory sequences from one of two duplicate copies of a gene provides an opportunity to deduce the function of individual conserved noncoding sequences *in planta*. The function of the lost regulatory site should be reflected in differences between expression patterns of the two duplicate genes within the same gene expression datasets. This naturally occurring form of promoter bashing – or fractionation mutagenesis (Freeling *et al.* 2012) – provides an opportunity to obtain detailed understanding of the function of individual gene regulatory elements, and unequivocal evidence that a particular noncoding element is functionally a part of a particular nearby gene.

Here we test the efficacy of fractionation mutagenesis. We focused on maize, both a major crop species and a genetic model system for which a significant quantity of published RNA-seq expression data is available (Wang *et al.* 2009; Li *et al.* 2010; Davidson *et al.* 2011; Waters *et al.* 2011; Bolduc *et al.* 2012; Chang *et al.* 2012). The maize lineage experienced a whole genome duplication between 5-12 million years ago, after the divergence of its close ancestor, sorghum (Gaut and Doebley 1997; Swigoňová *et al.* 2004). The large number of genome sequences available from diverse species in the grass family make it possible to identify pan-grass regulatory sequences (Turco et al, *in prep*), and each of these sequences may be followed through the whole genome duplication and subsequent fractionation present in the maize lineage.

## Results:

A previous study of the mRNA expression patterns of whole genome duplicates in arabidopsis reported only 43% of duplicate genes from the most recent whole genome duplication in that lineage remained significantlcorrelated when correlated is defined as an pearson's correlation r value >= to the 95$^{th}$ percentile of r values obtained by comparing the expression patterns of random pairs of genes (Blanc and Wolfe 2004). The same analysis was conducted using expression values calculated from published RNA-seq reads datasets. In maize 62% of duplicate gene pairs from the most recent whole genome duplication where each gene was expressed in at least one dataset show correlated expression (1837 gene pairs) (False Discovery Rate=5%) (Figure S1, Dataset S1). Compared to all gene pairs those with significantly correlated expression (Fig. 2A)

were enriched in the GO annotations “protein complex” (FDR corrected p-value: 0.032) and its parent annotation “macromolecular complex” (FDR corrected p-value: 0.001) when compared to maize gene pairs without significant correlation in gene expression (Fig. 2B), a result consistent with the predictions of the gene dosage hypothesis.

While the majority of maize gene pairs show correlated patterns of expression, a subset did show expression patterns more consistent with subfunctionalization. The homeologous maize genome pairs with the largest *negative* correlation in expression pattern were manually examined. Of these, twenty-six cases showed unambiguous developmental partitioning of expression (Dataset S2).

The availability of expression data from other grass species (Davidson *et al.* 2012) makes it possible to reconstruct the ancestral expression domain for the gene pair. Of the twenty six cases where manual examination confirmed the automated call of reciprocal expression, in nineteen the expression of both maize gene copies was necessary to provide expression in all tissues where data from other grasses indicated the ancestral gene gene was expressed (Dataset S2). In at least five of these cases the partitioning of expression could be linked to the reciprocal deletion of noncoding sequences. It was possible to assign hypothetical functions in individual conserved noncoding sequences by comparing the partitioning of expression domains between homeologous genes and the partitioning of conserved noncoding sequences it was possible to assign hypothetical functions in individual conserved noncoding sequences, as shown in Figure 3. Here the loss of one CNS is associated with expression in vegetative CNS and two other CNS with expression in infloresence tissues. The “subfunctionalization” of corn plants with separate male and female inflorescences – unlike the bisexual inflorescences of sorghum and rice -- further allows the assignment of these CNS to expression in male reproductive tissue.

Developing hypotheses about the function of individual conserved noncoding sequences requires, in principle, only the loss of a regulatory sequence and associated expression pattern from a single gene copy (Figure 2D). To test the effectiveness of this approach, we searched for homeologous gene pairs in maize where only one of the homeologs was expressed in pollen. Newly dehissed pollen was selected as a distinct cell type, and one for which two independent expression datasets were available in maize. Fifty-one genes pairs were identified where the ratio of expression between the duplicate gene copies was unambiguously different between vegetative organs and pollen and anthers, a pollen-containing organ whose expression was highly correlated with pollen (Figure S3). Of these 52 cases, 36 could be classified as ancestrally expressed in anthers/pollen based on expression data on the expression of orthologous genes in the anthers of rice, sorghum, and brachypodium (Davidson *et al.* 2012). In the remaining 16 cases, it appears likely that pollen-specific expression arose independently in the maize lineage following its whole genome duplication.

In six cases noncoding, ancestrally conserved, sequence lost only by the maize gene copy showing unexpectedly low expression in pollen and anthers could be identified. These sequences are candidate novel pollen-specific enhancers of gene expression. The candidate sequence inferred from one gene pair, one we used as an example previously (Freeling *et al.* 2012), was amplified from the sorghum genome and inserted into a GUS reporter construct (blue colored cells) alongside a minimal 35S promoter. The minimal 35S promoter is insufficient to drive expression of GUS in pollen/anther tissue, or anywhere else in the embryo, seedling or plant. However with the insertion of these conserved noncoding sequences, robust GUS expression was observed in both pollen grains and anther walls (Fig. 3). A survey of gene expression in additional organs, parts and tissues confirmed expression is largely confined to pollen/anthers, although the construct may also also drives expression at the floret base, an organ region not represented among current maize developmental expression datasets.

## Discussion:

**Mechanisms explaining gene pair retention after polyploidy**

The loss expression domains by one of the two homeologs (nonfuctionalization) appears to be significantly more common in among maize gene pairs generated by whole genome duplication than is subfunctionalization, where both copies are required to recapitulate the ancestral expression pattern. Current surveys of gene expression cover only a portion of development in maize and other grasses. It seems likely that many other gene pairs with reciprocal patterns of CNS deletion will show partially or completely partitioned expression domains as more complete and detailed expression RNA-seq gene expression atlases become available in maize and other grass species. Many CNS are believed to function in regulating the expression of genes in response to changes in environmental stimuli (Freeling *et al.* 2007; Spangler *et al.* 2011), an area neglected in the current set public gene expression dat.

The partitioning (subfunctionalization) of expression domains between duplicate genes will occasionally create a release from conflicting selective constraints, allowing protein and regulatory sequences to specialize and diverge. However, we agree with the theoretical model that all cases subfunctionalization are initially selectively neutral and could well remain that way (Force *et al.* 1999; Lynch and Force 2000). Unlike the predictions based on pure gene dosage theory [“balanced gene drive” (Freeling 2009)], retention of gene pairs where unequal expression of gene copies combined with repeated rounds of whole genome duplication can permit an eventual escape from constraints on deletion resulting from gene dosage interactions (Schnable *et al.* 2012c). With subfunctionalization there is no easy way to remove either copy of the retained gene pair, so the more diverged the expression patterns of a pair of duplicate genes are, the more likely both copies have become permanent fixtures in the maize genome

In our analysis of gene pairs differentially regulated in pollen, genes ancestrally expressed in pollen outnumbered by 2:1 genes which had acquired novel expression in pollen. Among arabidopsis whole genome duplicates differentially expressed in pollen, gains of novel pollen expression outnumbered losses of ancestral pollen expression (Liu *et al.* 2011). The higher proportion of ancestrally pollen-expressed gene pairs in maize could suggest that many of these gene pairs will ultimately be fractionated back to a single copy state. If so, the gene copy that has already lost a portion of its ancestral regulatory sequences is predestined for eventual deletion. The maize whole genome duplication is much more recent than arabidopsis alpha – the modal maize gene pair has experienced 0.15 synonymous substitutions per site compared to .76 model synonymous substations per site in arabidopsis -- and fractionation of genes is continuing in the modern maize genome at a significant rate (Schnable *et al.* 2011b).

**Studying gene regulation using natural deletions**

Comparisons of gene expression among differing grass species concluded that only a fraction of orthologs share conserved patterns of gene expression (Davidson *et al.* 2012). The existence of conserved noncoding sequences between a give set of orthologous genes does not guarantee that gene expression pattern will be conserved. A study of conserved noncoding elements from humans and zebrafish identified both cases of trans-regulatory divergence where upstream regulators had changed patterns between mammals and telost fish and cis-regulatory divergence where versions of the same conserved noncoding element from human and zebrafish drove different patterns of reporter gene expression in both mouse and fish (Ritter *et al.* 2010). In total only 30% of conserved noncoding elements examined drove comparable patterns of expression in both both lineages. A detailed investigation of conserved noncoding sequence function on the scale of the research conducted by Ritter and coworkers was beyond the resources and scope of this investigation. However the observation that sequences isolated from the sorghum genome based on aberrant maize gene expression patterns produced the expected patterns of expression when transformed into *Brachypodium* suggests that at least within the clade Poaceae (grasses), conserved noncoding sequences retain the same ancestral functions.

Conserved gene regulatory function among the grasses [a clade that originated an estimated 45-60 million years ago (The International Brachypodium Initiative 2010)] creates the opportunity to use individual polyploid species to inform our understanding of gene regulation in all grass lineages. Maize is currently the only sequenced grass genome that contains a whole genome duplication post-dating the divergence of the grass lineages. In the near future additional polyploid grass genomes (tef, switchgrass, and wheat) will become available. If significant amounts of RNA-seq expression data are made publicly available for these lineages, they will provide opportunities to investigate the function of additional conserved noncoding sequences not informed on by the ongoing fractionation of the maize genome.

**Conclusion**

It may turn out that the majority of genes already missing a portion of their ancestral regulatory sequence are functionally redundant with intact gene copies in maize, would produce no phenotype if knocked out, and are already fated for deletion. However, for geneticists interested in understanding how expression of different genes are regulated, these same genes – essentially mutants with wild type controls present in the same cells) represent priceless opportunities to bring some of the precision of the genetic investigation of the function of individual sequences to the traditionally "mile wide and inch deep" field of genomics. Fractionation mutagenesis is both higher throughput than traditional promoter bashing and addresses function more directly than approaches purely association based approaches to promoter characterization.

While the grasses, with maize acting as the deletion machine, are an excellent system for studying gene regulation by fractionation mutagenesis, there is no reason deletion machines should be confined to the grasses. The legumes and crucifers are also both represented by dense clusters of species with sequenced genomes adequately diverged for the discovery of conserved noncoding sequences, and each include at least one sequenced polyploid lineage. Given the lower rate of sequence divergence observed among vertebrate genomes, it may also be possible to modify our technique to computationally infer the function of vertebrate conserved noncoding elements by tracking the expression patterns of duplicated genes created by the 3R whole genome duplication in the teleost fish lineage.

