2009

Arrangement of gene pairs, retrotransposon insertions, and regulation of gene expression in plants

Nicholas D. Krom
Michigan Technological University

ARRANGEMENT OF GENE PAIRS, RETROTRANSPOSON INSERTIONS, AND REGULATION OF GENE EXPRESSION IN PLANTS

By

NICHOLAS D. KROM

A DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

(Biological Sciences)

MICHIGAN TECHNOLOGICAL UNIVERSITY

2009

This dissertation, "Arrangement of Gene Pairs, Retrotransposon Insertions, and Regulation of Gene Expression in Plants," is hereby approved in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in the field of Biological Sciences.

Department of Biological Sciences

Signatures:

Dissertation Advisor _____

Ramakrishna Wusirika

Committee: _____

John Adler

_____

Donald Leuking

_____

Chandrashekhar Joshi

Department Chair _____

K. Michael Gibson

Date _____

ACKNOWLEDGEMENTS

I would like to thank my advisor, Ramakrishna Wusirika, for all his help and guidance throughout my graduate career, as well as the other members of my committee, John Adler, Don Leuking, and Shekhar Joshi for their insights and feedback on my work.

Many thanks are due to my various employers over the years: John Adler, Tom Snyder, Heather Youngs, and Dave Poplawski. I am grateful for the learning experiences I gained while teaching for them. And for the money, of course.

I would be remiss if I neglected to thank the other always helpful members of the staff: Jeff Lewin and Mike Lebeau for keeping the labs working; Pat Asselin, Emily Betterly, and Emily Jackson for their aid in navigating the university's nightmarish bureaucratic underbelly; and Lori for always making the world sparkle.

I offer special thanks to all my friends and coworkers, for their help and companionship. I couldn't have done it without y'all. Well, maybe I could have, but I would have gone even more insane than I did. This august assemblage includes, but is not limited to: Tara, Emily, Foad, Steph, Zijun, Tracy, Patience, Louis, Ratul, Eric, Sarah, Kris, Tim, Leah, Surendar, the many Katies, Matt, John, Hien, Danielle, Deepak, Nari, Jen, Jill, Gunjan, Sam, Beth, Joe, Chris, Yeo, Matt, Sandy, Cory, and many others my weary mind cannot currently recall.

The most special thanks of all go to my family: Mom and Dad for all their love and support, and my brothers Ben and Steve for various brotherly things.

ABSTRACT

Plant genomes are extremely complex. Myriad factors contribute to their evolution and organization, as well as to the expression and regulation of individual genes. Here we present investigations into several such factors and their influence on genome structure and gene expression: the arrangement of pairs of physically adjacent genes, retrotransposons closely associated with genes, and the effect of retrotransposons on gene pair evolution.

All sequenced plant genomes contain a significant fraction of retrotransposons, including that of rice. We investigated the effects of retrotransposons within rice genes and within a 1 kb putative promoter region upstream of each gene. We found that approximately one-sixth of all rice genes are closely associated with retrotransposons. Insertions within a gene's promoter region tend to block gene expression, while retrotransposons within genes promote the existence of alternative splicing forms. We also identified several other trends in retrotransposon insertion and its effects on gene expression.

Several studies have previously noted a connection among genes between physical proximity and correlated expression profiles. To determine the degree to which this correlation depends on an exact physical arrangement, we studied the expression and interspecies conservation of convergent and divergent gene pairs in rice, *Arabidopsis*, and *Populus trichocarpa*. Correlated expression among gene pairs was quite common in all three species, yet conserved arrangement was rare. However, conservation of gene pair arrangement was significantly more common among pairs with strongly correlated expression levels.

In order to uncover additional properties of gene pair conservation and rearrangement, we performed a comparative analysis of convergent, divergent, and tandem gene pairs in rice, sorghum, maize, and *Brachypodium*. We noted considerable differences between gene pair types and species. We also constructed a putative evolutionary history for each pair, which led to several interesting discoveries.

To further elucidate the causes of gene pair conservation and rearrangement, we identified retrotransposon insertions in and near rice gene pairs. Retrotransposon-associated pairs are less likely to be conserved, although there are significant differences in the possible effect of different types and locations of retrotransposon insertions. The three types of gene pair also varied in their susceptibility to retrotransposon-associated evolutionary changes.

TABLE OF CONTENTS

LITERATURE REVIEW

Present-day genomes are the product of millions of years of change, selection, and divergence. Many different molecular processes introduce variation into a genome, at times producing phenotypic changes that affect the organism's survival and reproductive success, driving the process of evolution and creating the enormous diversity of living things in the world today.

Transposable elements (TEs) are one major source of genomic variation. They can be divided into two primary classes: retrotransposons, which employ an RNA intermediate during transposition, and DNA transposons, which do not (Wicker et al., 2007). Retrotransposons are by far the most common class in plants, making up a significant fraction of all sequenced plant genomes. Common plant retrotransposons are divided into three orders: Long Terminal Repeat (LTR) retrotransposons, Long Interspersed Nuclear Elements (LINEs), and Short Interspersed Nuclear Elements (SINEs). LTR-retrotransposons are flanked by LTR sequences at each end, and are further subdivided into two superfamilies, *Copia* and *Gypsy*, which differ primarily in the order of their protein coding regions (Wicker et al., 2007). The coding regions of *Copia* elements are arranged in the order {GAG, AP, INT, RT, RH}, while *Gypsy* elements are arranged {GAG, AP, RT, RH, INT}. Plant LINEs contain either three (ORF1, APE, and RT) or four (ORF1, APE, RT, and RH) coding regions, depending on their superfamily, and recognition sequences involved in the process of transposition. SINEs are non-autonomous, as they lack protein coding regions, and thus rely on enzymes encoded by LINEs to transpose. Non-autonomous versions of LTR-retrotransposons are also found in plant genomes, such as terminal repeat retrotransposons in miniature (TRIMs) and large retrotransposon derivatives (LARDs) (Witte et al, 201; Kalendar et al., 2004).

Overall retrotransposon content varies greatly among plant species, even across relatively short evolutionary distances, and is a major factor in determining overall genome size (SanMiguel et al., 1996; Bennetzen, 2002). Among the grasses, for instance, retrotransposon content ranges from approximately 8% of the *Brachypodium distachyon* genome (Huo et al., 2008) to 79% in *Zea mays* (Paterson et al., 2009). This broad range of genome sizes suggests that retrotransposon activity (i.e. sequence gain and loss) takes place at a very high rate. Vitte and colleagues (2007) hypothesized that in the ancestors of rice (*Oryza sativa* cv. Nipponbare) LTR-retrotransposon amplification occurred in bursts, with large numbers of copies being added to the genome in a relatively short time. Amplification is then followed by a longer period of relatively rapid loss of retrotransposon sequence. The rate of this sequence loss has been analyzed by several groups, resulting in an estimated half-life for LTR-retrotransposon sequence of less than 6 million (Ma et al., 2004) to 19 million years (Vitte et al., 2007). Assuming an intermediate value of 12 million years, an 8000 bp long LTR-retrotransposon present in the last common ancestor of the grasses (which diverged approximately 60 million years

ago (Wolfe et al., 1989; Buell, 2009)) would be expected to exist as a 250 bp fragment, having gone through five half-lives, in modern grass genomes. As a result of this high rate of turnover among retrotransposons, the majority of intact LTR-retrotransposons found in angiosperm genomes are believed to have been inserted less than 5 million years ago (Bennetzen, 2005). This can also result in major differences in the specific types of retrotransposons present in otherwise highly collinear regions of closely related species (Ramakrishna et al., 2002; Tikhonov et al., 1999). Much of the observed loss of retrotransposon sequence takes place through various types of recombination within and between LTR-retrotransposons, such as illegitimate recombination and unequal homologous recombination, which can also remove segments of the host genome as well as retrotransposon sequence (Ma and Bennetzen, 2004; Devos et al., 2002; Ma et al., 2004).

In addition to influencing genome size, retrotransposons inserted in or near a gene can alter that gene's expression. When an intragenic retrotransposon is included within an RNA transcript, splice sites within the insertion are sometimes employed, resulting in alternative gene products (Varagona et al. 1992; Marillonnet and Wessler 1997; Leprince et al. 2001). Parts of human *Alu* retrotransposons have been recruited as exons when present within introns (Sorek et al., 2002). The white skin color mutation in grapes is linked to the presence of a retrotransposon insertion in the promoter of a gene involved in pigment production (Kobayashi et al. 2004; Walker et al. 2007). In *Drosophila simulans* a retrotransposon insertion upstream of a gene resulted in higher levels of transcription (Schlenke and Begun, 2004), presumably due to interference with the proper function of regulatory elements. Retrotransposon promoters have also been used to initiate transcription of genes in the host genome (Van de Lagemaat et al. 2003), and alter the expression profiles of nearby genes (Kashkush et al., 2003).

Another common feature of plant genomes, in addition to high retrotransposon content, is the rapid loss of collinearity, or gene order, over time. This does not, however, imply similar differences in gene content. Among the grasses, a family that began to diverge 50-80 million years ago (Crepet and Feldman, 1991; Paterson et al., 2004; Prasad et al. 2005), genome sizes vary by 30-fold or more (Kellogg and Bennetzen, 2004), yet about 90% of genes are shared among most species (Bennetzen, 2007). However, in comparisons between maize and sorghum, which diverged approximately 12 million years ago, over one-third of all genes appear to have changed location since their divergence (Ilic et al., 2003; Lai et al., 2004). Multiple comparative analyses of orthologous regions of several grass genomes have identified numerous instances of inversions, deletions, and translocations involving small numbers of genes (Bennetzen and Ramakrishna, 2002; Ilic et al., 2003). A detailed comparison of the *Adh1* region in nine species within the genus *Oryza* identified many differences in gene gain and loss, several multi-kilobase segmental insertions and deletions, wide variation in repetitive

DNA content, and genes imported from other genomic regions, all of which arose in a span of approximately 15 million years (Ammiraju et al., 2008). In contrast, animal genomes maintain much higher levels of collinearity. For example, approximately 88% of the genes on mouse chromosome 16 have close matches within six different syntenic regions (one covering nearly one-half of chromosome 16) of the human genome, with near exact conservation of gene order, despite the fact that human and mouse lineages diverged over 80 million years ago (Mural et al., 2002). One major difference that may account for this disparity in collinearity between plant and animal genomes is polyploidization, which is rare in animals but occurs quite frequently in the lineages of plants. Nearly all angiosperms are either polyploid currently or are descended from some ancient polyploid (Paterson, 2004; Adams and Wendel, 2005; Bennetzen, 2005). Polyploidization can contribute to genome rearrangement and reduced collinearity through several mechanisms. First, by providing a duplicate of every gene, it allows for increased levels of sequence divergence or gene loss. Differential gene loss (i.e. losing different copies in related species) after polyploidization and divergence of lineages can effectively remove a gene from homologous regions, thus reducing collinearity, while retaining full function of that gene (Tian et al., 2005). Second, polyploidization has been known to stimulate transposon activity (Kashkush, 2002), with the potential for transposon-mediated rearrangements and gene inactivation. Segmental duplications can also produce many of the same effects as polyploidy, but on a smaller scale (Bennetzen, 2005).

Collinearity can also be interrupted by insertion of new genes. While there are many mechanisms capable of doing so, of particular interest are three types of transposon, common in plants, that capture genes and gene fragments and relocate them within the genome. The first of these, *Mutator*-like DNA elements (MULEs), are numerous in the rice genome (~3000 copies), and typically contain fragments (47-986 bp in length) of host genome sequence (in which case they are called "Pack-MULEs"), sometimes containing several rearrangements (Jiang et al., 2004). Approximately 5% of Pack-MULEs in rice are expressed, including their captured genome fragments, and thus may be considered novel genes themselves (Jiang et al., 2004). Another newly characterized class of transposons, *Helitrons*, replicate using a rolling-circle mechanism (Kapitonov and Jurka, 2001) and frequently contain pieces of multiple genes. These fragments are not always captured from a single locus, but appear to be added progressively over time. For example, a *Helitron* element in maize was found to contain pieces of 12 different genes (Lal et al., 2003). Like Pack-MULEs, *Helitron* transcripts have been identified, with introns spliced out to form a chimeric transcript composed of exons from the various genes. A third new type of transposon, terminal-repeat retrotransposons in miniature (TRIMs), are a non-autonomous relative of LTR-retrotransposons (Witte et al., 2001). TRIMs are involved in many kinds of genomic

rearrangement, including acting as target sites for insertion of other retrotransposons, promoting transduction of genes, and altering the internal structure of the genes into which they insert. These three types of genome-altering transposons, in conjunction with other, more common transposon families, may provide a significant contribution to plant genome diversity, especially given the overall high level of transposon activity in plants.

With so many mechanisms continually altering gene order and location, it may seem reasonable to assume that a gene's position has no effect on its function, and that as long as their internal structure and promoters are intact, genes could be distributed at random along an organism's chromosomes with no significant change in expression. However, gene order/location and expression appear to be linked, with coexpressed genes frequently being located in close proximity to one another in a wide range of eukaryotes (Hurst et al., 2004). This coexpression takes the form of both similar quantitative expression data across various conditions and shared involvement in a specific metabolic pathway or physiological process. These clusters of coexpressed genes vary considerably in size, with cluster of up to 20 genes identified in *Arabidopsis* (Williams and Bowles, 2004), and a 1,000 kb long region of coexpressed genes in the human genome (Lercher et al., 2002). Hurst and colleagues (2004) list three levels of co-regulation, each providing a general mechanism for coordinating expression across various distances. The primary level consists of *cis*-acting regulatory elements, such as bidirectional promoters, that are shared by within a small area (~10 kb or less). The secondary level involves regions of similarly modified histones controlled by Locus Control Regions (LCRs) and Boundary Elements, creating an area of somewhat uniform expression that spans ~100 kb. At the tertiary level, large stretches of chromatin are arranged into loops extending out from an "active chromatin hub", with genes near the hub being more accessible for transcription. Another possible tertiary level mechanism, chromosome territories, involves chromatin being formed into three dimensional structures, with genes on the surface being expressed while those in the interior are generally inactive. Tertiary level mechanisms affect expression over a span of up to several million bases (Hurst et al., 2004).

In plants, most studies of coexpression clusters involve relatively few genes. In *Arabidopsis*, pairs of adjacent genes are frequently coexpressed, especially when both genes are in the same functional category (Williams and Bowles, 2004). Also in *Arabidopsis*, Ren and colleagues (2005) identified numerous clusters of two to four coexpressed genes. Pairs of genes arranged in a divergent manner have been found to be controlled by a single bidirectional promoter, although this is currently believed to be more common in animal genomes (Trinklein et al., 2004) than in plants (Mitra et al., 2009). Bidirectional promoters may also be common in fungi, due to higher rates of conservation among divergent gene pairs (Kensche et al., 2008).

The enormous complexity of plant genomes provides an endless selection of topics for investigation. Due to their prevalence and wide variety of effects on all aspects of their host genome, retrotransposons are a perennial favorite, and are far from being fully understood. The coexpression and evolution of pairs of adjacent genes is a relatively new and promising area of study, with the potential to help shed light on many related aspects of genome structure and function as well.

LITERATURE CITED

Adams, K.L., J.F. Wendel. 2005. Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8: 135-141.

Ammiraju, J.S.S., F. Lu, A. Sanyal, Y. Yu, X. Song, et al. 2008. Dynamic evolution of *Oryza* genomes is reveal by comparative genomic analysis of a genus-wide vertical data set. Plant Cell 20: 3191-3209.

Bennetzen, J.L. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115: 29-36.

Bennetzen, J.L. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Cur Opin Genet & Dev 15: 621-627.

Bennetzen, J.L. 2007. Patterns in grass genome evolution. Cur Opin Plant Bio 10:176-181.

Bennetzen, J. L. and W. Ramakrishna. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol 48: 821-827.

Buell, C. R. 2009. Poaceae genomes: Going from unattainable to becoming a model clade for comparative plant genomics. Plant Physiol 149: 111–116.

Crepet, W.L., and G.D. Feldman. 1991. The earliest remains of grasses in the fossil record. Am J Bot 78: 1010-1014.

Devos, K.M., J.K.M. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12: 1075-1079.

Huo, N., G.R. Lazo, J.P. Vogel, F.M. You, et al. 2008. The nuclear genome of *Brachypodium distachyon:* analysis of BAC end sequences. Funct Integr Genomics 8: 135-147.

Hurst, L.D., C. Pal, and M.J. Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299-310.

Ilic, K., P. J. SanMiguel, and J. L. Bennetzen. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes. Proc Natl Acad Sci 100: 12265–12270.

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

Kalendar, R., C.M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy, A.H. Schulman. 2004. LARD retroelements: novel, non-autonomous components of barley and related genomes. Genetics 166: 1437-1450.

Kapitonov, V.V., and J. Jurka. 2001. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 98: 8714-8719.

Kashkush, K., M. Feldman, A.A. Levy. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics 160: 1651-1659.

Kashkush, K., M. Feldman, A.A. Levy. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33: 102–106.

Kensche, P.R., M. Oti, B.E. Dutilh, and M.A. Huynen. 2008. Conservation of divergent transcription in fungi. Trends in Genet 24: 207-211.

Kobayashi, S., N. Yamamoto, H. Hirochika. 2004. Retrotransposon-induced mutations in grape skin color. Science 304: 982.

Lai, J., J. Ma, Z. Swigonova, W. Ramakrishna, et al. 2004. Gene loss and movement in the maize genome. Genome Res 14: 1924-1931.

Lal, S.K., M.J. Giroux, V. Brendel, E. Vallejos, and C. Hannah. 2003. The maize genome contains a Helitron insertion. Plant Cell 15: 381-391.

Leprince, A.S., M.A. Grandbastien, and C. Meyer. 2001. Retrotransposons of the Tnt1B family are mobile in Nicotiana plumbaginifolia and can induce alternative splicing of the host gene upon insertion. Plant Mol Biol 47: 533–541.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genet 31: 180-183.

Ma, J., and Bennetzen J.L. 2004. Recent rapid growth and divergence of the rice nuclear genome. Proc Natl Acad Sci 101:12404-12410.

Ma, J., K.M. Devos, J.L. Bennetzen. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860-869.

Marillonnet, S., and S.R. Wessler. 1997. Retrotransposon insertion into the maize waxy gene results in tissue-specific RNA processing. Plant Cell 9:967–978.

Mitra, A., J. Han, Z.J. Zhang, and A. Mitra. 2009. The intergenic region of *Arabidopsis thaliana* cab1 and cab2 divergent genes functions as a bidirectional promoter. Planta 229: 1015-1022.

Mural, R.J., M.D. Adams, E.W. Adams, H.O. Smith, et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 296:1661-1671.

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

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

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

Ren, X.-Y., M. Fiers, W.J. Stiekema, and J.-P. Nap. 2005. Local coexpression domains of two to four genes in the genome of Arabidopsis. Plant Physiol 138: 923-934.

SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Schlenke, T.A., D.J. Begun. 2004. Strong selective sweep associated with a transposon insertion in Drosophila simulans. Proc Natl Acad Sci USA 101: 1626–1631.

Sorek, R., G. Ast, and D. Graur. 2002. Alu-containing exons are alternatively spliced. Genome Res 12: 1060–1067.

Tian, C.G., Y.Q. Xiong, T.Y. Liu, S.H. Sun, L.B. Chen, M.S. Chen. 2005. Evidence for an ancient whole genome duplication event in rice and other cereals. Yi Chuan Xue Bao 32: 519-527.

Tikhonov, A.P., P.J. SanMiguel, Y. Nakajima, N.M. Gorenstein, et al. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409-7414.

Trinklein, N.D., S.F. Aldred, S.J. Hartman, D.I. Schroeder, et al. 2004. An abundance of bidirectional promoters in the human genome. Genome Res 14: 62-66.

Van de Lagemaat, L.N., J.R. Landry, D.L. Mager, P. Medstrand. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19: 530–536.

Varagona, M.J., M. Purugganan, and S.R. Wessler 1992. Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. Plant Cell 4: 811–820.

Vitte, C., O. Panaud, and H. Quesneville. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8: 218.

Walker, A.R., E. Lee, J. Bogs, D.A.J. McDavid, M.R. Thomas, S.P. Robinson. 2007. White grapes arose through the mutation of two similar and adjacent regulatory genes. Plant J 49: 772–785.

Wicker, T., F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, et al. 2007. A unified classification system of eukaryotic transposable elements. Nat Rev Genet 8: 973-982.

Williams, E.J.G., and D.J. Bowles. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. 14: 1060-1067.

Witte, C.-P., Q.H. Le, T. Bureau, A. Kumar. 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13778-13783.

Wolfe, K.H., M. Gouy, Y.W. Yang, P.M. Sharp, and W.H. Li. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86: 6201–6205.

CHAPTER 1:

ANALYSIS OF GENES ASSOCIATED WITH RETROTRANSPOSONS IN THE RICE GENOME

Nicholas Krom, Jill Recla*, and Wusirika Ramakrishna

Previously published online in *Genetica*, December 9, 2007.

* Ms. Recla participated in a preliminary analysis related to this study. However, no data she produced remains in the final version.

1.1 ABSTRACT

Retrotransposons comprise a significant fraction of the rice genome. Despite their prevalence, the effects of retrotransposon insertions are not well understood, especially with regard to how they affect the expression of genes. In this study, we identified one sixth of rice genes as being associated with retrotransposons, with insertions either in the gene itself or within its putative promoter region. Among genes with insertions in the promoter region, the likelihood of the gene actually being expressed was shown to be directly proportional to the distance of the retrotransposon from the translation start site. In addition, retrotransposon insertions in the transcribed region of the gene were found to be positively correlated with the presence of alternative splicing forms. Furthermore, preferential association of retrotransposon insertions with genes in several functional classes was identified. Some of the retrotransposons that are part of full-length cDNA (fl-cDNA) contribute splice sites and give rise to novel exons. Several interesting trends concerning the effects of retrotransposon insertions on gene expression were identified. Taken together, our data suggests that retrotransposon association with genes have a role in gene regulation. The data presented in this study provides a foundation for experimental studies to determine the role of retrotransposons in gene regulation.

1.2 INTRODUCTION

A large fraction of complex plant genomes are composed of transposable elements (TEs). Transposable elements are present in nearly all sequenced genomes, both prokaryotic and eukaryotic. The function of TEs in diverse genomes has been debated for many years (Wessler 2001; Brookfield and Johnson 2006). It has been suggested that TEs play an important role in gene and genome evolution (Kazazian 2004; Bennetzen 2000, 2005; Vitte and Bennetzen 2006). The organization and insertion patterns of mobile elements have been well studied in various genomes. The current data suggests that transposable elements underwent a rapid turnover in the recent past that include their insertions and deletions in the genome (Prak and Kazazian 2000; Devos et al. 2002; Ma et al. 2004). Retrotransposons, a major class of TEs, are abundant in plant genomes. However, very little is known about their function in the genome.

Transposable elements have been divided into two main classes according to their method of transposition (Wicker et al. 2007). Class I elements move to new locations in the genome through an RNA intermediate that is converted into DNA by the enzyme reverse transcriptase. Retrotransposons belong to this class. They consist of long terminal repeat (LTR) and non-LTR-retrotransposons. LTR-retrotransposons are divided into two major superfamilies, *Copia* and *Gypsy*. They differ in sequence similarity and the order of their encoded gene products. Other LTR-retrotransposons present in plants include terminal repeat retrotransposons in miniature (TRIM) and large retrotransposon derivatives (LARD), which lack the coding domains required for their mobility (Witte et al. 2001; Kalendar et al. 2004). Non-LTR-retrotransposons are mainly divided into long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Class II elements (DNA transposons) are divided into two subclasses (Wicker et al. 2007). Subclass I include TEs that transpose by an excision and repair (cut and paste) method using a transposase that recognizes their terminal inverted repeat (TIR) sequences. Plant TEs that belong to superfamilies, *Tc1-Mariner*, *hAT*, *Mutator*, *P*, *PIF-Harbinger*, and *CACTA* are part of this subclass. *Helitron*s, which replicate by rolling circle mechanism and are capable of capturing gene fragments, belong to subclass II. Furthermore, *Tc1-Mariner* and *PIF-Harbinger* gave rise to miniature inverted-repeat transposable elements (MITEs), which are preferentially associated with genes (Jiang et al. 2004a).