## Methods:

**Identification of orthologous genes:**

Orthologous genes were identified using the combination of syntenic block identification and aggregate synonymous substitution rate analysis previously described (Schnable *et al.* 2012b).

**Identification of conserved noncoding sequences:**

Pairwise orthologous CNS were identified in comparisons between the rice genome (MSU 6.1) (Ouyang *et al.* 2007) and each of sorghum (Sbi1.4) (Paterson *et al.* 2009), setaria (Phytozome 2.1) (Bennetzen *et al.* 2012), and brachypodium (1.0) (The International Brachypodium Initiative 2010) using the CNS Discovery Pipeline 3.0 (Turco et al, *in prep*). Pan-grass CNS were defined as groups of CNS representing all three pairwise comparisons with overlapping genomic positions in rice. Tracing the fate of pan-grass CNS into maize for gene pairs identified as carrying candidate pollen enhancer sequences was performed manually using GEvo (Lyons *et al.* 2008a), part of the CoGe toolkit, the B73 RefGen_v2 maize genome assembly (Schnable *et al.* 2009), and a special version of the sorghum genome carrying annotations for pan-grass CNS as well as the Sbi1.4 gene models.

**Expression data**
Raw RNA-seq data was downloaded from the sequence read archive for six published papers (Wang *et al.* 2009; Li *et al.* 2010; Davidson *et al.* 2011; Waters *et al.* 2011; Bolduc *et al.* 2012; Chang *et al.* 2012) as well as project SRP006965 from the Maize Gametophyte project. Data was processed using GSNAP (Wu and Nacu 2010), SAMtools (Li *et al.* 2009), and Cufflinks (Trapnell *et al.* 2010). All visualization of gene expression data was carried out using qTeller https://github.com/jschnable/qTeller.

**Gene pair expression categories**
Triplets of sorghum, maize1 and maize2 genes are provided as supplemental dataset S2. Maize1 and maize2 are defined using the previously published subgenomes of maize (Schnable *et al.* 2011b). First, datasets where both genes were expressed at less than 1/10th that genes maximum level of expression in the total database were excluded from the analysis. If both genes showed an average of < 1 FPKM in expression after the above exclusions the gene pair was classified as “dead both.” It it likely that many of these apparently dead genes are simply expressed in tissues not included in the database or only under environmental conditions not encountered by the plants in the database. Among the remaining genes, if one gene copy showed an average expression which was less than either 1 FPKM or 1/10 the average expression of the other gene copy the gene was classified as dead1 or dead2 (depending on the subgenome of the non-expressed gene). For the remaining gene pairs, Spearman's rho was calculated using the datasets not excluded based on low expression of both gene pairs. Gene pairs with a spearman's rho greater than .65 were classified as “correlated.” Gene pairs with a spearman's rho less than -.65 were classified as “inversely correlated.” Genes in between these two values where classified as “not correlated.” Genes from the first category were further subdivided into “correlated maize1” if the average expression of maize1 was at least two times the average of maize2 with “correlated maize2” having the inverse definition. Genes in between these two extremes were classified as “correlated even.”

**Identification of pollen specific genes**
A gene showing expression in pollen was defined as a case where the average expression in Pollen1 (SRP006965 Maize Gametophyte Project), Pollen2 and Anthers (Davidson *et al.* 2011) was at least 2-fold greater than the average expression of Shoots field1-3 (SRP006965 Maize Gametophyte Project), Seedling leaves field, and Seedling leaves growth chamber (Davidson *et al.* 2011) and the minimum value of the pollen and anther datasets was greater than the maximum value of the shoots and seedling leaves datasets. Verification of the ancestral nature of pollen specific expression was carried out manually in qTeller though the comparison of anther and vegetative tissue expression from syntenic orthologous genes in rice and sorghum, using the gene expression data for these species published by the Buell lab (Davidson *et al.* 2012) and processed using the qTeller RNA-seq quantification pipeline (https://github.com/jschnable/qTeller).

**CNS Enhancer Cloning and Vector Construction**
Candidate enhancer sequenced identified in silico were PCR-amplified from *Sorghum bicolor* genomic DNA using Promega GoTaq® DNA polymerase (Promega Corporation, Madison, WI). The two (2) CNSs predicted to control pollen/anther-specific expression in the maize gene GRMZM2G014240 (Sb10g005250) were amplified using the forward primer 240F1: 5'-GGGCTTTGGCTTTGGATGCTTGTA-3' and the reverse primer 240R1: 5'-ACACGTGAGTGACAGATGGCAGAA-3'. The resulting 367bp product was cloned into TOP10 electrocompetent *E. coli* using the Invitrogen TOPO TA cloning kit (Life Technologies, Carlsbad, CA). Cloned CNSs were confirmed through sequencing with an ABI3730 DNA Analyzer (Life Technologies, Carlsbad, CA).

For *Brachypodium* whole-plant transformation, CNS enhancer sequences were cloned into a modified version of the pGPro8 (GenBank JN593327) binary expression vector which is a derivative of pGPro1 (Thilmony *et al.* 2006), with the rice *ubiquitin2* promoter driving the *hptII* hygromycin resistance and *GUSPlus* as the reporter gene. The pGPro8 vector was modified by restriction digestion with EcoRI and NcoI to insert a 35S minimal promoter (35SMin) sequence just proximal to the *GUSPlus* coding region (Broothaerts *et al.* 2005). The resulting vector was named pGPro8-35SMin and used to evaluate CNS enhancer candidates in transgenic *Brachypodium distachyon* plants.

Cloned CNSs were digested from TOPO vectors with EcoRI and then inserted into the corresponding EcoRI site of the pGPro8-35SMin vector, forming a transcriptional fusion with the 35S minimal promoter sequence. All binary plasmid vectors were transformed into *A. tumefaciens* strain AGL1 containing the helper plasmid pSoup from the pGreen series (Hellens *et al.* 2000) to use for plant transformation.

**Agrobacterium-mediated Transformation**
The *Agrobacterium*-mediated transformation protocol used to generate transgenic Bd21-3 plants with potential CNS enhancer sequences was optimized from the protocol previously described (Vogel and Hill 2008). Embryos (0.3–0.7 mm) were dissected from immature seeds and transferred to callus initiation media (CIM, per L: 4.43 g Linsmaier & Skoog basal medium (Phytotechnology, Shwanee Mission, KS #L689), 30 g sucrose, 1 ml 0.6 mg/ml CuSO4, pH 5.8. For plates, add 2 g phytagel (Sigma #P-8169). After autoclaving, add 0.5 ml of 5 mg/ml 2,4-D stock solution.) Following 3–4 weeks incubation in the dark at 28ºC, embryogenic callus was subcultured onto fresh CIM plates. A second subculture was performed after two more weeks. The calluses from the second subculture were grown for one week before being used for transformation. On the day of transformation, calluses were bathed for 5 minutes in a suspension of *A. tumefaciens* strain AGL1 containing the helper plasmid pSoup and the desired pGPro8-35SMin vector with CNS sequence. This suspension (OD600 = 0.6) was prepared in liquid CIM containing 200 mM 2,4-D and 0.1% Synperonic PE/F68 (Sigma #81112,

formerly Pluronic F68). After removing as much of the *Agrobacterium* suspension as possible, the calluses were transferred to petri dishes containing a piece of sterile filter paper for co-cultivation for 3 days in the dark at 22ºC. Note that co-cultivation under desiccating conditions is critical to the success of the transformation protocol. Next the callus pieces were moved to CIM plates containing 150 mg/L timentin and 40 mg/L hygromycin B (Phytotechnology H397), and incubated in the dark at 28ºC for 1 week. Healthy sectors of hygromycin resistant transgenic callus were subcultured to fresh CIM plates one time for an additional two weeks of selection. Around 3 weeks after co-cultivation, calluses were transferred to regeneration media (per L: 4.43 g Linsmaier & Skoog (LS) basal medium, 30 g maltose, 2 g phytagel, pH 5.8; after autoclaving, 1.0 ml of sterile 0.2 mg/ml kinetin stock solution was added) containing 150 mg/L timentin and 40 mg/L hygromycin. Plates were incubated in the light (cool-white fluorescent lighting at a level of 65 $\mu Em^{-2}\ s^{-1}$ with a 16 hr light : 8 hr dark cycle) at 28ºC. Callus pieces began to turn green and shoots appeared between 2–4 weeks. Individual $T_0$ plantlets were moved to tissue culture boxes (we used sundae cups made for food service applications from Solo Corporation, Lake Forest, IL Cat. # SOL-TS5 (cups) and SOL-DL-100 (dome lids)) containing MS sucrose medium (per L: 4.42 g Murashige & Skoog (MS) basal medium with vitamins (Phytotechnology M519), 30 g sucrose, and 2 g phytagel, pH 5.7) and incubated in the light (coolwhite fluorescent lighting at a level of 65 $\mu Em^{-2}\ s^{-1}$ with a 16 hr light : 8 hr dark cycle) at 28ºC. After $T_0$ plantlets had formed roots and were approximately 2–5 cm tall, they were transplanted to soil and placed in a growth chamber for flowering (20 hr light, 4 hr dark, 24ºC during the day and 18ºC at night, cool-white fluorescent lighting at a level of 150 $\mu Em^{-2}\ s^{-1}$).

**GUS Histochemical Staining**

Detection of ß-glucuronidase activity in various plant parts was conducted as previously described (Jefferson 1987; Rueb and Hensgens 1989) in a GUS staining solution (0.1M sodium phosphate pH 7.0, 0.5mM potassium ferrocyanide, 0.5mM potassium ferricyanide, 1.5g/L of X-Gluc (5-bromo-4-chloro-3-indolyl-ß-D-glucuronic acid), 0.5% (v/v) Triton X-100) and incubated overnight at 37ºC. Whole flowers, anthers, leaves and stems were dissected from transgenic Bd21-3 plants and viewed under bright field conditions using an Olympus BX51 microscope with a DP70 CCD digital camera (Olympus, Melville, NY).

**Plant growth conditions**

*Brachypodium distachyon* inbred line Bd21-3 was used for all transgenic plant analyses. Plants were grown in a soil mix of 1 part sandy loam, 2 parts sand, 3 parts peat moss, and 3 parts medium grade (#3) vermiculite. A time release fertilizer containing micronutrients (Osmocote Plus 15-9-12, Scotts Co., Marysville, OH) was added at the time of planting. Plants were grown in both greenhouses and growth chambers. Growth chambers conditions were 20 hr light : 4 hr dark photoperiod, cool-white fluorescent lighting at a level of 150 $\mu Em^{-2}\ s^{-1}$, and temperatures of 24ºC during the day and 18ºC at night.

Greenhouse conditions were no shading, 24ºC in the day and 18ºC at night with the day length extended to 16 hours by supplemental lighting.

**Figures:**

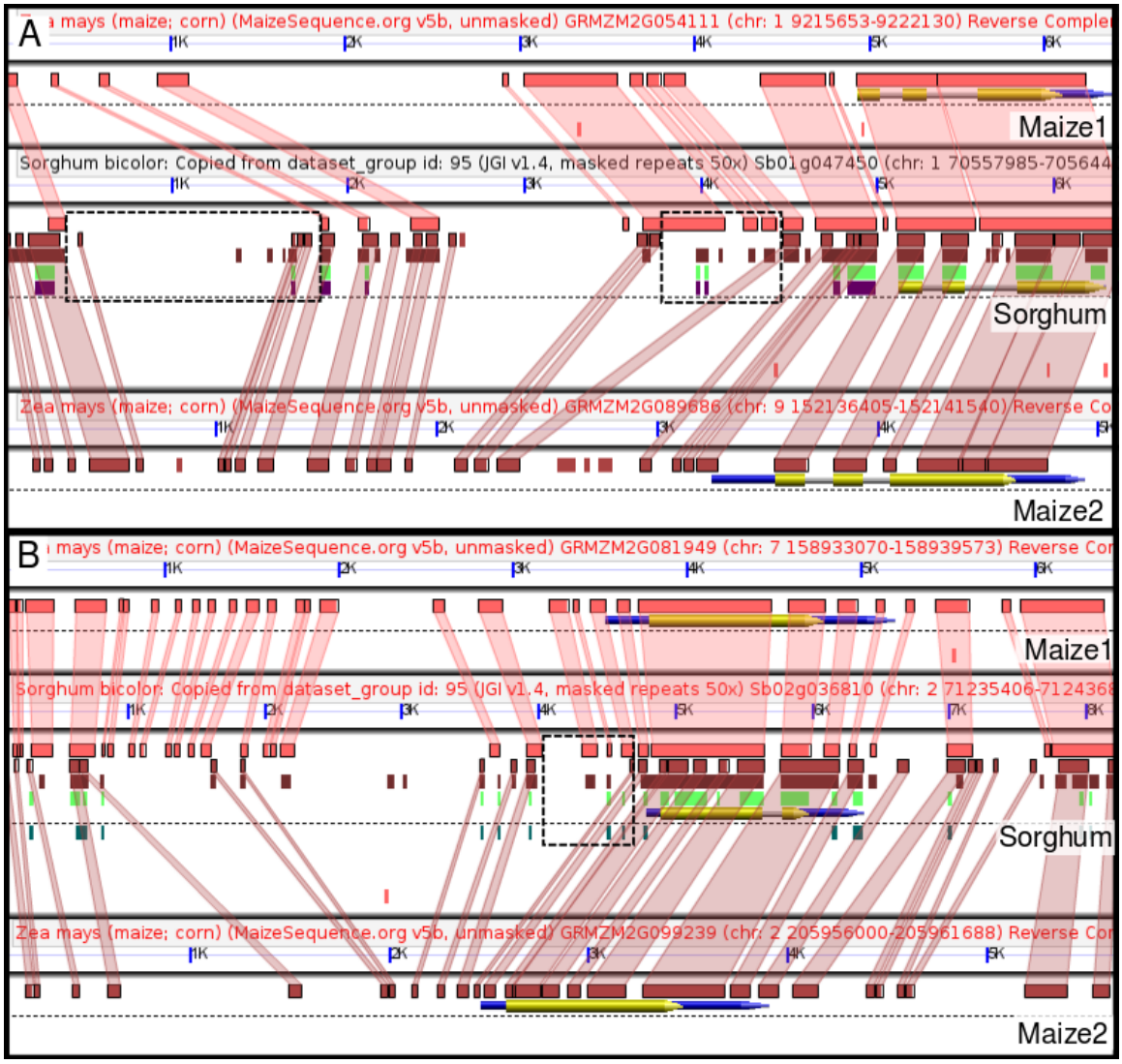



Figure 1: Two GEvo panels (Lyons et al 2008) showing examples of reciprocal and one-sided deletion of conserved noncoding sequences. Protein coding exons are shown as yellow rectangles and arrows and untranslated regions as blue rectangles and arrows. Colored boxes connected by shaded regions mark sequenced identified as homologous by BLASTN between sorghum and the two duplicate homeologs in maize. From top to bottom in each sorghum panel colored boxes represent sequences with detectable sequence conservation when compared to the orthologous region in maize1, maize2, setaria, and rice. The final set of colored boxed (purple in A and green in B) mark conserved noncoding sequences annotated by the CNS Discovery Pipeline (Turco in

prep). Dashed boxes mark promoter regions in sorghum which contain at least one conserved noncoding sequence which has been lost from the region surrounding one maize homeolog.

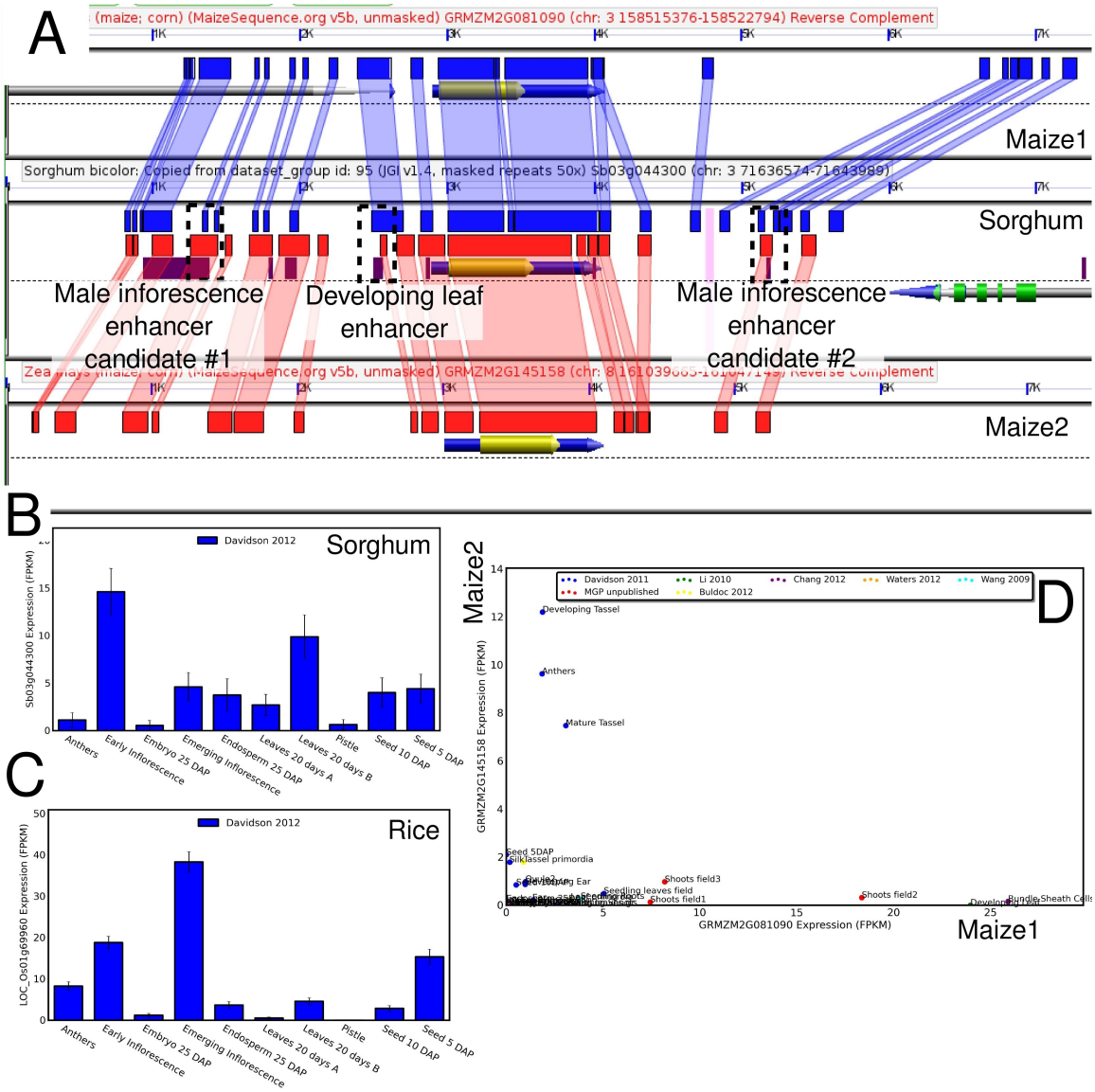



Figure 2: An example of how reciprocal expression pattern of a duplicate gene pair in maize may be combined with deletions of conserved noncoding sequences to develop hypotheses regarding the function of individual conserved noncoding sequences. A) A GEvo panel comparing the sequence of a sorghum gene (Sb03g044300) to its two co-orthologs in maize (GRMZM2G081090 and GRMZM2G145158). Conserved noncoding sequences are marked by purple boxes. CNS deleted from one of the region adjacent to one of the two maize genes are marked by dashed boxes. B & C) qTeller bar charts showing the expression of the sorghum gene and its rice syntenic ortholog (LOC_Os01g69960). Note that peak expression occurs in emerging inflorescence in rice

and early inflorescence in sorghum, however this likely reflects the difficulty of selecting developmentally equivalent timepoints between species rather than a shift in gene expression. D) qTeller scatterplot comparing the expression patterns of both gene copies in maize. The maize2 copy retains the ancestral inflorescence expression pattern (revealed to be male specific in maize where seperate imperfect flowers are born on specialized male and female inflorescences) and the maize1 copy is expressed in vegetative tissue which may also constitute a portion of the ancestral expression domain (supported by sorghum but not rice).