Gene regulation is central to the genotype-phenotype relationship in all organisms. TE insertions can regulate genes to enhance gene expression, change the temporal and/or spatial patterns of expression, or give rise to a new combination of genes by alternative splicing (Varagona et al. 1992; Davis et al. 1998; Zheng et al. 2005; Medstrand et al. 2005). The use of a splice site within an inserted TE can result in the production of a novel protein. For instance, a mutated waxy allele, *wxG* in maize,

showed altered tissue-specific expression resulting in a 30-fold higher enzymatic activity in pollen than in endosperm because of alternate splicing caused by a retrotransposon insertion (Varagona et al. 1992; Marillonnet and Wessler 1997). Induction of alternative splicing has also been shown by retrotransposon insertion in a gene in *Nicotiana plumbaginifolia* (Leprince et al. 2001). Furthermore, low copy number retrotransposons such as *Bs1* provide mechanisms for the evolution of new genes by acquiring part of another gene and transposing to a new genomic location (Jin and Bennetzen 1994; Elrouby and Bureau 2001).

Retrotransposon insertions can cause a change in phenotype. For instance, *brown midrib* mutation in maize is caused by a retrotransposon insertion in the coding region of the gene COMT, which codes for O-methyl transferase (Vignols et al. 1995). Another example is the insertion of a retrotransposon in the promoter of *VvMYBA1* and two non-conservative mutations in *VvMYBA2*, the two regulatory genes controlling anthocyanin biosynthesis, which result in white skin color in grapes (Kobayashi et al. 2004; Walker et al. 2007).

In order to study the contribution of retrotransposons to the regulation of genes, we have chosen to focus on rice, a major crop species whose genome is fully sequenced (International Rice Genome Sequencing Project 2005). Investigating the association of retrotransposons with genes will provide a foundation for investigating their role in gene regulation. Here we identify retrotransposon insertions in genes from the rice genome, analyze the expression patterns of these genes and discuss possible role of retrotransposons in gene regulation.

1.3 RESULTS

Higher frequency of LTR-retrotransposon insertions compared to LINE and SINE insertions in rice genes

For this study, a "gene" was defined as a sequence from the start to the stop codon, and a "promoter" was defined as the region 1-kb upstream of the translation start site. This will include any regulatory elements between transcriptional and translational start sites. A distance of 1-kb was chosen because the majority of promoter and *cis*-regulatory elements essential for gene regulation are expected to be present within this region. With this approach, most of the regulatory elements will be recovered, although a small percentage of regulatory elements that exhibit long-range regulation will be missed.

LTR-retrotransposons belonging to *Gypsy* superfamily were the most abundant retrotransposons found inserted in genes compared to *Copia* LTR-retrotransposons, LINEs, and SINEs (Table 1). The number of genes with *Gypsy* and SINE insertions in their promoters was about 1.5 fold higher than the insertion of these elements within genes. In contrast, LINE insertions were about 1.5 fold more common in genes than in promoters. *Copia* insertions appeared in genes and promoters with approximately the same frequency. A total of 714 genes with *Gypsy* insertions and 478 genes with *Copia* insertions were identified in TIGR release 4 of the rice pseudomolecules (Table 1). In addition, 506 and 628 genes with LINE and SINE insertions, respectively, were identified. Furthermore, 1097 and 467 genes with *Gypsy* and *Copia* LTR-retrotransposon insertions in promoters were identified. Likewise, 348 and 929 genes with LINE and SINE insertions, respectively, in their promoters were identified. A total of 1556 (5.5% of rice genes), 818 (2.9%), 815 (2.9%) and 1502 (5.3%) genes appear to be associated with *Gypsy* LTR-retrotransposons, *Copia* LTR-retrotransposons, LINEs and SINEs, respectively. Altogether, this accounts for about one-sixth of rice genes being associated with retrotransposons.

Non-random chromosomal distribution of retrotransposon inserted genes in the rice genome

The chromosomal distribution of genes with retrotransposon insertions was investigated in order to detect any bias for a specific chromosome. Table 2 shows the number of genes with each type of retrotransposon found on the twelve rice chromosomes. Number of genes expected with retrotransposon insertions for each chromosome was calculated based on the expected fraction for each chromosome which was estimated using the number of genes on that particular chromosome. The binomial test with Bonferroni correction was used to show that *Gypsy* LTR-retrotransposon

insertions in both promoters and genes were significantly under-represented on chromosomes 1 and 3, their insertions in genes were over-represented on chromosomes 4 and 8, and insertions in promoters were over-represented on chromosome 11. *Copia* LTR-retrotransposon insertions in promoters were under-represented on chromosomes 1 and 3 and over-represented on chromosomes 11 and 12. *Copia* insertions in genes were under-represented on chromosome 3 and over-represented on chromosomes 4 and 12. Similarly, LINE insertions in genes were under-represented on chromosome 3. SINE insertions in both promoters and genes were over-represented on chromosome 9. This data indicates that retrotransposons show preferential insertion in genes on some chromosomes.

We further investigated the ratio of retrotransposon insertions in genes to that of promoters. There seems to be a correlation between the number of retrotransposons found in promoters and those in genes. Interestingly, on all the chromosomes *Gypsy* LTR-retrotransposon and SINE insertions in genes were found to be lower compared to those in promoters (Table 2). However, the number of LINE insertions in genes was found to be higher compared to promoters on all chromosomes except chromosome 3. No clear pattern is noticeable with regard to the ratio of *Copia* insertions in promoters and genes across chromosomes. The mean ratio of *Gypsy* LTR-retrotransposon and SINE insertions in genes to that in promoters was approximately 0.7 and that for LINE insertions was 1.5. This suggests that LINEs show preferential insertion in genes compared to promoters. Alternately, selection had prevented insertions in promoters because they could prove to be deleterious.

Distribution of retrotransposon insertions upstream of translation start site

The distribution of retrotransposon insertions relative to the distance from the translation start site (TLS) was examined in 100 bp segments up to 1 kb upstream of the TLS. The number of genes with *Gypsy* LTR-retrotransposon insertions increases gradually from 43 to 136 as the distance from the TLS increases from 101 bp to 800 bp (Table 3). However, from 1 to 100 bp upstream of the TLS, there is a spike in the number of insertions to 245, more than 5-fold compared to the next 100 bp interval. *Copia* LTR-retrotransposon insertions also show a spike in the first 100 bp upstream of the TLS, but vacillate between 27 and 56 insertions per 100 bp interval afterwards. The number of LINE and SINE insertions increased steadily with increasing distance from the TLS, leveling off at 600 bp and 500 bp upstream of the TLS for LINEs and SINEs, respectively. All retrotransposon types display marked decreases in the number of insertions from 901 bp to 1 kb compared to the number of insertions in the interval from

801 bp to 900 bp, ranging from 40% fewer *Gypsy* LTR-retrotransposon insertions to 70% fewer LINE insertions.

In order to estimate the effect of retrotransposon insertions in promoters on gene expression, we determined the number of genes with retrotransposon insertions in their promoters that had full-length cDNAs (fl-cDNAs). Only 11%, 22%, 30% and 37% of the genes with *Copia*, *Gypsy*, LINE and SINE insertions, respectively, in the first 100 bp upstream of the start codon had matching fl-cDNAs (Table 3). About 2-3 fold increase in the percentage of genes having full-length cDNAs with retrotransposon insertions between 100 bp and 200 bp upstream of the start codon was observed. The general trend appears to be an increase in the percentage of genes with fl-cDNAs as the distance of retrotransposon insertion increases from the TLS, with the highest percentage of genes with retrotransposons having full length cDNAs showing insertions between 900 bp and 1 kb upstream of the TLS.

Preferential association of retrotransposons with genes belonging to different functional categories

Gene Ontology (GO) classification was used to investigate the possible functions of genes associated with retrotransposons. GO classification identified genes belonging to several categories that were over- or under-represented compared to GO data for the whole genome (Table 4). Statistical significance of this data was determined using the binomial test with Bonferroni correction. Genes containing both types of LTR-retrotransposon insertions were quite frequently under-represented in various GO classes. *Copia* insertions in both genes and promoters were under-represented among genes encoding proteins involved in regulation of biological processes, showing transcription regulator activity, and those localized in organelles. In addition, *Copia* insertions in promoters were over-represented among genes encoding proteins with signal transducer activity and those localized in extracellular regions, and under-represented in the GO classes "physiological process" and "catalytic activity". *Gypsy* insertions in both promoters and genes were found significantly less frequently than expected among genes in the classes "physiological process," "binding," "transcription regulator activity," "cell part," and "organelle." Genes containing *Gypsy* insertions in their promoters were also under-represented in the GO classes "regulation of biological processes," "reproduction," and "transporter activity," while *Gypsy* insertions in genes were under-represented among proteins localized in organelle parts and possessing catalytic activity.

The numbers of promoters and genes containing LINE and SINE insertions in the various GO classes are generally closer to genomic averages than those containing LTR-

retrotransposon insertions. For LINE insertions, only those in promoters of genes involved in catalytic activity deviated significantly from the expected value. These showed significant over-representation. SINE insertions in genes were over-represented in several GO categories that include "physiological process," "interaction between organisms," "catalytic activity," "transporter activity," and "cell part." Over-representation was not observed for *Gypsy* insertions in either promoters or genes.

Expression analysis of genes associated with retrotransposons

In order to evaluate whether the rice genes with retrotransposon insertions show expression, they were analyzed for the presence of corresponding fl-cDNAs and MPSS data (Kikuchi et al. 2003; Nakano et al. 2006; Nobuta et al. 2007). A total of 193, 596, 232, and 649 genes with *Copia, Gypsy*, LINE, and SINE insertions, respectively, in their promoters showed evidence of expression based on either fl-cDNA and/or MPSS data (Table 5). These account for about 41%, 54%, 67%, and 70% of the genes with *Copia, Gypsy*, LINE, and SINE insertions, respectively, in their promoters. Similarly, 258, 361, 359, and 474 genes, which account for about 54%, 50%, 71%, and 75% with *Copia, Gypsy*, LINE and SINE insertions, respectively, in genes had either fl-cDNAs and/or MPSS data. The absence of fl-cDNA and/or MPSS data for a given gene indicates either the absence of expression or that the specific developmental stage/tissue type was not assayed where the gene is expressed. Alternately, the level of expression was below the detection limit of the techniques used to generate MPSS or fl-cDNA data. Thus, the lower percentage of genes expressed with *Copia* or *Gypsy* LTR-retrotransposon insertions compared to LINE or SINE insertions using the same expression data suggests that LINEs and SINEs are less likely to eliminate the expression of genes compared to LTR-retrotransposons.

Next, we investigated the presence of retrotransposons in gene transcripts. A total of 55, 108, 53, and 38 genes with *Copia* LTR-retrotransposon, *Gypsy* LTR-retrotransposon, LINE, and SINE insertions, respectively, were found to have retrotransposons as part of fl-cDNAs (Table 5). This analysis showed that a higher percentage (15%) of genes with *Gypsy* LTR-retrotransposon insertions have the retrotransposon as part of fl-cDNAs compared to 11.5%, 10.5% and 6% of the total genes that had *Copia,* LINE and SINE insertions, respectively, as part of fl-cDNAs. These percentages were estimated using the data from Tables 1 and 5. This constitutes about 26%, 38%, 8% and 10% of the 212, 283, 277 and 380 genes with *Copia, Gypsy*, LINE and SINE insertions, respectively, that have fl-cDNAs (Table 5). This data suggests that both types of LTR-retrotransposon insertions in genes are more likely to become part of exons compared to LINE and SINE insertions.