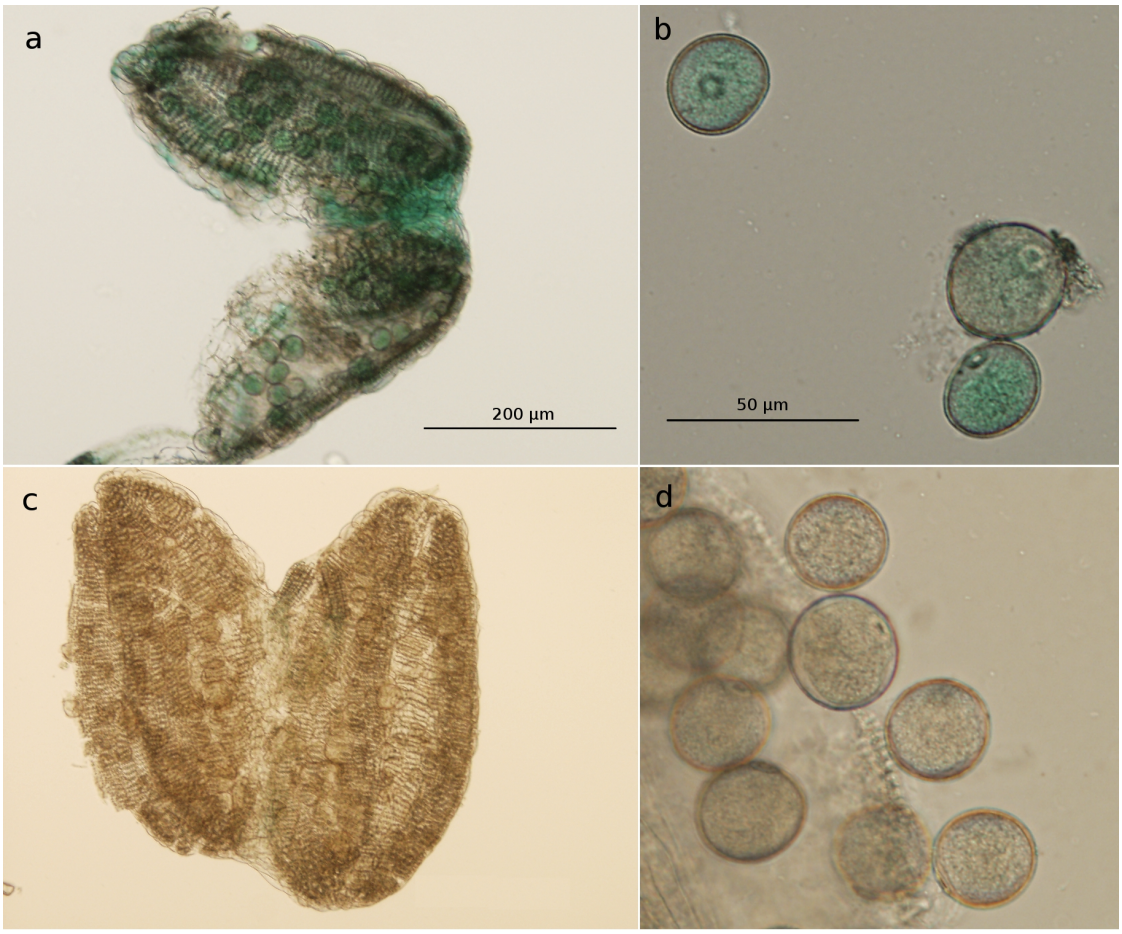



Figure 3: Expression of GUS driven by a minimal 35S promoter plus a putative pollen/anther enhancer from Sb10g005250 identified by fractionation mutagenesis. A) Anthers from transgenic brachypodium carrying the reporter construct. B) Individual pollen grains from the same transgenic line. C) Anthers from a transgenic brachypodium line carrying the GUS reporter construct with the minimal 35S promoter but lacking the enhancer from Sb10g005250. D) Individual pollen grains from the same plant.

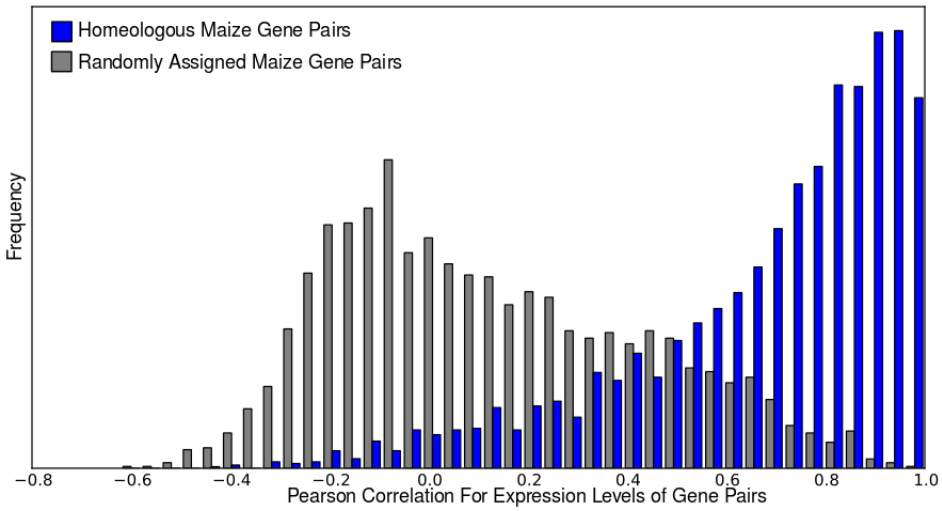



Figure S1: Comparison of the level of correlation of expression pattern observed for homeologous maize gene pairs and, as a control, randomly assigned gene pairs.

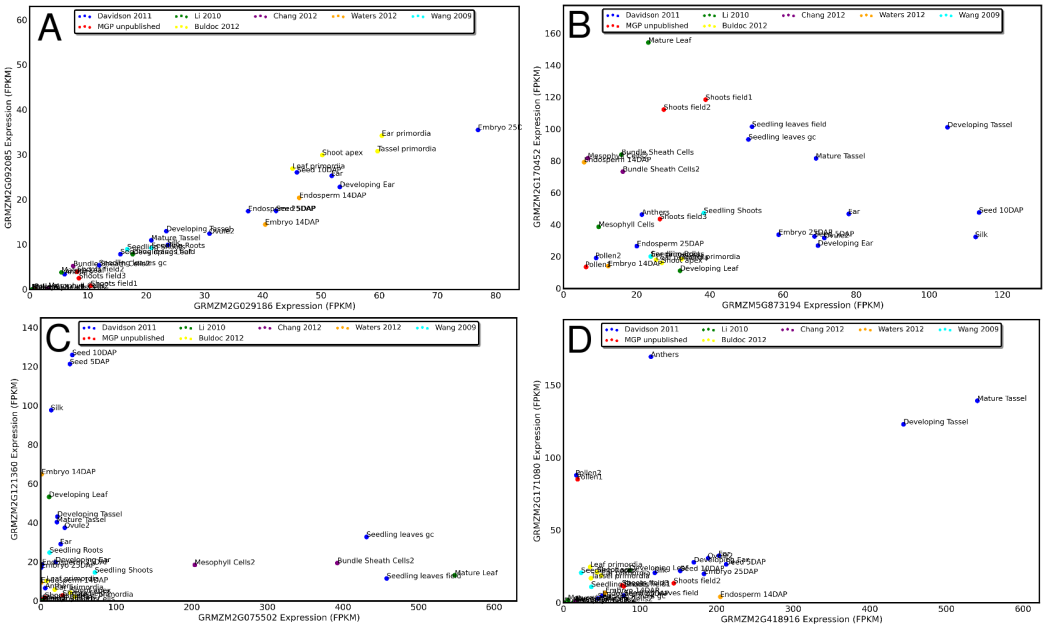



Figure S2: Four qTeller panels summarizing different categories into which patterns of gene expression between duplicate homeologs in maize may be classified. A) Expression of both gene copies is largely correlated across all datasets. Note that the absolute level of expression is often quite different even in gene pairs where the pattern of expression is highly correlated. In this example the difference in expression level is roughly 2-fold. B) Uncorrelated expression. In this example the highest and lowest observed expression levels of both gene copies are comparable, but the pattern of expression for the two genes is wildly different. C) Reciprocal expression of gene copies. Here one gene copy is expressed to high levels in mature leaf tissue (seedling leaves and “Mature Leaf” datasets) while the other gene shows higher expression in developing seeds, silks, and tassels. D) Outlier tissue. In this case the expression of the two gene copies are roughly correlated in most tissues, but only one copy shows high levels of expression in pollen and (to a lesser extent) anthers).

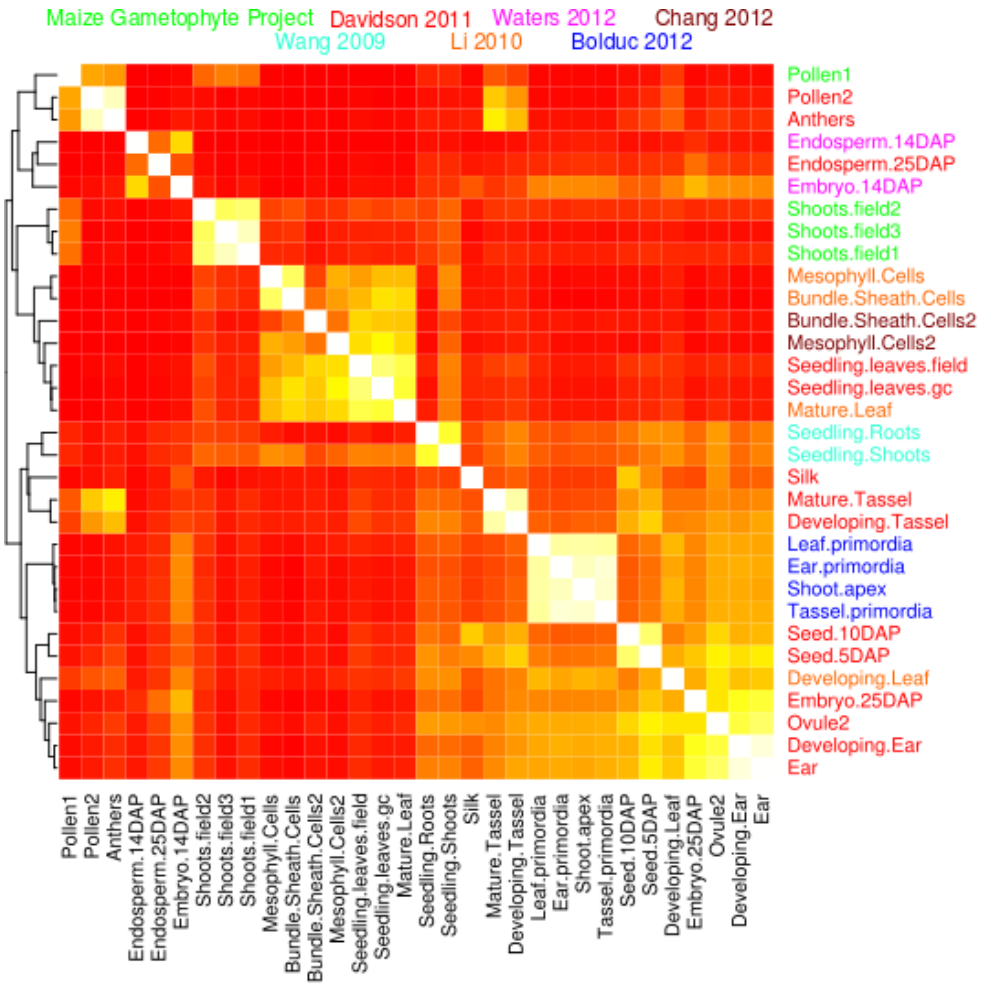



Figure S3: Heatmap and denodrogram reflecting the spearman's rho correlation of gene expression patterns among all the maize expression datasets currently available within qTeller.


Figure S4: Lack of GUS expression from the pollen/anther enhancer construct in vegetative tissue. Some possible expression at the floret base.

## Chapter 9: Wrap Up

At the birth of maize genetics a mere century ago geneticists did their research in a world that was, by todays standards, strikingly data poor. The result of each test cross represented the hours of long fieldwork and setting up an informative experiment might require years of effort making the correct crosses to create the required genetic lines. Today researchers live in a world of easy and abundant data (see Chapter 1). Data, by itself, cannot substitute for intellectually rigorous experiments, but with the proper experimental design abundant data can greatly speed up the business of developing and testing hypotheses. This dissertation provides two examples of how real biological questions can be addressed through the use of existing genomic and gene expression datasets.

The first question I sought to address was reason for the observed bias in the deletion of gene copies following whole genome duplications. The identification of this bias was possible using a single genome assembly (Thomas *et al.* 2006), yet identifying the mechanism responsible for the bias required comparisons between the sequenced genomes of three grass species (rice, sorghum, and maize) and the analysis of gene expression data generated by multiple research groups. The final model I was able to propose to explain the biased gene loss grew out of an observation of unequal expression between duplicated genes from different parental subgenomes which was first reported in tetraploid cotton (Flagel *et al.* 2008; Flagel and Wendel 2010).

Developing models which fit known observations is an important step in the scientific process but to be truly rigorous it is also necessary to test the predictions of those models. In Chapter 3 I took advantage of a dataset likely unique to maize, the avaliability of characterized mutants accumulated over the last century of maize genetics to test the prediction of my model that genes where knockouts result in noticeable phenotypes should be disproportionately found on the dominant subgenome within maize. In addition to conforming to the prediction of my model, this analysis created a dataset of historically studied “classical” maize genes which has been widely used by the broader maize genetics community.

The ability to accurately identify and classify orthologous or homeologous genes among related species is a necessary foundation to most comparative genetics projects which attempt to take advantage of the wealth of sequenced plant genomes now available (Chapter 1, Figure 1). In Chapter 4 I outline the approach I employed for this task throughout the research described in this dissertation. The combination of synteny and synonymous substitution rate analysis is limited to comparisons between species where synteny has not been scrambled by too many whole genome duplications and synonymous substitutions have not yet saturated. However, within this window of useful comparison my combined approach is both precise and accurate and also provides sets of

putative pseudogenes and high confidence orthologous gene deletions which are not identified by transitional sequence-based or tree-based methods for identifying orthology.

Rigorous scientific investigation also depends on the ability to perform replicated experiments, a trait *in silico* genomic investigation is often unable to provide. However as I demonstrated in Chapter 5, the increasing number of plant genome sequences available is beginning to address this traditional weak point in genomics. I was able to provide much more support for my conclusions by identifying parallel arrangements of species and whole genome duplications in the grasses and the crucifers and showing the same patterns of gene expression and deletion in both clades. As the number of species with sequenced genomes continue to increase I expect that this approach with become commonplace for many types of genomic research.

Another aspect which risks being lost in the transition from genetics to genomic research are opportunities for fortuitous discovery. Analysis of large datasets require computational tools yet performing automated analyses at the command line limits the chances of spotting completely unexpected patterns in data. While computers are powerful tools they remain limited to identifying patterns a user has previously programmed them to look for. The human brain is still the best tool a scientist possesses for spotting unexpected patterns in datasets. In this age of abundant data it remains important to develop tools that present data in forms the human brain can work with. The analytical pipeline and web interface I developed and describe in Chapter 7, qTeller, is an attempt to enable these sorts of fortuitous discoveries by enabling researchers from around the world to pull out and visualize comparable expression values for their favorite from a wide range of experiments without first having to become an expert at bioinformatic techniques for dealing with RNA-seq datasets.

Finally, a key question in the investigation of many genes continues to be which sequences determine where the gene will be turned off and where the gene will be turned on. In Chapter 8 I demonstrated how a number of different pieces could can be combined to address this question at the level of individual sequences regulating individual genes under specific sets of environmental and/or developmental conditions. These pieces include:

1) Pairs of genes in maize or other ancient polyploid species
2) Sequence deletions of the sort described in our previously published PLoS Biology paper
3) The functionally constrained conserved noncoding sequences discussed in Chapter 6.
4) Existing RNA-seq gene expression datasets
5) The RNA-seq processing and visualization tools which make up qTeller described in Chapter 7.

Putting all these pieces together allows researchers to to use polyploid species with

plentiful RNA-seq datasets as “deletion machines” to understand the function of conserved regulatory sequences shared by orthologous genes in many species. As a result wet lab biologists can develop specific and testable hypotheses about the function of individual regulatory sequences before they ever perform a single PCR reaction or create a transgenic tester line, result in faster and more efficient characterization of developmental or environmental specific enhancers and repressors of gene expression.

**Conclusion**

The research described in this document has contributed the answers to specific biological questions such as the link between biased gene loss and genome dominance or the identification of a novel pollen/anther specific enhancer. However, it is my hope that what anyone reading my dissertation will take these as examples to demonstrate that genomics research does not need to be purely associative. Instead, developing testable hypotheses, and in some cases even testing them is possible *in silico*. As biological data becomes ever cheaper and more abundant the ability to combine different analytical approaches and different types of data in unexpected ways will become an even more important part of genetics and genomics.

**Bibliography**


Acosta I. F., Laparra H., Romero S. P., Schmelz E., Hamberg M., Mottinger J. P., Moreno M. A., Dellaporta S. L., 2009 tasselseed1 Is a Lipoxygenase Affecting Jasmonic Acid Signaling in Sex Determination of Maize. Science **323**: 262–265.

Alexandrov N. N., Brover V. V., Freidin S., Troukhan M. E., Tatarinova T. V., Zhang H., Swaller T. J., Lu Y.-P., Bouck J., Flavell R. B., Feldmann K. A., 2009 Insights into corn genes derived from large-scale cDNA sequencing. Plant Mol. Biol **69**: 179–194.

Barker M. S., Kane N. C., Matvienko M., Kozik A., Michelmore R. W., Knapp S. J., Rieseberg L. H., 2008 Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years. Mol Biol Evol **25**: 2445 –2455.

Baucom R. S., Estill J. C., Chaparro C., Upshaw N., Jogi A., Deragon J.-M., Westerman R. P., SanMiguel P. J., Bennetzen J. L., 2009 Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. PLoS Genet **5**: e1000732.

Bekaert M., Edger P. P., Pires J. C., Conant G. C., 2011 Two-Phase Resolution of Polyploidy in the Arabidopsis Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints. Plant Cell **23**: 1719–1728.

Bennetzen J. L., 2007 Patterns in grass genome evolution. Curr Opin Plant Biol **10**: 176–181.

Bennetzen J. L., Freeling M., 1993 Grasses as a single genetic system: genome composition, collinearity and compatibility. Trends Genet **9**: 259–261.

Bennetzen J. L., Schmutz J., Wang H., Percifield R., Hawkins J., Pontaroli A. C., Estep M., Feng L., Vaughn J. N., Grimwood J., Jenkins J., Barry K., Lindquist E., Hellsten U., Deshpande S., Wang X., Wu X., Mitros T., Triplett J., Yang X., Ye C.-Y., Mauro-Herrera M., Wang L., Li P., Sharma M., Sharma R., Ronald P. C., Panaud O., Kellogg E. A., Brutnell T. P., Doust A. N., Tuskan G. A., Rokhsar D., Devos K. M., 2012 Reference genome sequence of the model plant Setaria. Nat. Biotechnol. **30**: 555–561.

Bikard D., Patel D., Metté C. Le, Giorgi V., Camilleri C., Bennett M. J., Loudet O., 2009 Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science **323**: 623–626.

Birchler J. A., Riddle N. C., Auger D. L., Veitia R. A., 2005 Dosage balance in gene regulation: biological implications. Trends Genet **21**: 219–226.

Birchler J. A., Veitia R. A., 2007 The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. Plant Cell **19**: 395 –402.

Birchler J. A., Veitia R. A., 2010 The Gene Balance Hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol **186**: 54–62.

Birchler J. A., Yao H., Chudalayandi S., 2007 Biological consequences of dosage dependent gene regulatory systems. Biochim. Biophys. Acta **1769**: 422–428.

Blanc G., Wolfe K. H., 2004 Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. Plant Cell **16**: 1679–1691.

Bodt S. De, Maere S., Peer Y. Van de, 2005 Genome duplication and the origin of angiosperms. Trends Ecol Evol **20**: 591–597.

Bolduc N., Yilmaz A., Mejia-Guerra M. K., Morohashi K., O'Connor D., Grotewold E., Hake S., 2012 Unraveling the KNOTTED1 regulatory network in maize meristems. Genes Dev. **26**: 1685–1690.

Bomblies K., Doebley J. F., 2005 Molecular Evolution of FLORICAULA/LEAFY Orthologs in the Andropogoneae (Poaceae). Mol Biol Evol **22**: 1082–1094.

Bowers J. E., Chapman B. A., Rong J., Paterson A. H., 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**: 433–438.

Broothaerts W., Mitchell H. J., Weir B., Kaines S., Smith L. M. A., Yang W., Mayer J. E., Roa-Rodríguez C., Jefferson R. A., 2005 Gene transfer to plants by diverse species of bacteria. Nature **433**: 629–633.

Buggs R. J., Chamala S., Wu W., Gao L., May G. D., Schnable P. S., Soltis D. E., Soltis P., Barbazuk W. B., 2010 Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol Ecol **19**: 132–146.

Buggs R. J. A., Zhang L., Miles N., Tate J. A., Gao L., Wei W., Schnable P. S., Barbazuk W. B., Soltis P. S., Soltis D. E., 2011 Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr. Biol. **21**: 551–556.