Higher proportion of alternate splicing models of genes with LINE and SINE insertions

Genes with retrotransposon insertions were investigated for the presence of alternate splicing models. 82 (17%), 112 (16%), 113 (22%), and 148 (24%) genes with *Copia*, *Gypsy*, LINE, and SINE insertions, respectively, had alternate splicing models compared to 4648 (16%) genes in the entire rice genome that had alternate splicing models. The statistical significance of the effect of retrotransposon insertion on alternative splicing was evaluated using the binomial test (normal approximation), and genes containing LINE or SINE insertions were significantly more likely ($p < 0.000001$ and $p < 0.000172$, respectively) to have alternative splicing models compared to the genome as a whole, suggesting a role for LINE and SINE insertions in generating alternate transcripts. Analysis of promoter regions identified 23 (5%), 97 (9%), 39 (11%), and 131 (14%) genes with *Copia, Gypsy*, LINE and SINE insertions in their promoter regions which showed alternate splicing models. The binomial test was again applied to test for significant over- or under-representation. Genes whose promoters contain *Copia* or *Gypsy* insertions are far less likely ($p < 0.000001$) to have alternate transcripts, while LINE insertions appear to have a weaker, but still significant ($p < 0.004$) effect. SINE insertions in promoters do not appear to have a significant effect on alternative splicing. This suggests that LTR-retrotransposons and LINEs may reduce the likelihood of alternative splicing when present in promoters.

Different patterns of retrotransposon insertions in genes

Retrotransposon insertions appear to be part of exons as well as introns. Different patterns were observed in genes where retrotransposon insertions were part of cDNAs.

LTR-Retrotransposons: Genes with LTR-retrotransposons that showed alternate transcripts had the retrotransposon as part of either one cDNA (Fig. 1A-D) or more than one fl-cDNA of varying lengths (Fig. 1E-F). In some cases, where the retrotransposon was part of a cDNA, intron-exon or exon-intron splice junctions were present within a retrotransposon (Fig. 1B-D, F). Figure 1A shows a gene encoding a protein similar to hexose carrier protein that can perform diverse functions including sugar transport and sensing (Lalonde et al. 1999). Alternative splicing generates three cDNAs, and one of them, AK069891, ends in a retrotransposon. Figure 1B shows a gene whose putative protein product is similar to nonspecific lipid-transfer protein thought to be involved in diverse biological processes such as cutin formation and embryogenesis, response to pathogens, and adaptation to environmental stresses (Kader 1996). The second exon and

5' part of the third exon of the gene represented in the cDNA, AK070414 was generated from part of the retrotransposon. This implies that the splice junctions of exon 2 and the intron-exon splice junction of exon 3 arose from the retrotransposon. Figure 1C shows a gene whose putative protein product is closest to a protein encoded by a maize defense inducible gene (Simmons et al. 2003). The first exon in the cDNA, AK100888, is contributed by LTR-retrotransposon. Figure 1D shows a gene encoding a protein similar to Mov34 family protein. Members of this family are found in proteasome regulatory subunits and regulators of transcription factors (Aravind and Ponting 1998). Figure 1E shows a gene with two transcripts. The cDNA, AK065384, ends in a SINE. The second cDNA, AK067477, includes both a SINE and an LTR-retrotransposon. The putative protein product encoded by this gene shows homology to GOS9, which is probably involved in cell cycle regulation (Rey et al. 1993). Figure 1F shows a gene encoding a putative protein product similar to aspartyl protease involved in proteolysis. A *copia*-type LTR-retrotransposon is part of the second exon in the cDNA AK100338, whereas a *gypsy*-type retrotransposon is part of the last exon including the intron-exon junction corresponding to the cDNA AK109756.

LINEs: Genes with LINE insertions also had alternate transcripts as part of either one cDNA (Fig. 2A-E) or more than one cDNA (Fig. 2F). In addition, intron-exon or exon-intron splice junctions of some genes with LINEs were present within a retrotransposon (Fig. 2E-G). Figure 2A shows a gene encoding a putative protein product similar to the leucine zipper transcription factor HBP-1b. One of the cDNAs, AK069158, has a LINE as part of an internal exon. Figure 2B shows a gene that codes for the rice blast resistance protein Pib. Pib gene on rice chromosome 2 confers race specific resistance to the fungal pathogen *Magnaporthe grisea* (Wang et al. 1999). A cDNA, AB013449, codes for the rice Pib protein. A second cDNA, AK067225, includes a LINE. Figure 2C shows a gene whose protein product is similar to maize nitrate transporter (Quaggiotii et al. 2004), which belongs to the POT protein family. Most of the POT family members are involved in peptide transport. A full-length cDNA, AK065457, corresponding to this gene has a LINE in the first exon. Figure 2D shows a gene whose predicted protein product is similar to flavonol 3-sulfotransferase involved in regulating auxin transport and signaling, and response to stress in plants (Varin et al. 1997). Figure 2E shows a gene encoding an unknown protein. One of the exons present in cDNA, AK070590, is part of a LINE with splice junctions contributed by the LINE. Figure 2F shows a gene whose putative protein product is similar to LEC14B whose function is not known. However, this protein has a WD40 domain, which is present in proteins that are involved in signal transduction, pre-mRNA processing, and cytoskeleton assembly. Two of the three cDNAs have a LINE with an exon entirely contributed by the

LINE in the cDNA NM_001049630. Figure 2G shows a gene whose putative protein product is similar to cell wall associated kinases, which are involved in pathogen response and cell elongation (Verica and He 2002). The entire last exon and the intron-exon splice junction are contributed by a LINE.

SINEs: Genes with SINE insertions showed alternate transcripts as part of either one cDNA (Fig. 3A-C) or more than one cDNA (Fig. 3D). Figure 3A shows a gene whose putative protein product is similar to prolyl endopeptidase, which acts as a proteolytic enzyme. One cDNA, AK069664, ends before SINE insertion whereas a second cDNA, AK065693, includes the SINE. Figure 3B shows a gene whose putative protein product shows homology to glycosyl hyrolase family 17 proteins. In the cDNA AK067284, the SINE is spliced out whereas in AK072943 the SINE is part of the last exon. Figure 3C shows a gene whose putative protein product shows homology to a pectinesterase inhibitor, which controls post-translational regulation of pectin methylesterase (PME). Plant PMEs play a role in several processes that include microsporogenesis, pollen growth, seed germination, root development, stem elongation, fruit ripening, and response to fungal pathogens (Di Matteo et al. 2005). In the cDNA, AK072310, the last exon compared to the cDNA AK071817, is extended to include a SINE. Figure 3D shows a gene which codes for an unknown protein. One cDNA (AK065202) includes a SINE as part of a 1.5 kb transcript whereas a second cDNA (AK121914) starts with a SINE.

1.4 DISCUSSION

The abundance of TEs in large scale genome sequence data has resulted in renewed efforts to understand their function. In the present study, we discovered that about one-sixth of the genes in the rice genome are associated with LTR-retrotransposons, LINEs, and/or SINEs. This information can serve as an estimate of the degree to which TEs act as a source of functional changes in the rice genome. It has been proposed that a substantial fraction (about 25%) of human regulatory sequences arose from TEs, based on analysis of human genome data (Jordan et al. 2004; Jordan 2006). Furthermore, the involvement of LTR-retrotransposons in the structural and/or regulatory evolution of *C. elegans*, *Drosophila*, human and mouse genes was suggested due to their close association with genes (Nekrutenko and Li 2001; Ganko et al. 2003; Van de Lagemaat et al. 2003; Franchini et al. 2004; DeBarry et al. 2006; Ganko et al. 2006). Transposable elements such as *Mutator*-like elements (MULEs) have been suggested to capture genes, provide novel protein coding regions and contribute to the evolution of genes in rice (Jiang et al. 2004b). A recent report in rice suggests that retrotransposition generated chimeric genes that perform novel functions (Wang et al. 2006). However, only 27 (2%) of the primary retrogenes were found within LTR-retrotransposons. Another study surveyed transcriptional activity of TE-related genes in rice (Jiao and Deng, 2007). The data obtained in the present study supports the hypothesis that retrotransposons associated with genes in rice play a role in gene regulation and evolution. By building upon the foundation of data presented here, detailed analyses of retrotransposon-mediated gene regulatory can be accomplished.

Lack of selection pressure on retrotransposon insertions in promoters and genes would result in their random distribution on rice chromosomes. However, we found either an over- or under-representation of *Copia* and *Gypsy* LTR-retrotransposon insertions in promoters and genes on six different rice chromosomes. Although the reason for differential association of retrotransposons with genes is not known, it is possible that some chromosomal regions provide a favorable environment for their insertions and/or illegitimate or homologous recombination in the case of LTR-retrotransposons (Ma et al. 2004) generating truncated elements in genic regions. As a result of differential insertion patterns, some genes in the GO subclasses were also under- or over-represented. It is likely that these retrotransposons are under selection pressure. Insertions of retrotransposons in genes, could lead to loss/reduction in plant viability and a decrease in efficiency of plant survival in competitive environments. Such insertions would not be selected. This can result in an under-representation of retrotransposons in genes belonging to some GO subclasses. Conversely, frequent insertions of retrotransposons in other GO subclasses may lead to the creation of novel gene functions that would confer an adaptive advantage for the over-all fitness of the plant. Such genes would be over-represented in the GO subclasses.

Insertions in the core promoter region close to the transcription start site might affect the transcription of a gene. In the present study, we found a spike in *Copia* and *Gypsy* LTR-retrotransposon insertions in the first 100 bp upstream the translation start site which could be due to the ability of genes to tolerate these insertions in the 5' untranslated region (5' UTR) than in the region 5' to the transcriptional start site. This is supported by the average length of 106 bp of 5' UTR reported in vascular plants (Lynch et al. 2005).

Insertions in the promoter region may impact the regulation of a gene, either by up-regulation or down-regulation. For instance, insertion of a non-LTR-retrotransposon, *Doc*, in the 5' flanking region of a cytochrome P450 gene in *Drosophila simulans*, is associated with increased transcript abundance (Schlenke and Begun 2004). In the current study, more than half of the genes with LTR-retrotransposons and two thirds of the genes with LINEs and SINEs in their promoters were found to be expressed suggesting that they are functional.

Retrotransposons have the ability to use their own promoter for the transcription of host genes via insertion within the host gene's promoter (Van de Lagemaat et al. 2003). For example, wheat WIS2 retrotransposon LTRs have been shown to activate or silence neighboring genes (Kashkush et al. 2003). Here, we have shown that there is an increase in the proportion of genes expressed with increase in the distance of retrotransposon insertions from the translation start site. Excision of known retrotransposon promoter sequences, sequence modification by site directed mutagenesis and/or making deletion constructs, and their insertion into an expression vector will facilitate the identification of regulatory sequences within these promoters that are essential for gene expression.

Insertion of a transposable element in a gene or a regulatory region can induce or suppress alternative splicing and/or change gene expression patterns, which can result in a relatively rapid change in the function of a gene. In the primate anthropoids, *SETMAR*, a new gene evolved by the fusion of a *SET* histone methyltransferase gene with a downstream transposase gene, was suggested to shape novel regulatory networks (Jordan 2006; Cordaux et al. 2006). In the human genome, parts of Alu retrotransposons have been found to be recruited as exons when inserted in intronic regions, creating novel alternative transcripts (Sorek et al. 2002; Sorek et al. 2004). Our study in rice has identified several potential instances of LTR-retrotransposons, LINEs, and SINEs acting as exon donors. In addition, genes containing retrotransposon insertions especially LINEs and SINEs in rice appear more likely to have alternate splicing models compared to insertions in promoters whose genes appear to have less than expected alternate transcripts. It is possible that the generation of alternate transcripts by retrotransposon inserted genes may lead to the evolution of new functions.

Our studies suggest that retrotransposons may act as important regulators of gene expression and functional diversification in rice. This study serves as a foundation for in-depth analyses of retrotransposon inserted genes and promoters and their roles in the evolutionary and environmental adaptation of plants.

1.5 METHODS

Identification of rice genes associated with retrotransposons

Gene sequence and annotation data (version 4) for the *Oryza sativa* ssp. *japonica* (cultivar Nipponbare) genome were downloaded from the Rice Genome Annotation (version 4) Database at The Institute for Genomic Research (TIGR) (Yuan et al. 2005). Genes annotated as hypothetical, pseudogenes or transposon-related were excluded, leaving 28,287 genes for further analysis. The unspliced genomic and 1 kb upstream sequences of the remaining genes were analyzed for retrotransposon insertions using RepeatMasker (http://www.repeatmasker.org) with the latest Repbase repeat sequence library (http://www.girinst.org/repbase/index.html; Jurka 2000). The RepeatMasker output was then parsed to identify genes containing LTR-retrotransposons, LINEs, or SINEs, within the gene, 1 kb upstream, or both. Most of the LTR-retrotransposons and LINEs associated with genes were truncated. The binomial test (normal approximation) with Bonferroni correction was used to determine which chromosomes contain greater than expected numbers of promoters and genes with retrotransposon insertions compared to the overall chromosomal distribution of genes.

Functional classification of genes

Gene Ontology (GO) classification data for all previously identified genes containing retrotransposon insertions was downloaded from the TIGR Rice Genome Annotation Database (Yuan et al. 2005; http://www.tigr.org/tdb/e2k1/osa1). Using the GO classification tree from the Gene Ontology Database (http://www.genedb.org), a full list of GO classes to which each gene belongs was created. This list was then analyzed to determine the number of genes belonging to each of the second level classes in the overall GO hierarchy. The binomial test (normal approximation) with Bonferroni correction was used to determine which individual classes were over- and under-represented among genes with retrotransposon insertions.

Expression analysis

Sequences for 28,469 *Oryza sativa* ssp. *japonica* full-length cDNA were obtained from the Rice Full-length cDNA Consortium (http://cdna01.dna.affrc.go.jp/cDNA). A BLASTN (Atlschul et al. 1997) search comparing the coding sequence of the genes containing retrotransposon insertions with the full-length cDNA (fl-cDNA) database was performed, and a list of matching fl-cDNAs was compiled for each gene. These

matching fl-cDNA sequences were then analyzed with RepeatMasker to determine if any retrotransposon sequence was included in the transcript.

Massively Parallel Signature Sequencing (MPSS) (Nobuta et al. 2007; http://mpss.udel.edu/rice) data was compiled for each gene containing retrotransposon insertions. Only class 1 signatures (located within an exon) found in a single gene were used in further analysis. The MPSS data for each gene was then analyzed to determine if the gene is expressed as represented by the presence of MPSS signatures(s).

Alternative splicing analysis

Gene splicing model data from the TIGR Rice Genome Annotation Database (Yuan et al. 2005) was compiled for all genes with retrotransposon insertions, and genes with multiple splicing models were identified. In addition, genes shown in figures were analyzed manually, using BLAST searches and the data available on the TIGR web site, for the presence of multiple unique fl-cDNAs which represent alternate transcripts.

1.6 ACKNOWLEDGEMENTS

We thank Dr. Aparna Deshpande for her critical review of the manuscript and help in the development of the final version. Preliminary analysis done by Matthew McCormick and Zijun Xu is greatly appreciated.

1.7 LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl Acids Res 25:3389-3402

Aravind L, Ponting CP (1998) Homologues of 26S proteasome subunits are regulators of transcription and translation. Protein Sci 7:1250-1254

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42:251-269

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621-627

Brookfield JFY, Johnson LJ (2006) The evolution of mobile DNAs: When will transposons create phylogenies that look as if there is a master gene? Genetics 173:1115-1123

Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA 103:8101-8106

Davis MB, Dietz J, Standiford DM, Emerson CP (1998) Transposable element insertions respecify alternative exon splicing in three *Drosophila* myosin heavy chain mutants. Genetics 150:1105-1114

DeBarry JD, Ganko EW, McCarthy EM, McDonald JF (2006) The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. Mol Biol Evol 23:479-481

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res 12:1075-1079

Di Matteo A, Giovane A, Raiola A, Camardella L, Bonivento D, De Lorenzo G, Cervone F, Bellincampi D, Tsernoglou D (2005) Structural basis for the interaction between pectin methylesterase and a specific inhibitor protein. Plant Cell 17:849-858

Elrouby N, Bureau TE (2001) A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. J Biol Chem 276:41963-41968

Franchini LF, Ganko EW, McDonald JF (2004) Retrotransposon-gene associations are widespread among *D. melanogaster* populations. Mol Biol Evol 21:1323-1331

Ganko EW, Bhattacharjee V, Schliekelman P, McDonald JF (2003) Evidence for the contribution of LTR retrotransposons to *C. elegans* gene evolution. Mol Biol Evol 20:1925-1931

Ganko EW, Greene CS, Lewis JA, Bhattacharjee V, McDonald JF (2006) LTR retrotransposon-gene associations in *Drosophila melanogaster*. J Mol Evol 62:111-120

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436: 793-800

Jiang N, Feschotte C, Zhang X, Wessler SR (2004a) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol 7: 115-119

Jiang N, Zhirong B, Zhang X, Eddy SR, Wessler SR (2004b) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569-573

Jiao Y, Deng XW (2007) A genome-wide transcriptional activity survey of rice transposable element-related genes. Genome Biol 8: R28

Jin YK, Bennetzen JL (1994) Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. Plant Cell 6:1177-1186

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68-72

Jordan IK (2006) Evolutionary tinkering with transposable elements. Proc Natl Acad Sci USA 103:7941-7942

Jurka J (2000) Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet 9:418-420

Kader J-C (1996) Lipid-transfer proteins in plants. Annual Rev Plant Phys Plant Mol Biol 47:627-654

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) LARD retroelements: novel, non-autonomous components of barley and related genomes. Genetics 166:1437-1450

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102-106

Kazazian HH (2004) Mobile elements. Drivers of genome evolution. Science 303:1626-1632

Kikuchi S, *et al.* (2003) Collection, mapping, and annotation of 28,000 full-length cDNA clones from *Japonica* rice. Science 301:376-379

Kobayashi S, Yamamoto N, Hirochika H (2004) Retrotransposon-induced mutations in grape skin color. Science 304:982

Lalonde S, Boles E, Hellmann H, Barker L, Patrick JW, Frommer WB, Ward JM (1999) The dual function of sugar carriers. Transport and sugar sensing. Plant Cell 11:707-726

Leprince AS, Grandbastien MA, Meyer C (2001) Retrotransposons of the Tnt1B family are mobile in *Nicotiana plumbaginifolia* and can induce alternative splicing of the host gene upon insertion. Plant Mol Biol 47:533-541

Lynch M, Scofield DG, Hong X (2005) The evolution of transcription-initiation sites. Mol Biol Evol 22:1137-1146

Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14:860-869

Marillonnet S, Wessler SR (1997) Retrotransposon insertion into the maize *waxy* gene results in tissue-specific RNA processing. Plant Cell 9:967-978

Medstrand P, van de Lagemaat LN, Dunn CA, Landry J-R, Svenback D, Mager DL (2005) Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res 110:342-352

Nakano M, Nobuta K, Vemaraju K, Tej S, Skogen JW, Meyers BC (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 34:D731-D735

Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619-621

Nobuta K, Venu RC, Lu C, Belo´ A, Vemaraju K, Kulkarni K, Wang W, Pillay M, Green PJ, Wang G, Meyers BC (2007) An expression atlas of rice mRNAs and small RNAs. Nat Biotechnol 25:473-477

Prak ETL, Kazazian H (2000) Mobile elements and the human genome. Nature Rev Genet 1:134-144

Quaggiotti S, Ruperti B, Pizzeghello D, Francioso O, Tugnoli V, Nardi S (2004) Effect of low molecular size humic substances on nitrate uptake and expression of genes involved in nitrate transport in maize (*Zea mays* L.) J Exp Bot 55:803-813

Rey P, Diaz C, Schilperoort RA, Hensgens LAM (1993) Cell-type specific expression of three rice genes GOS2, GOS5 and GOS9. Plant Mol Biol 23:889-894

Schlenke TA, Begun DJ (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans.* Proc Natl Acad Sci USA 101:1626-1631

Simmons CR, Fridlender M, Navarro PA, Yalpani N (2003) A maize defense-inducible gene is a major facilitator superfamily member related to bacterial multidrug resistance efflux antiporters. Plant Mol Biol 52:433-446

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12:1060-1067

Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D, Ast G (2004) Minimal conditions for exonization of intronic sequences:′5splice site formation in alu exons. Mol Cell 14:221-231

Van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530-536

Varagona MJ, Purugganan M, Wessler SR (1992) Alternative splicing induced by insertion of retrotransposons into the maize *waxy* gene. Plant Cell 4:811-820

Varin L, Marsolais F, Richard M, Rouleau M (1997) Biochemistry and molecular biology of plant sulfotransferases. FASEB J 11:517-525

Verica JA, He Z-H (2002) The cell wall-associated kinase (*WAK*) and *WAK*-like kinase gene family. Plant Physiol 129:455-459

Vignols F, Rigau J, Torres MA, Capellades M, Puigdomenech P (1995) The *brown midrib3* (*bm3*) mutation in maize occurs in the gene encoding caffeic acid o-methyltransferase. Plant Cell 7:407-416

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638-17643

Walker AR, Lee E, Bogs J, McDavid DAJ, Thomas MR, Robinson SP (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. Plant J 49:772-785

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK, Long M, Wang J (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791-1802

Wang ZX, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, Katayose Y, Sasaki T (1999) The Pib gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. Plant J 19:55-64

Wessler SR (2001) Plant transposable elements. A hard act to follow. Plant Physiol 125:149-151

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet doi:10.1038/nrg2165

Witte C-P, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13778-13783

Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, Hamilton J, Haas B, Sultana R, Cheung F, Wortman J, Buell CR (2005) The Institute for Genomic Research Osa1 rice genome annotation database. Plant Physiol 138:18-26

Zheng CL, Fu XD, Gribskov M (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. RNA 11:1777-1787

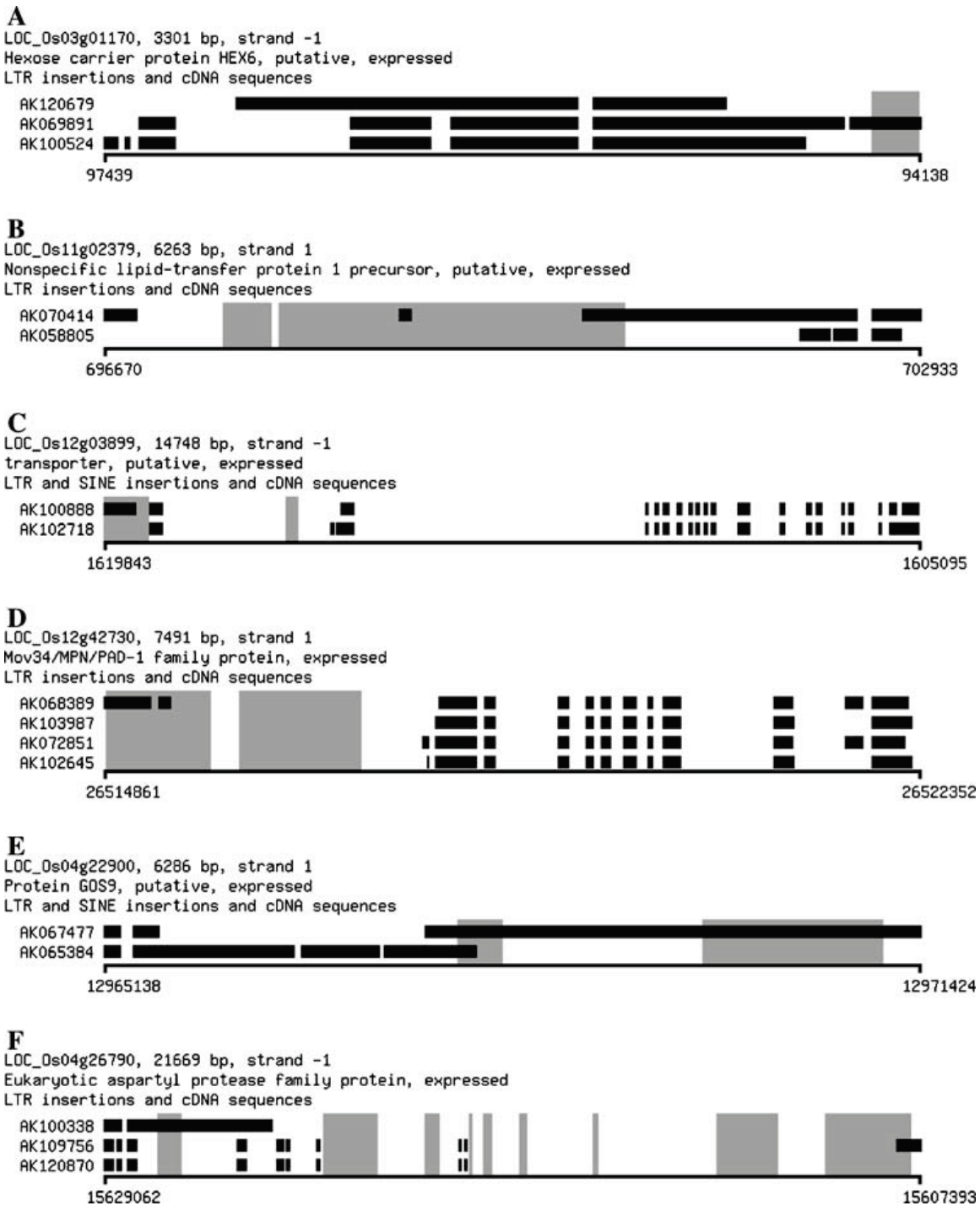


Figure 1: Examples of genes with LTR-retrotransposons. Black colored rectangles represent exons and grey colored rectangles represent LTR-retrotransposons except in Fig. 1C, where the second grey rectangle and Fig. 1E, where the first grey rectangle represent a SINE. The unique TIGR locus identifier is shown for each gene. Coordinates shown below correspond to the positions on the chromosome. Fig. 1A-E. LTR-

retrotransposon is part of only one fl-cDNA. Fig. 1E. SINE is part of two fl-cDNAs with AK067477 harboring both a SINE and an LTR-retrotransposon. Fig. 1F. Two different LTR-retrotransposons are part of two fl-cDNAs of different lengths. Fig. 1B-D, F. Some exon-intron or exon-intron splice junctions are contributed by a retrotransposon.