Cannon S. B., Mitra A., Baumgarten A., Young N. D., May G., 2004 The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol **4**: 10.

Chang P. L., Dilkes B. P., McMahon M., Comai L., Nuzhdin S. V., 2010 Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol **11**: R125.

Chang Y.-M., Liu W.-Y., Shih A. C.-C., Shen M.-N., Lu C.-H., Lu M.-Y. J., Yang H.-W., Wang T.-Y., Chen S. C.-C., Chen S. M., Li W.-H., Ku M. S. B., 2012 Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol. **160**: 165–177.

Chen Z. J., 2007 Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu Rev Plant Biol **58**: 377–406.

Chen S., Zhang Y. E., Long M., 2010 New Genes in Drosophila Quickly Become Essential. Science **330**: 1682 –1685.

Davidson R. M., Gowda M., Moghe G., Lin H., Vaillancourt B., Shiu S.-H., Jiang N., Robin Buell C., 2012 Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J **71**: 492–502.

Davidson R. M., Hansey C. N., Gowda M., Childs K. L., Lin H., Vaillancourt B., Sekhon R. S., Leon N. de, Kaeppler S. M., Jiang N., Buell C. R., 2011 Utility of RNA Sequencing for Analysis of Maize Reproductive Transcriptomes. Plant Genome **4**: 191–203.

Dehal P., Boore J. L., 2005 Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol **3**: e314.

Deng X., Gu L., Liu C., Lu T., Lu F., Lu Z., Cui P., Pei Y., Wang B., Hu S., Cao X., 2010 Arginine methylation mediated by the Arabidopsis homolog of PRMT5 is essential for proper pre-mRNA splicing. Proc Natl Acad Sci USA **107**: 19114–19119.

D'Hont A., Denoeud F., Aury J.-M., Baurens F.-C., Carreel F., *et al.*, 2012 The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature **488**: 213–217.

Duarte J. M., Wall P. K., Edger P. P., Landherr L. L., Ma H., Pires J. C., Leebens-Mack J., dePamphilis C. W., 2010 Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol Biol **10**: 61.

Dugas D. V., Monaco M. K., Olsen A., Klein R. R., Kumari S., Ware D., Klein P. E., 2011 Functional Annotation of the Transcriptome of Sorghum bicolor in Response to Osmotic Stress and Abscisic Acid. BMC Genomics **12**: 514.

EDGER P. P., PIRES J. C., 2009 Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res **17**: 699–717.

EMERSON R. A., 1920 Heritable Characters of Maize II.-Pistillate Flowered Maize Plants. J Hered **11**: 65–76.

EVELAND A. L., SATOH-NAGASAWA N., GOLDSHMIDT A., MEYER S., BEATTY M., SAKAI H., WARE D., JACKSON D., 2010 Digital Gene Expression Signatures for Maize Development. Plant Physiol **154**: 1024–1039.

FLAGEL L. E., UDALL J., NETTLETON D., WENDEL J., 2008 Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. BMC Biol **6**: 16.

FLAGEL L. E., WENDEL J. F., 2010 Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol **186**: 184–193.

FORCE A., LYNCH M., PICKETT F. B., AMORES A., YAN Y., POSTLETHWAIT J., 1999 Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics **151**: 1531–1545.

FREELING M., 2009 Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol **60**: 433–453.

FREELING M., LYONS E., PEDERSEN B., ALAM M., MING R., LISCH D., 2008 Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Res **18**: 1924 –1937.

FREELING M., RAPAKA L., LYONS E., PEDERSEN B., THOMAS B. C., 2007 G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis. Plant Cell **19**: 1441–1457.

FREELING M., SUBRAMANIAM S., 2009 Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol **12**: 126–132.

FREELING M., THOMAS B. C., 2006 Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res **16**: 805–814.

FREELING M., WOODHOUSE M. R., SUBRAMANIAM S., TURCO G., LISCH D., SCHNABLE J. C., 2012 Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol **15**: 131–139.

GAETA R. T., PIRES J. C., INIGUEZ-LUY F., LEON E., OSBORN T. C., 2007 Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. Plant Cell **19**: 3403 –3417.

GAUT B. S., DOEBLEY J. F., 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci USA **94**: 6809 –6814.

GOFF S. A., RICKE D., LAN T.-H., PRESTING G., WANG R., *et al.*, 2002 A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). Science **296**: 92 –100.

GRANT G. R., FARKAS M. H., PIZARRO A. D., LAHENS N. F., SCHUG J., BRUNK B. P., STOECKERT C. J., HOGENESCH J. B., PIERCE E. A., 2011 Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM). Bioinformatics **27**: 2518–2528.

GRASS PHYLOGENY WORKING GROUP, 2001 Phylogeny and Subfamilial Classification of the Grasses (Poaceae). Ann Mo Bot Gard **88**: 373–457.

GROVER C. E., GALLAGHER J. P., SZADKOWSKI E. P., YOO M. J., FLAGEL L. E., WENDEL J. F., 2012 Homoeolog expression bias and expression level dominance in allopolyploids. New Phytologist: n/a–n/a.

GUMUCIO D. L., HEILSTEDT-WILLIAMSON H., GRAY T. A., TARLÉ S. A., SHELTON D. A., TAGLE D. A., SLIGHTOM J. L., GOODMAN M., COLLINS F. S., 1992 Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. Mol. Cell. Biol. **12**: 4919–4929.

GUMUCIO D. L., SHELTON D. A., BAILEY B. A., SLIGHTOM J. L., GOODMAN M., 1993 Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. PNAS **90**: 6018–6022.

GUMUCIO D. L., SHELTON D. A., ZHU W., MILLINOFF D., GRAY T., BOCK J. H., SLIGHTOM J. L., GOODMAN M., 1996 Evolutionary Strategies for the Elucidation ofcisandtransFactors That Regulate the Developmental Switching Programs of the β-like Globin Genes. Molecular Phylogenetics and Evolution **5**: 18–32.

GUMUCIO D. L., WIEBAUER K., CALDWELL R. M., SAMUELSON L. C., MEISLER M. H., 1988 Concerted evolution of human amylase genes. Mol. Cell. Biol. **8**: 1197–1205.

GUO H., MOOSE S. P., 2003 Conserved Noncoding Sequences among Cultivated Cereal Genomes Identify Candidate Regulatory Sequence Elements and Patterns of Promoter Evolution. Plant Cell **15**: 1143–1158.

GU Z., STEINMETZ L. M., GU X., SCHARFE C., DAVIS R. W., LI W.-H., 2003 Role of

duplicate genes in genetic robustness against null mutations. Nature **421**: 63–66.

HARDISON R. C., 2000 Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. **16**: 369–372.

HARRIS R. S., 2007 Improved Pairwise Alignment of Genomic Data.

HELLENS R. P., EDWARDS E. A., LEYLAND N. R., BEAN S., MULLINEAUX P. M., 2000 pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. Plant Mol. Biol. **42**: 819–832.

HU T. T., PATTYN P., BAKKER E. G., CAO J., CHENG J.-F., CLARK R. M., FAHLGREN N., FAWCETT J. A., GRIMWOOD J., GUNDLACH H., HABERER G., HOLLISTER J. D., OSSOWSKI S., OTTILAR R. P., SALAMOV A. A., SCHNEEBERGER K., SPANNAGL M., WANG X., YANG L., NASRALLAH M. E., BERGELSON J., CARRINGTON J. C., GAUT B. S., SCHMUTZ J., MAYER K. F. X., PEER Y. VAN DE, GRIGORIEV I. V., NORDBORG M., WEIGEL D., GUO Y.-L., 2011 The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet **43**: 476–481.

INADA D. C., BASHIR A., LEE C., THOMAS B. C., KO C., GOFF S. A., FREELING M., 2003 Conserved Noncoding Sequences in the Grasses. Genome Res **13**: 2030 –2041.

JAILLON O., AURY J.-M., NOEL B., POLICRITI A., CLEPET C., *et al.*, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**: 463–467.

JEFFERSON R., 1987 Assaying chimeric genes in plants: The GUS gene fusion system. Plant Molecular Biology Reporter **5**: 387–405.

JIA Y., LISCH D. R., OHTSU K., SCANLON M. J., NETTLETON D., SCHNABLE P. S., 2009 Loss of RNA–Dependent RNA Polymerase 2 (RDR2) Function Causes Widespread and Unexpected Changes in the Expression of Transposons, Genes, and 24-nt Small RNAs. PLoS Genet **5**: e1000737.

JIAO Y., WICKETT N. J., AYYAMPALAYAM S., CHANDERBALI A. S., LANDHERR L., RALPH P. E., TOMSHO L. P., HU Y., LIANG H., SOLTIS P. S., SOLTIS D. E., CLIFTON S. W., SCHLARBAUM S. E., SCHUSTER S. C., MA H., LEEBENS-MACK J., DEPAMPHILIS C. W., 2011 Ancestral polyploidy in seed plants and angiosperms. Nature **473**: 97–100.

KAPLINSKY N. J., BRAUN D. M., PENTERMAN J., GOFF S. A., FREELING M., 2002 Utility and distribution of conserved noncoding sequences in the grasses. Proc Natl Acad Sci USA **99**: 6147 –6151.

KASAHARA M., 2007 The 2R hypothesis: an update. Curr. Opin. Immunol **19**: 547–552.

KASAHARA M., NARUSE K., SASAKI S., NAKATANI Y., QU W., AHSAN B., YAMADA T., NAGAYASU Y., DOI K., KASAI Y., JINDO T., KOBAYASHI D., SHIMADA A., TOYODA A., KUROKI Y., FUJIYAMA A., SASAKI T., SHIMIZU A., ASAKAWA S., SHIMIZU N., HASHIMOTO S., YANG J., LEE Y., MATSUSHIMA K., SUGANO S., SAKAIZUMI M., NARITA T., OHISHI K., HAGA S., OHTA F., NOMOTO H., NOGATA K., MORISHITA T., ENDO T., SHIN-I T., TAKEDA H., MORISHITA S., KOHARA Y., 2007 The medaka draft genome and insights into vertebrate genome evolution. Nature **447**: 714–719.

KAZAN K., MANNERS J. M., 2008 Jasmonate Signaling: Toward an Integrated View. Plant Physiol. **146**: 1459–1468.

KELLIS M., BIRREN B. W., LANDER E. S., 2004 Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**: 617–624.

LAI J., LI R., XU X., JIN W., XU M., ZHAO H., XIANG Z., SONG W., YING K., ZHANG M., JIAO Y., NI P., ZHANG J., LI D., GUO X., YE K., JIAN M., WANG B., ZHENG H., LIANG H., ZHANG X., WANG S., CHEN S., LI J., FU Y., SPRINGER N. M., YANG H., WANG J., DAI J., SCHNABLE P. S., WANG J., 2010 Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet **42**: 1027–1030.

LANGHAM R. J., WALSH J., DUNN M., KO C., GOFF S. A., FREELING M., 2004 Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. Genetics **166**: 935–945.

LANGMEAD B., TRAPNELL C., POP M., SALZBERG S. L., 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10**: R25.

LANG D., WEICHE B., TIMMERHAUS G., RICHARDT S., RIAÑO-PACHÓN D. M., CORRÊA L. G. G., RESKI R., MUELLER-ROEBER B., RENSING S. A., 2010 Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. Genome Biol Evol **2**: 488 – 503.

LARKIN D. M., PAPE G., DONTHU R., AUVIL L., WELGE M., LEWIN H. A., 2009 Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genet Res **19**: 770 –777.

LAWRENCE C. J., DONG Q., POLACCO M. L., SEIGFRIED T. E., BRENDEL V., 2004 MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Res **32**: D393–397.

LAWRENCE C. J., HARPER L. C., SCHAEFFER M. L., SEN T. Z., SEIGFRIED T. E., CAMPBELL D. A., 2008 MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research. Int J Plant Genomics **2008**: 496957.

Lee H.-S., Chen Z. J., 2001 Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. Proc Natl Acad Sci USA **98**: 6753 –6758.

Levine M., 2010 Transcriptional enhancers in animal development and evolution. Curr Biol **20**: R754–763.

Lewis E. B., 1951 Pseudoallelism and gene evolution. Cold Spring Harb Symp Quant Biol **16**: 159–174.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

Lin J.-Y., Stupar R. M., Hans C., Hyten D. L., Jackson S. A., 2010 Structural and functional divergence of a 1-Mb duplicated region in the soybean (Glycine max) genome and comparison to an orthologous region from Phaseolus vulgaris. Plant Cell **22**: 2545–2561.

Li P., Ponnala L., Gandotra N., Wang L., Si Y., Tausta S. L., Kebrom T. H., Provart N., Patel R., Myers C. R., Reidel E. J., Turgeon R., Liu P., Sun Q., Nelson T., Brutnell T. P., 2010 The developmental dynamics of the maize leaf transcriptome. Nat Genet **42**: 1060–1067.

Liu S.-L., Baute G. J., Adams K. L., 2011 Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in Arabidopsis thaliana. Genome Biol Evol.

Li L., Wang X., Stolc V., Li X., Zhang D., Su N., Tongprasit W., Li S., Cheng Z., Wang J., Deng X. W., 2006 Genome-wide transcription analyses in rice using tiling microarrays. Nat Genet **38**: 124–129.

Loots G. G., Locksley R. M., Blankespoor C. M., Wang Z. E., Miller W., Rubin E. M., Frazer K. A., 2000 Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science **288**: 136–140.

Loots G., Ovcharenko I., 2007 ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. Bioinformatics **23**: 122–124.

Lynch M., Force A. G., 2000 The Origin of Interspecific Genomic Incompatibility via Gene Duplication. Am Naturalist **156**: 590–605.

Lyons E., Pedersen B., Kane J., Alam M., Ming R., Tang H., Wang X., Bowers J. E., Paterson A. H., Lisch D., Freeling M., 2008a Finding and Comparing Syntenic

Regions among Arabidopsis and the Outgroups Papaya, Poplar, and Grape: CoGe with Rosids. Plant Physiol **148**: 1772–1781.

LYONS E., PEDERSEN B., KANE J., FREELING M., 2008b The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. Tropical Plant Biol **1**: 181–190.

LYSAK M. A., KOCH M. A., PECINKA A., SCHUBERT I., 2005 Chromosome Triplication Found Across the Tribe Brassiceae. Genome Res **15**: 516–525.

MAERE S., BODT S. DE, RAES J., CASNEUF T., MONTAGU M. VAN, KUIPER M., PEER Y. VAN DE, 2005 Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA **102**: 5454–5459.

MALLET J., 2007 Hybrid speciation. Nature **446**: 279–283.

MARTIN M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10–12.

MASSA A. N., WANJUGI H., DEAL K. R., O'BRIEN K., YOU F. M., MAITI R., CHAN A. P., GU Y. Q., LUO M. C., ANDERSON O. D., RABINOWICZ P. D., DVORAK J., DEVOS K. M., 2011 Gene Space Dynamics during the Evolution of Aegilops tauschii, Brachypodium distachyon, Oryza sativa, and Sorghum bicolor Genomes. Mol Biol Evol **28**: 2537–2547.

MCCLINTOCK B., A Short Biographical Note: Barbara McClintock.

MILLER W., MAKOVA K. D., NEKRUTENKO A., HARDISON R. C., 2004 Comparative genomics. Annu Rev Genomics Hum Genet **5**: 15–56.

MIZUTA Y., HARUSHIMA Y., KURATA N., 2010 Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proc Natl Acad Sci USA **107**: 20417 – 20422.

MOORE G., DEVOS K. M., WANG Z., GALE M. D., 1995 Cereal genome evolution. Grasses, line up and form a circle. Curr Biol **5**: 737–739.

NEI M., GOJOBORI T., 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**: 418–426.

OHNO S., 1970 *Evolution by gene duplication*. Springer-Verlag, Berlin, New York,.

OONO Y., KAWAHARA Y., KANAMORI H., MIZUNO H., YAMAGATA H., YAMAMOTO M., HOSOKAWA S., IKAWA H., AKAHANE I., ZHU Z., WU J., ITOH T., MATSUMOTO T., 2011 mRNA-Seq

Reveals a Comprehensive Transcriptome Profile of Rice under Phosphate Stress. Rice **4**: 50–65.

OSBORN T. C., CHRIS PIRES J., BIRCHLER J. A., AUGER D. L., JEFFERY CHEN Z., LEE H.-S., COMAI L., MADLUNG A., DOERGE R. W., COLOT V., MARTIENSSEN R. A., 2003 Understanding mechanisms of novel gene expression in polyploids. Trends Genet **19**: 141–147.

OUYANG S., ZHU W., HAMILTON J., LIN H., CAMPBELL M., CHILDS K., THIBAUD-NISSEN F., MALEK R. L., LEE Y., ZHENG L., ORVIS J., HAAS B., WORTMAN J., BUELL C. R., 2007 The TIGR Rice Genome Annotation Resource: improvements and new features. Nucl Acids Res **35**: D883–D887.

OWNBEY M., 1950 Natural Hybridization and Amphiploidy in the Genus Tragopogon. Am J Bot **37**: 487–499.

PATERSON A. H., BOWERS J. E., BRUGGMANN R., DUBCHAK I., GRIMWOOD J., GUNDLACH H., HABERER G., HELLSTEN U., MITROS T., POLIAKOV A., SCHMUTZ J., SPANNAGL M., TANG H., WANG X., WICKER T., BHARTI A. K., CHAPMAN J., FELTUS F. A., GOWIK U., GRIGORIEV I. V., LYONS E., MAHER C. A., MARTIS M., NARECHANIA A., OTILLAR R. P., PENNING B. W., SALAMOV A. A., WANG Y., ZHANG L., CARPITA N. C., FREELING M., GINGLE A. R., HASH C. T., KELLER B., KLEIN P., KRESOVICH S., MCCANN M. C., MING R., PETERSON D. G., MEHBOOB-UR-RAHMAN, WARE D., WESTHOFF P., MAYER K. F. X., MESSING J., ROKHSAR D. S., 2009 The Sorghum bicolor genome and the diversification of grasses. Nature **457**: 551–556.

PATERSON A. H., BOWERS J. E., CHAPMAN B. A., 2004 Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101**: 9903–9908.

PATERSON A. H., FREELING M., TANG H., WANG X., 2010 Insights from the Comparison of Plant Genome Sequences. Annu Rev Plant Biol **61**: 349–372.

PEER Y. VAN DE, MAERE S., MEYER A., 2009 The evolutionary significance of ancient genome duplications. Nat Rev Genet **10**: 725–732.

PRASAD V., STRÖMBERG C. A. E., ALIMOHAMMADIAN H., SAHNI A., 2005 Dinosaur coprolites and the early evolution of grasses and grazers. Science **310**: 1177–1180.

RAATZ B., EICKER A., SCHMITZ G., FUSS E., MÜLLER D., ROSSMANN S., THERES K., 2011 Specific expression of LATERAL SUPPRESSOR is controlled by an evolutionarily conserved 3′ enhancer. The Plant Journal **68**: 400–412.

RAMSEY J., SCHEMSKE D. W., 1998 Pathways, Mechanisms, and rates of polyploid

formation in flowering plants. Annu Rev Ecol Syst **29**: 467–501.

Reineke A. R., Bornberg-Bauer E., Gu J., 2011 Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucl. Acids Res. **39**: 6029–6043.

Rieseberg L. H., Willis J. H., 2007 Plant Speciation. Science **317**: 910 –914.

Ritter D. I., Li Q., Kostka D., Pollard K. S., Guo S., Chuang J. H., 2010 The Importance of Being Cis: Evolution of Orthologous Fish and Mammalian Enhancer Activity. Mol Biol Evol **27**: 2322–2332.

Rueb S., Hensgens L., 1989 Improved histochemical staining for B-D-glucuronidase activity in monocotyledonous plants. Rice Genetics Newletter **6**: 56.

Salama R. A., Stekel D. J., 2010 Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. Nucleic Acids Res **38**: e135.

Salse J., Abrouk M., Bolot S., Guilhot N., Courcelle E., Faraut T., Waugh R., Close T. J., Messing J., Feuillet C., 2009 Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proceedings of the National Academy of Sciences **106**: 14908 –14913.

Salse J., Bolot S., Throude M., Jouffe V., Piegu B., Quraishi U. M., Calcagno T., Cooke R., Delseny M., Feuillet C., 2008 Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution. Plant Cell **20**: 11–24.

Sankoff D., Zheng C., Zhu Q., 2010 The collapse of gene complement following whole genome duplication. BMC Genomics **11**: 313.

Scannell D. R., Byrne K. P., Gordon J. L., Wong S., Wolfe K. H., 2006 Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature **440**: 341–345.

Schnable J. C., Freeling M., 2011 Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PLoS One **6**: e17855.

Schnable J. C., Freeling M., Lyons E., 2012a Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol Evol **4**: 265–277.

Schnable J. C., Freeling M., Lyons E., 2012b Genome-Wide Analysis of Syntenic Gene Deletion in the Grasses. Genome Biol Evol **4**: 265–277.