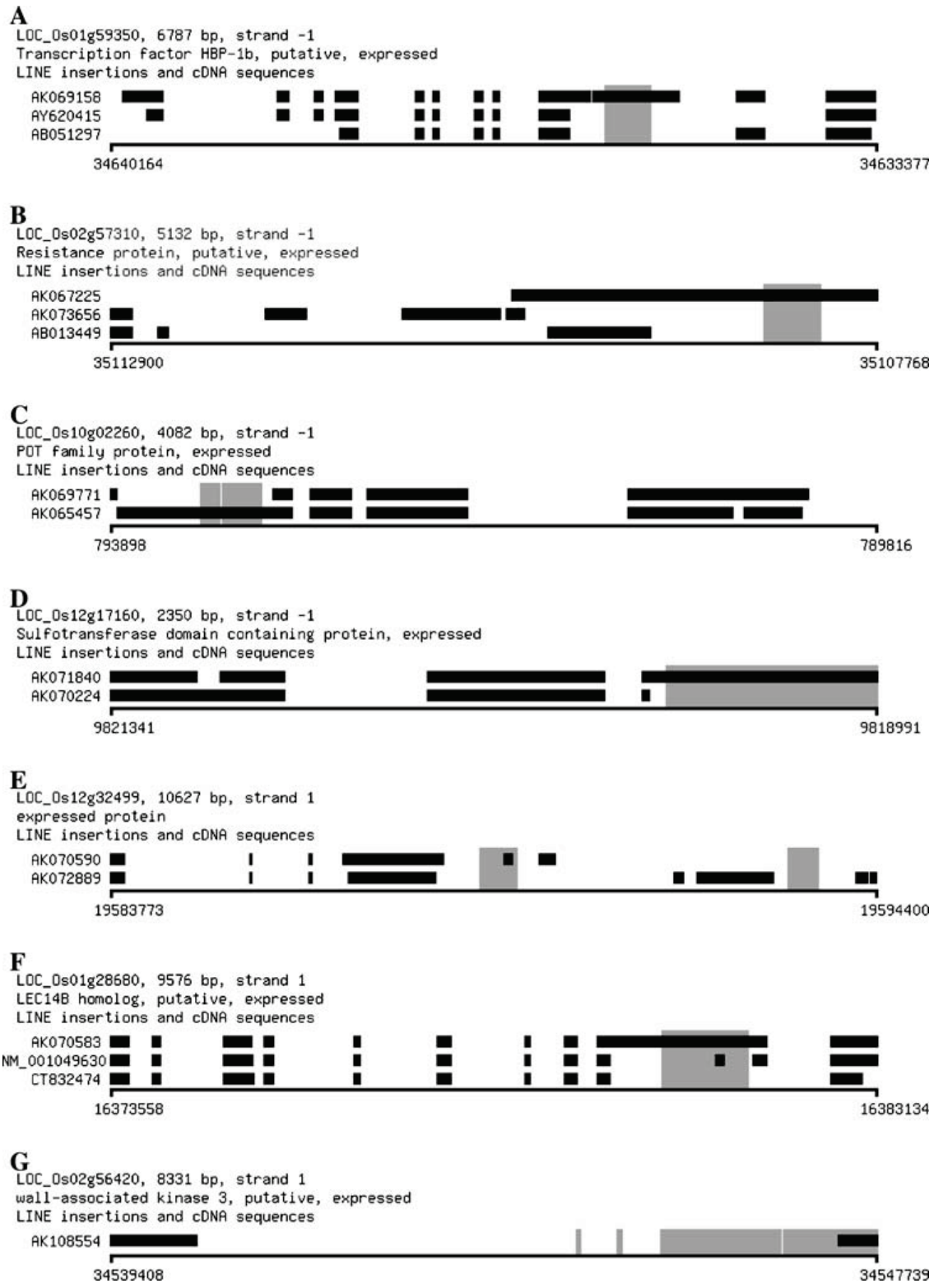


Figure 2: Examples of genes with LINE insertions. Black colored rectangles represent exons and grey colored rectangles represent LINEs. Fig. 2A-E, G. LINE is part of only

one fl-cDNA. Fig. 2F. LINE is part of two fl-cDNAs. Fig. 2E-G. Some intron-exon or exon-intron splice junctions are contributed by a LINE.

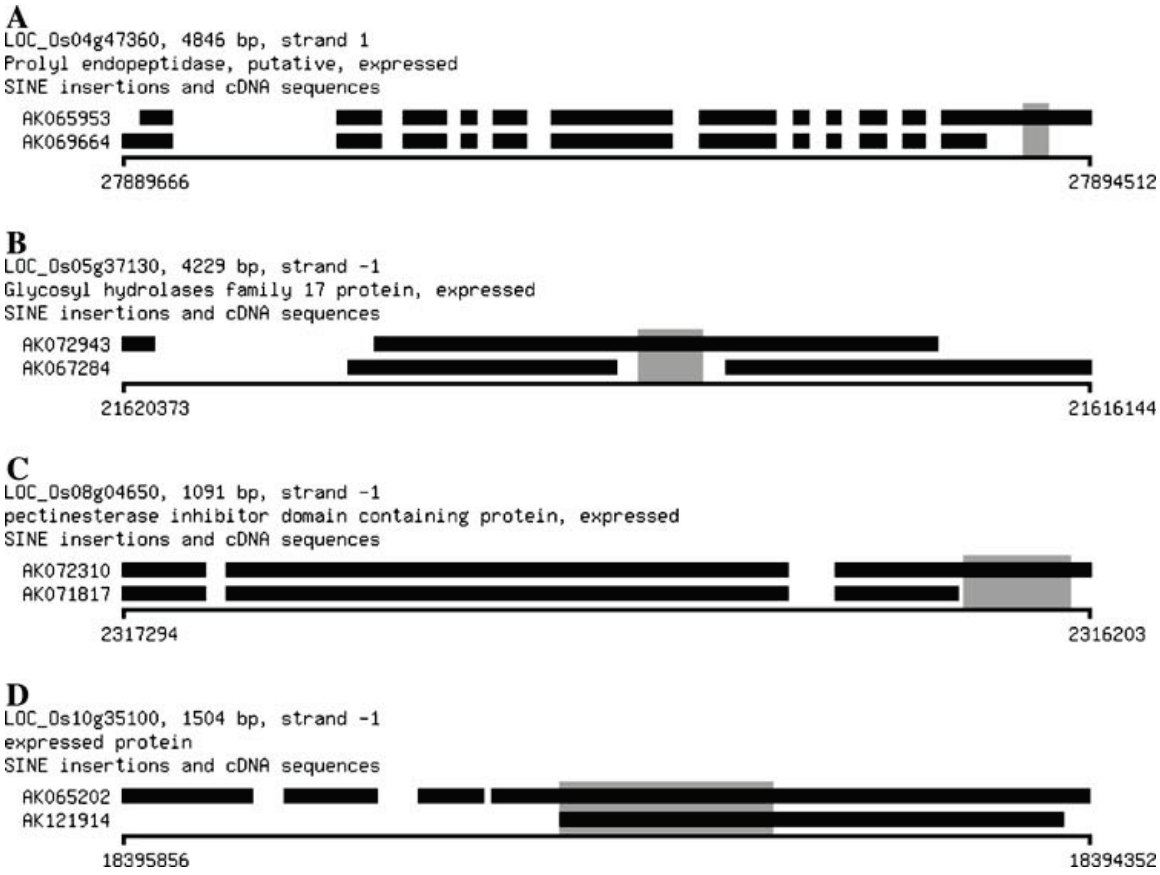


Figure 3: Examples of genes with SINE insertions. Black colored rectangles represent exons and grey colored rectangles represent SINEs. Fig. 3A-C. SINE is part of one fl-cDNA. Fig. 3D. SINE is part of more than one fl-cDNA.

Table 1 Genes associated with retrotransposons

Type of retrotransposon inserted	Number of genes with position of the retrotransposon			
	Promoter	Gene	Both	Total
LTR-retrotransposon				
Copia	467	478	127	818
Gypsy	1,097	714	255	1,556
LINE	348	506	39	815
SINE	929	628	55	1502

Total annotated genes were 56,551. The above analysis was conducted using 28,287 genes which were not annotated as TEs or hypothetical genes

Table 2 Chromosomal distribution of genes associated with retrotransposons

Chr	LTR-retrotransposon—*Copia*				LTR-retrotransposon—*Gypsy*				LINE				SINE				Genome	Expected fraction
	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z		
1	**44**	**−2.80**	49	−2.31	**97**	**−4.84**	**64**	**−3.81**	34	−2.22	77	0.87	103	−2.46	83	−0.48	3,927	0.139
2	39	−1.95	42	−1.67	*108*	−1.42	64	−1.89	32	−1.16	46	−1.48	*112*	0.87	75	0.63	3,156	0.112
3	**30**	**−3.76**	**38**	**−2.78**	**103**	**−2.75**	**56**	**−3.49**	*47*	0.82	**39**	**−3.02**	104	−0.82	72	−0.47	3,416	0.121
4	46	0.80	**57**	**2.41**	106	1.01	**82**	**2.53**	28	−0.51	47	0.37	74	−0.93	54	−0.2	2,497	0.088
5	39	0.36	32	−0.98	89	0.26	52	−0.61	28	0.07	30	−1.68	68	−0.71	43	−1.02	2,248	0.079
6	37	−0.22	36	−0.53	101	1.22	66	1.02	29	0.07	40	−0.26	79	0.31	64	1.79	2,327	0.082
7	40	0.62	31	−1.07	94	0.95	56	0.04	37	1.96	40	0.08	81	1.03	58	1.33	2,210	0.078
8	37	0.96	32	−0.09	76	0.17	**67**	**2.74**	19	−1.01	41	1.13	76	1.63	44	0.18	1,932	0.068
9	29	0.67	37	2.15	76	2.07	41	0.28	19	−0.02	40	2.39	**71**	**2.89**	**51**	**2.9**	1,551	0.055
10	34	1.58	35	1.64	73	1.52	49	1.47	24	1.08	35	1.32	49	−0.39	44	1.57	1,576	0.056
11	**43**	**2.48**	41	1.95	**96**	**3.18**	60	2.19	24	0.37	39	1.19	56	−0.48	46	0.93	1,814	0.064
12	**49**	**4.34**	**48**	**3.97**	78	1.86	57	2.50	27	1.59	32	0.53	56	0.33	48	2.01	1,633	0.058

Bolded values vary significantly from expected fraction. This data was normalized to reflect over- or under-representaion of genes or promoters with retrotransposon insertions for each of the 12 rice chromosomes, separately. Significance was determined using the binomial test with Bonferroni correction (cut-off P value of 0.00416). P value estimates: $|Z| > 3.7190$: $P < 0.0001$, $|Z| > 3.0902$: $P < 0.001$, $|Z| > 2.3263$: $P < 0.01$, $|Z| > 1.6449$: $P < 0.05$. Negative Z: less than expected. Positive Z: more than expected. Values in italics and underlined represent the highest and the lowest number of genes, respectively, on a chromosome not expected compared to the distribution of all genes in the genome

Table 3 Distance distribution of retrotransposons upstream of the start codon

Distance (bp) from translation start site	Number of genes with				fl-cDNA				Proportion of genes with cDNAs			
	Copia	*Gypsy*	LINE	SINE	*Copia*	*Gypsy*	LINE	SINE	*Copia*	*Gypsy*	LINE	SINE
1–100	136	245	20	19	15	54	6	7	0.11	0.22	0.30	0.37
101–200	27	43	24	54	8	17	17	33	0.30	0.40	0.71	0.61
201–300	30	66	36	98	13	29	27	65	0.43	0.44	0.75	0.66
301–400	45	85	34	132	29	47	23	98	0.64	0.55	0.68	0.74
401–500	29	90	43	142	19	59	29	100	0.66	0.66	0.67	0.70
501–600	46	114	53	116	24	73	36	83	0.52	0.64	0.68	0.72
601–700	29	121	41	116	15	72	25	84	0.52	0.60	0.61	0.72
701–800	42	136	40	116	24	98	25	84	0.57	0.72	0.63	0.72
801–900	56	123	44	91	30	91	33	61	0.54	0.74	0.75	0.67
901–1000	27	74	13	45	16	56	11	34	0.59	0.76	0.85	0.76

Table 4 Gene ontology classification of genes associated with retrotransposons

Biological process	ID	LTR-retrotransposon—*Copia*				LTR-retrotransposon—*Gypsy*				LINE				SINE				Genome	Expected fraction
		Promoter	Z	Gene	Z	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z		
Cellular process	GO:0009987	18	0.67	15	−0.20	34	−0.37	28	0.93	10	−0.44	12	−1.17	31	0.07	21	0.06	933	0.03
Development	GO:0007275	13	−1.05	10	−1.86	26	−2.33	19	−1.47	15	0.62	19	0.09	28	−1.08	23	−0.03	1,042	0.04
Growth	GO:0040007	0	−0.68	0	−0.69	1	−0.09	2	1.52	0	−0.55	2	2.32	2	1.30	2	1.94	25	0.00
Interaction between organisms	GO:0051704	0	−0.68	0	−0.69	0	−1.05	0	−0.85	0	−0.44	1	1.33	0	−0.73	**2**	**2.76**	16	0.00
Physiological process	GO:0007582	**99**	**−3.89**	113	−2.76	**270**	**−3.48**	**170**	**−3.28**	109	0.80	154	0.53	293	1.46	**229**	**3.91**	8,306	0.29
Regulation of biological process	GO:0050789	**15**	**−3.40**	**15**	**−3.50**	**44**	**−4.19**	34	−2.61	14	−2.35	33	−0.67	62	−0.73	38	−1.20	2,065	0.07
Reproduction	GO:0000003	8	−0.44	4	−1.82	**9**	**−2.79**	7	−1.95	7	0.06	12	0.66	21	0.65	13	0.19	556	0.02
Response to stimulus	GO:0050896	68	0.16	66	−0.31	136	−1.80	83	−2.04	59	1.39	77	0.56	132	−0.12	107	1.93	4,057	0.14
Viral life cycle	GO:0016032	0	−0.39	0	−0.39	0	−0.59	0	−0.48	0	−0.33	0	−0.40	1	1.30	1	1.79	9	0.00
Molecular function																			
Antioxidant activity	GO:0016209	1	2.06	0	−0.41	1	0.98	0	−0.50	0	−0.35	0	−0.42	0	−0.57	0	−0.47	10	0.00
Binding	GO:0005488	105	−0.87	92	−2.53	**201**	**−4.55**	**135**	**−3.30**	76	−1.02	122	−0.04	218	−0.52	148	−0.37	6,843	0.24
Catalytic activity	GO:0003824	**108**	**−3.15**	129	−1.34	299	−1.84	**177**	**−2.93**	**132**	**3.33**	175	2.37	295	1.32	**248**	**5.33**	8,422	0.30
Enzyme regulator activity	GO:0030234	1	−1.08	2	−0.51	8	0.55	1	−1.59	3	0.54	2	−0.67	6	0.05	2	−0.99	179	0.01
Motor activity	GO:0003774	0	−1.19	0	−1.20	4	0.39	2	−0.10	1	0.03	1	−0.35	0	−1.61	2	0.19	79	0.00
Nutrient reservoir activity	GO:0045735	1	0.07	0	−0.98	3	0.54	0	−1.20	2	1.47	0	−1.04	1	−0.69	1	−0.29	60	0.00
Signal transducer activity	GO:0004871	**16**	**3.73**	1	−2.22	12	−0.86	6	−1.27	4	−0.41	8	0.33	6	−1.97	5	−1.30	398	0.01
Structural molecule activity	GO:0005198	5	−0.76	1	−2.32	12	−1.11	7	−1.14	5	−0.06	4	−1.28	13	−0.20	5	−1.42	418	0.02
Transcription regulator activity	GO:0030528	**12**	**−3.75**	**9**	**−4.38**	**35**	**−4.95**	**27**	**−3.37**	12	−2.61	22	−2.35	53	−1.57	36	−1.26	1,985	0.07
Translation regulator activity	GO:0045182	2	−0.22	2	−0.25	4	−0.64	3	−0.30	2	0.18	2	−0.35	6	0.60	1	−1.22	143	0.01
Transporter activity	GO:0005215	13	−2.35	24	−0.18	**33**	**−3.27**	25	−2.04	15	−0.73	35	1.77	49	0.14	**59**	**4.77**	1,464	0.05
Cellular component																			
Cell part	GO:0044464	55	−1.03	63	−0.14	**110**	**−3.28**	**65**	**−3.37**	49	0.37	70	0.29	140	1.50	**109**	**2.91**	3,790	0.13
Extracellular matrix	GO:0031012	1	0.07	0	−0.98	2	−0.13	1	−0.36	1	0.56	0	−0.92	1	−0.44	2	0.94	47	0.00

Table 4 continued

Biological process	ID	LTR-retrotransposon—*Copia*				LTR-retrotransposon—*Gypsy*				LINE				SINE				Genome	Expected fraction
		Promoter	Z	Gene	Z	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z	Promoter	Z	Gene	Z		
Extracellular region	GO:0005576	7	**3.06**	0	−1.55	6	0.22	0	−1.89	2	0.09	3	0.16	3	−0.91	4	0.33	153	0.01
Extracellular region part	GO:0044421	0	−0.68	0	−0.69	0	−1.05	0	−0.85	0	−0.53	0	−0.64	1	0.28	1	0.69	23	0.00
Organelle	GO:0043226	**35**	**−3.48**	**36**	**−3.50**	**93**	**−4.37**	**60**	**−3.58**	37	−1.29	55	−1.39	119	−0.12	88	0.80	3,662	0.13
Organelle part	GO:0044422	13	−1.62	17	−0.80	36	−1.66	**16**	**−2.71**	12	−0.77	19	−0.59	49	1.49	19	−1.56	1,212	0.04
Protein complex	GO:0043234	17	−0.99	14	−1.74	49	−0.21	20	−2.29	13	−0.78	22	−0.28	38	−0.76	24	−0.95	1,305	0.05

Only those classes containing at least one gene of interest are included. Bolded values vary significantly from expected fraction. Significance was determined using binomial test with Bonferroni correction (cut-off P value of 0.0055 for biological processes, 0.0045 for molecular functions, and 0.0071 for cellular components). P value estimates: $|Z| > 3.7190$: $P < 0.0001$, $|Z| > 3.0902$: $P < 0.001$, $|Z| > 2.3263$: $P < 0.01$, $|Z| > 1.6449$: $P < 0.05$. 'Genome' refers to the total set of rice genes annotated as non-TE or hypothetical genes and classified using gene ontology

Table 5 Expression analysis of genes associated with retrotransposons

Genes with	*Copia* insertion		*Gypsy* insertion		LINE insertion		SINE insertion	
	Promoter	Gene	Promoter	Gene	Promoter	Gene	Promoter	Gene
Full-length (fl) cDNA	164	212	438	283	183	277	522	380
MPSS	104	140	354	206	134	230	390	286
MPSS and fl-cDNA	75	94	196	128	85	148	263	192
MPSS and/or fl-cDNA	193	258	596	361	232	359	649	474
Retro in fl-cDNA	15	55	38	108	11	53	9	38

CHAPTER 2:

COMPARATIVE ANALYSIS OF DIVERGENT AND CONVERGENT GENE PAIRS AND THEIR EXPRESSION PATTERNS IN RICE, *ARABIDOPSIS*, AND *POPULUS*

Nicholas Krom and Wusirika Ramakrishna

Previously published in *Plant Physiology* (www.plantphysiol.org), 2008, 147: 1763-1773. Published online May 30, 2008.

This work was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number 2007-35301-18036.

2.1 ABSTRACT

Comparative analysis of the organization and expression patterns of divergent and convergent gene pairs in multiple plant genomes can identify patterns that are shared by more than one species or are unique to a particular species. Here, we study the coexpression and inter-species conservation of divergent and convergent gene pairs in three plant species: rice, *Arabidopsis*, and *Populus*. Strongly correlated expression levels between divergent and convergent genes were found to be quite common in all three species, and the frequency of strong correlation appears to be independent of intergenic distance. Conservation of divergent or convergent arrangement among these species appears to be quite rare. However, conserved arrangement is significantly more frequent when the genes display strongly correlated expression levels or have one or more Gene Ontology (GO) classes in common. A correlation between intergenic distance in divergent and convergent gene pairs and shared GO classes was observed, in varying degrees, in rice and *Populus* but not in *Arabidopsis*. Furthermore, multiple GO classes were either over-represented or under-represented in *Arabidopsis* and *Populus* gene pairs while only two GO classes were under-represented in rice divergent gene pairs. Three *cis*-regulatory elements common to both *Arabidopsis* and rice were over-represented in the intergenic regions of strongly correlated divergent gene pairs compared to those of non-correlated pairs. Our results suggest that shared as well as unique mechanisms operate in shaping the organization and function of divergent and convergent gene pairs in different plant species.

2.2 INTRODUCTION

Gene rearrangements occur frequently during the evolution of prokaryotic and eukaryotic genomes. The number of rearrangements appears to be a function of the phylogenetic distance between the organisms being studied. Rice and *Arabidopsis* are the model monocot and dicot genomes that have been fully sequenced (Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005). Recently, a second dicot plant genome, *Populus trichocarpa,* has been sequenced (Tuskan et al., 2006). Divergence time between *Populus* and *Arabidopsis* is estimated to be 100-120 million years ago (mya) and that of *Arabidopsis* and rice is 130 to 200 mya (Wolfe et al., 1989; Chaw et al., 2004; Tuscan et al., 2006). Very little collinearity in gene order has been observed between *Arabidopsis* and rice due to the large evolutionary distance that separates them (Devos et al., 1999; Liu et al., 2001; Vandepoele et al., 2002). Despite this lack of collinearity, at the level of single genes, 71% of protein coding rice genes had homologs in *Arabidopsis* genome compared to 90% of *Arabidopsis* genes with homologs in the rice genome (International Rice Genome Sequencing Project 2005).