Schnable J. C., Pedersen B. S., Subramaniam S., Freeling M., 2011a Dose-sensitivity, conserved noncoding sequences and duplicate gene retention through multiple tetraploidies in the grasses. Front Plant Sci **2**.

Schnable J. C., Springer N. M., Freeling M., 2011b Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci USA **108**: 4069 –4074.

Schnable J. C., Wang X., Pires J. C., Freeling M., 2012c Escape from Preferential Retention Following Repeated Whole Genome Duplications in Plants. Front Plant Sci **3**: 94.

Schnable P. S., Ware D., Fulton R. S., Stein J. C., Wei F., *et al.*, 2009 The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326**: 1112–1115.

Sémon M., Wolfe K. H., 2007a Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. Trends in Genetics **23**: 108–112.

Sémon M., Wolfe K. H., 2007b Consequences of genome duplication. Curr Opin Genet Dev **17**: 505–512.

Senchina D. S., Alvarez I., Cronn R. C., Liu B., Rong J., Noyes R. D., Paterson A. H., Wing R. A., Wilkins T. A., Wendel J. F., 2003 Rate Variation Among Nuclear Genes and the Age of Polyploidy in Gossypium. Mol Biol Evol **20**: 633 –643.

Seoighe C., Gehring C., 2004 Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet **20**: 461–464.

Shin J. T., Priest J. R., Ovcharenko I., Ronco A., Moore R. K., Burns C. G., MacRae C. A., 2005 Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucl. Acids Res. **33**: 5437–5445.

Soderlund C., Bomhoff M., Nelson W. M., 2011 SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. **39**: e68.

Soderlund C., Descour A., Kudrna D., Bomhoff M., Boyd L., Currie J., Angelova A., Collura K., Wissotski M., Ashley E., Morrow D., Fernandes J., Walbot V., Yu Y., 2009 Sequencing, Mapping, and Analysis of 27,455 Maize Full-Length cDNAs (JR Ecker, Ed.). PLoS Genet **5**: e1000740.

Soltis D. E., Albert V. A., Leebens-Mack J., Bell C. D., Paterson A. H., Zheng C., Sankoff D., dePamphilis C. W., Wall P. K., Soltis P. S., 2009 Polyploidy and angiosperm diversification. Am J Bot **96**: 336–348.

Soltis P. S., Soltis D. E., 2009 The Role of Hybridization in Plant Speciation. Annu Rev Plant Biol **60**: 561–588.

Spangler J. B., Subramaniam S., Freeling M., Feltus F. A., 2011 Evidence of Function for Conserved Noncoding Sequences in Arabidopsis thaliana. New Phytol **Accepted**.

Springer N. M., Ying K., Fu Y., Ji T., Yeh C.-T., Jia Y., Wu W., Richmond T., Kitzman J., Rosenbaum H., Iniguez A. L., Barbazuk W. B., Jeddeloh J. A., Nettleton D., Schnable P. S., 2009 Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content (JR Ecker, Ed.). PLoS Genet **5**: e1000734.

Sun X., Zou Y., Nikiforova V., Kurths J., Walther D., 2010 The complexity of gene expression dynamics revealed by permutation entropy. BMC Bioinformatics **11**: 607.

Swanson-Wagner R. A., Eichten S. R., Kumari S., Tiffin P., Stein J. C., Ware D., Springer N. M., 2010 Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res **20**: 1689–99.

Swigoňová Z., Lai J., Ma J., Ramakrishna W., Llaca V., Bennetzen J. L., Messing J., 2004 Close Split of Sorghum and Maize Genome Progenitors. Genome Res **14**: 1916–1923.

Tang H., Bowers J. E., Wang X., Paterson A. H., 2010 Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci USA **107**: 472–477.

Tang H., Lyons E., Pedersen B., Schnable J. C., Paterson A. H., Freeling M., 2011 Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics **12**: 102.

Tang H., Woodhouse M. R., Cheng F., Schnable J. C., Pedersen B. S., Conant G., Wang X., Freeling M., Pires J. C., 2012 Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model for paleohexaploidy. Genetics **190**: 1563–1574.

The Brassica rapa Genome Sequencing Project Consortium, 2011 The genome of the mesopolyploid crop species Brassica rapa. Nature Genetics **43**: 1035–1039.

The International Brachypodium Initiative, 2010 Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature **463**: 763–768.

THOMAS B. C., PEDERSEN B., FREELING M., 2006 Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res **16**: 934–946.

THOMAS B. C., RAPAKA L., LYONS E., PEDERSEN B., FREELING M., 2007 Arabidopsis intragenomic conserved noncoding sequence. PNAS **104**: 3348–3353.

THROUDE M., BOLOT S., BOSIO M., PONT C., SARDA X., QURAISHI U. M., BOURGIS F., LESSARD P., ROGOWSKY P., GHESQUIERE A., MURIGNEUX A., CHARMET G., PEREZ P., SALSE J., 2009 Structure and expression analysis of rice paleo duplications. Nucleic Acids Res **37**: 1248 –1259.

TRAPNELL C., WILLIAMS B. A., PERTEA G., MORTAZAVI A., KWAN G., BAREN M. J. VAN, SALZBERG S. L., WOLD B. J., PACHTER L., 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech **28**: 511–515.

TUSKAN G. A., DIFAZIO S., JANSSON S., BOHLMANN J., GRIGORIEV I., *et al.*, 2006 The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science **313**: 1596 –1604.

VANDEPOELE K., SIMILLION C., PEER Y. VAN DE, 2003 Evidence that rice and other cereals are ancient aneuploids. Plant Cell **15**: 2192–2202.

VANDERSLUIS B., BELLAY J., MUSSO G., COSTANZO M., PAPP B., VIZEACOUMAR F. J., BARYSHNIKOVA A., ANDREWS B., BOONE C., MYERS C. L., 2010 Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol Syst Biol **6**: 429.

VEITIA R. A., 2010 A generalized model of gene dosage and dominant negative effects in macromolecular complexes. FASEB J **24**: 994–1002.

VEITIA R. A., BOTTANI S., BIRCHLER J. A., 2008 Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet **24**: 390–397.

VENKATESH B., KIRKNESS E. F., LOH Y.-H., HALPERN A. L., LEE A. P., JOHNSON J., DANDONA N., VISWANATHAN L. D., TAY A., VENTER J. C., STRAUSBERG R. L., BRENNER S., 2006 Ancient Noncoding Elements Conserved in the Human Genome. Science **314**: 1892–1892.

VENKATESH B., KIRKNESS E. F., LOH Y.-H., HALPERN A. L., LEE A. P., JOHNSON J., DANDONA N., VISWANATHAN L. D., TAY A., VENTER J. C., STRAUSBERG R. L., BRENNER S., 2007 Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome. PLoS Biol. **5**: e101.

VOGEL J., HILL T., 2008 High-efficiency Agrobacterium-mediated transformation of Brachypodium distachyon inbred line Bd21-3. Plant Cell Rep. **27**: 471–478.

WANG X., ELLING A. A., LI X., LI N., PENG Z., HE G., SUN H., QI Y., LIU X. S., DENG X. W., 2009 Genome-Wide and Organ-Specific Landscapes of Epigenetic Modifications and Their Relationships to mRNA and Small RNA Transcriptomes in Maize. Plant Cell **21**: 1053–1069.

WANG X., TANG H., PATERSON A. H., 2011 Seventy Million Years of Concerted Evolution of a Homoeologous Chromosome Pair, in Parallel, in Major Poaceae Lineages. Plant Cell **23**: 27–37.

WANG J., TIAN L., LEE H.-S., WEI N. E., JIANG H., WATSON B., MADLUNG A., OSBORN T. C., DOERGE R. W., COMAI L., CHEN Z. J., 2006 Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172**: 507–517.

WATERS A. J., MAKAREVITCH I., EICHTEN S. R., SWANSON-WAGNER R. A., YEH C.-T., XU W., SCHNABLE P. S., VAUGHN M. W., GEHRING M., SPRINGER N. M., 2011 Parent-of-Origin Effects on Gene Expression and DNA Methylation in the Maize Endosperm. Plant Cell **23**: 4221–4233.

WEI F., COE E., NELSON W., BHARTI A. K., ENGLER F., BUTLER E., KIM H., GOICOECHEA J. L., CHEN M., LEE S., FUKS G., SANCHEZ-VILLEDA H., SCHROEDER S., FANG Z., MCMULLEN M., DAVIS G., BOWERS J. E., PATERSON A. H., SCHAEFFER M., GARDINER J., CONE K., MESSING J., SODERLUND C., WING R. A., 2007 Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. PLoS Genet **3**: e123.

WEI F., ZHANG J., SCHWARTZ D. C., ZHOU S., WING R., 2010 Maize Genome Sequence Release 2: B73RefGen_v2. Maize Genome Sequence Release 2: B73RefGen_v2.

WILKERSON M. D., SCHLUETER S. D., BRENDEL V., 2006 yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes. Genome Biol **7**: R58.

WOODHOUSE M. R., SCHNABLE J. C., PEDERSEN B. S., LYONS E., LISCH D., SUBRAMANIAM S., FREELING M., 2010 Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol **8**: e1000409.

WOOD T. E., TAKEBAYASHI N., BARKER M. S., MAYROSE I., GREENSPOON P. B., RIESEBERG L. H., 2009 The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA **106**: 13875–13879.

WOOLFE A., GOODSON M., GOODE D. K., SNELL P., MCEWEN G. K., VAVOURI T., SMITH S. F., NORTH P., CALLAWAY H., KELLY K., WALTER K., ABNIZOVA I., GILKS W., EDWARDS Y. J. K., COOKE J. E., ELGAR G., 2004 Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. PLoS Biol **3**: e7.

WRAY G. A., HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V., ROMANO L. A., 2003 The Evolution of Transcriptional Regulation in Eukaryotes. Mol Biol Evol **20**: 1377 –1419.

WU T. D., NACU S., 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26**: 873 –881.

XIONG Z., GAETA R. T., PIRES J. C., 2011 Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc Natl Acad Sci USA **108**: 7908–13.

YANG Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**: 1586–1591.

YU J., WANG J., LIN W., LI S., LI H., *et al.*, 2005 The Genomes of Oryza sativa: A History of Duplications. PLoS Biol **3**: e38.

ZEMACH A., MCDANIEL I. E., SILVA P., ZILBERMAN D., 2010 Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. Science **328**: 916–919.

ZHANG W., WU Y., SCHNABLE J. C., ZENG Z., FREELING M., CRAWFORD G. E., JIANG J., 2012 High-resolution mapping of open chromatin in the rice genome. Genome Research **22**: 151–162.