Eukaryotic genes appear to be distributed in a nonrandom fashion with clustered genes exhibiting coordinated expression patterns (Hurst et al., 2004). Different trends of coexpression were observed depending on the types of genes and organisms. Strong positive correlation was observed in the expression patterns of divergent gene pairs compared to weak or no correlation in those of convergent gene pairs in *C. elegans* (Chen and Stein 2006). This was attributed to RNA transcripts from convergent genes obstructing each other by base pairing at their 3' ends (Katayama et al., 2005). Although coexpression patterns were observed in both divergent as well as convergent genes in yeast, divergent gene pairs displayed higher correlation than convergent gene pairs (Cohen et al., 2000; Kruglyak and Tang 2000). Significant numbers of pairs of adjacent genes have been found to have strongly correlated expression levels in *Arabidopsis* (Williams and Bowles 2004). Local domains of two to four highly coexpressed genes have also been identified in *Arabidopsis* (Ren et al., 2005), as have higher-order domains corresponding to regions of euchromatin (Zhan et al., 2006). Additionally, correlated expression of neighboring genes appears to be more common when both genes in a pair are classified in the same functional category (Williams and Bowles 2004). Correlated expression patterns of divergent or convergent genes might result due to *cis*-acting enhancers and/or their involvement in the same or related biological process/pathway as determined by Gene Ontology classification. Furthermore, chromatin organization can regulate coexpression as seen in case of coordinated expression of two transgenes in tobacco due to an artificial chromatin domain (Mlynarova et al., 2002). Although the tendency for neighboring genes to be coexpressed is well documented in *Arabidopsis*, little is known about this phenomenon in other plant species.

In the present study, bioinformatic analysis was performed to identify divergent and convergent gene pairs, using the three completely sequenced plant genomes, *Oryza sativa, Arabidopsis thaliana,* and *Populus trichocarpa*. Coexpression of gene pairs was determined based upon Pearson correlation coefficients calculated using Massively Parallel Signature Sequencing (MPSS) and microarray expression data. Gene pair conservation of each species' divergent and convergent genes with the whole genome sequences of the other two species was determined using BLASTP and TBLASTN. Furthermore, the effect of intergenic distance on the likelihood of both genes in a pair to be expressed (as evidenced by MPSS and/or microarray data) was investigated. Subsequently, GO classification of these gene pairs was used to identify over- and under-represented classes. Finally, we identified regulatory elements over-represented in the intergenic regions of gene pairs whose expression levels are strongly correlated to determine the basis of the observed coexpression.

2.3 RESULTS

Differential Variation in Divergent and Convergent Gene Numbers with Intergenic Distances in Rice, *Arabidopsis*, and *Populus*

Rice, *Arabidopsis*, and *Populus* gene annotation data was analyzed for pairs of adjacent genes arranged divergently (← →) and convergently (→ ←). Release 4 of the TIGR rice (*Oryza sativa* ssp. *japonica*) pseudomolecules contains a total of 56,563 annotated genes. Discarding hypothetical or transposon-related genes leaves 28,287 genes for further analysis. Among these, a total of 8,742 divergent and 8,772 convergent gene pairs were identified. Only in a minority of these pairs are the two genes separated by a short distance, with approximately one seventh of divergent pairs and one third of convergent pairs having 1 kb or less between them (Table I).

In *Arabidopsis thaliana*, the analysis was performed on 24,019 genes after filtering out hypothetical and transposon-related genes from 30,001 annotated genes. A total of 5,763 divergent gene pairs were identified, of which about 36% are separated by 1 kb or less. Among the 4949 convergent pairs discovered, 71% were separated by less than 1 kb. Version 1.1 of the JGI annotation of the *Populus trichocarpa* genome lists 45,554 genes. This dataset was not filtered for hypothetical or transposon-related genes, as no predicted functions were given. In all, 8823 divergent gene pairs were identified, accounting for 39% of the genome. Of these, 613 pairs (7%) were separated by less than 1 kb. A total of 8967 convergent gene pairs were identified, of which 2212 (25%) were separated by less than 1 kb. These results show a similar trend in the decrease in the fraction of divergent genes with decreasing intergenic distance from <1 kb to <250 bp in all the three species. However, *Populus* showed significant decrease in the fraction of convergent genes when compared to rice and *Arabidopsis* when the intergenic distance was decreased from <1 kb to <250 bp. Similarly, rice showed a significantly smaller decrease in the fraction of convergent genes compared to *Arabidopsis* and *Populus*. Furthermore, convergent genes were found to be 2 to 4-fold higher compared to divergent genes separated by <500 bp in all the three plant genomes.

An interesting observation was made when the results for the three species were compared. The fraction of gene pairs separated by a small distance (<1 kb) appears to be proportional to genome size. In *Arabidopsis*, with a 115 Mb genome, more than one third of divergent pairs are separated by <1 kb, compared to rice, with a 450 Mb genome, where only one seventh of divergent gene pairs show the same pattern, despite there being only 14% more genes under consideration in rice. This trend is even more pronounced in *Populus*, where about 7% of divergent gene pairs are separated by <1 kb, almost half as many as in rice, despite having far more genes (45,554 vs. 28,287) and a substantially larger genome (550 Mb). Similar observations were made when

comparisons involved convergent gene pairs. This relationship between gene pair 'compactness' and genome size is clearly non-linear which may be a result of the genome biology of the three plant species or differences in gene annotation methods.

Expression and Coexpression Patterns of Divergent and Convergent Genes Differ in Rice, *Arabidopsis*, and *Populus*

Several types of expression data were compiled for divergent and convergent gene pairs. Our analysis had both qualitative and quantitative aspects. The qualitative analysis confirmed that both genes were in fact expressed and not annotation artifacts. The goal of the quantitative analysis was to determine which gene pairs showed correlated expression levels across multiple tissues and treatments.

Divergent and convergent gene sequences were aligned with EST and full-length cDNA (fl-cDNA) sequences using BLASTN. In rice, the fraction of gene pairs for which matching EST/fl-cDNA sequences were found for both genes increases with decreasing intergenic distance (Figures 1A and 1B). This trend is more pronounced in case of divergent gene pairs. Both the strong negative correlation between intergenic distance and EST/fl-cDNA matches seen in rice and the weak correlation in *Arabidopsis* are non-existent in *Populus*. This phenomenon may be due to differences in regulatory mechanisms or the availability of fewer EST/fl-cDNA sequences for *Populus* compared to rice and *Arabidopsis*.

Analysis of Massively Parallel Signature Sequencing (MPSS) and microarray data revealed that the fraction of divergent and convergent pairs with expression data for both genes increases significantly with reduced intergenic distance in rice (Tables II and III). Interestingly, a pronounced increase in the fraction (32% to 65% for MPSS and 53% to 81% for microarray data) of divergent pairs with expression data for both genes was observed compared to a modest increase in the fraction (43% to 58% for MPSS and 69% to 79% for microarray data) of convergent genes in rice when the intergenic distance was reduced from 1 kb to 250 bp. A similar trend is seen in *Arabidopsis*, although the increases observed are not as pronounced as in rice and are only statistically significant for MPSS data. For *Populus*, microarray expression data coverage actually decreases somewhat for divergent pairs, and increases only slightly for convergent pairs. Altogether, there was a significant increase in the fraction of rice divergent and convergent gene pairs with fl-cDNA/EST, MPSS or microarray expression data with decreasing intergenic distance. This trend was not seen in the other two genomes except in the case of *Arabidopsis* gene pairs with MPSS data.

Correlated expression of genes in divergent and convergent pairs was examined based on the Pearson correlation of their MPSS expression levels. Gene pairs with correlation coefficients greater than 0.5 were considered to be significantly coexpressed, while those with coefficients less than -0.5 were considered antiregulated (i.e. expression of one gene precludes expression of the other). Strong positive correlation was observed in approximately 2% of rice divergent and convergent pairs. In *Arabidopsis,* 12% of divergent pairs and 10% of convergent pairs showed strong positive correlation, while less than 1% of either pair type was antiregulated. No statistically significant connection between intergenic distance and frequency of correlated expression was noted in either species. The mean Pearson correlation of all rice divergent pairs for which MPSS data was available was 0.112, and the same figure for convergent pairs was 0.108. The mean correlation for 3,000 randomly paired rice genes was found to be 0.013, far lower than that of divergent or convergent pairs. In *Arabidopsis*, the average correlation for divergent pairs was 0.247, and 0.235 for convergent pairs. The mean correlation of 3,000 random gene pairs was 0.098, again significantly lower than the average correlation of divergent and convergent gene pairs. These data support the hypothesis that genes in divergent or convergent arrangement are more likely to be coexpressed than random pairs of genes.

The Pearson correlation of divergent and convergent gene pairs was also calculated using the mean intensity levels of each gene's corresponding probes across multiple microarray hybridizations. In rice, 26% of divergent pairs with microarray data for both genes had correlation coefficients greater than 0.5 compared to 10% and 49%, respectively, of *Arabidopsis* and *Populus* pairs (Table III). Similar results were found for convergent gene pairs, with 26%, 9%, and 48% of rice, *Arabidopsis,* and *Populus* pairs, respectively, showing high levels of correlation. While slight increases in the fraction of pairs showing strongly correlated expression can be seen as intergenic distance decreases, these changes are not statistically significant. A great deal of variation between species can be noted with regard to the frequency with which gene pairs are strongly correlated. About half of *Populus* divergent and convergent gene pairs show strong correlation, compared to one-fourth of rice and one-tenth of *Arabidopsis* gene pairs. A partial list of strongly correlated gene pairs is given in Supplemental Tables S1 and S2. Strong negative correlation appears to be quite rare, with only 1.6% of all *Arabidopsis* divergent gene pairs and 1.5% of *Populus* divergent gene pairs having Pearson correlation coefficients less than -0.5. Similar percentages of convergent gene pairs display strong negative correlation in *Arabidopsis* and *Populus*.

To determine the degree to which divergent and convergent arrangement affects coexpression, the mean correlation levels of divergent and convergent gene pairs were calculated and compared with that of a set of 8,000 randomly selected pairs of genes.

The mean correlation of rice divergent gene pairs calculated with microarray expression data was 0.390, and for convergent pairs the same figure was 0.392. The mean correlation of the random set was 0.103, approximately one-quarter that of either type of gene pair. In *Arabidopsis*, the mean correlation for divergent pairs was 0.163, and 0.144 for convergent pairs. Both values are significantly higher than the mean correlation for the *Arabidopsis* random pairs, which was calculated to be 0.044. This pattern was repeated in *Populus*, where the mean correlation for divergent and convergent gene pairs was 0.486 and 0.481, respectively, compared to 0.155 for the set of random gene pairs. These results indicate that in all three species, divergent and convergent gene pairs display significantly higher levels of correlated expression than randomly paired genes. For all three organisms, the mean correlation of both pair types is about three to four-fold higher than that of the random sets, suggesting that the degree to which divergent and convergent arrangement affects coexpression of neighboring genes compared to random sets does not vary greatly among these species. Interestingly, the mean Pearson correlation for the expression of divergent or convergent genes are about 3 and 2.5 times higher in *Populus* and rice, respectively compared to *Arabidopsis*. While it is possible that this may reflect biological differences between the species, it is more likely an artifact of the variation in the number of microarray hybridizations analyzed for each species (2829 in *Arabidopsis*, 446 in rice, and 150 in *Populus*). Since the Pearson correlation was calculated using paired data points from each gene across all hybridizations, a larger number of hybridizations would lower the probability of obtaining a high correlation coefficient.

Differential Inter-species Conservation of Divergent and Convergent Gene Arrangement

Conserved divergent or convergent arrangement of genes across species separated by vast evolutionary distances suggests conserved functional interaction between the proteins encoded by the genes in a pair. Six sets of BLASTP and TBLASTN searches were performed, aligning divergent and convergent genes from rice, *Arabidopsis*, and *Populus* with the genomes of the other two species. Conserved rice divergent gene pairs were found to be rare in both *Arabidopsis* and *Populus*, with only 26 pairs conserved in *Arabidopsis* and 77 in *Populus* (Table IV). For convergent pairs, conservation levels were found to be slightly higher at 42 pairs in *Arabidopsis* and 111 pairs in *Populus*. Examining only those pairs with short intergenic distances showed slight increase in conserved divergent pairs, while the fraction of conserved convergent pairs nearly doubled (0.8% in *Arabidopsis* and 2.6% in *Populus*) when intergenic distance is <250 bp. The frequency of *Arabidopsis* gene pair conservation varied greatly in rice and *Populus*. In rice, only 29 divergent gene pairs were found to be conserved compared to 52

convergent pairs. *Arabidopsis* gene pairs conserved in *Populus* were found to be far more common, with 355 divergent pairs and 401 convergent pairs having conserved gene order and orientation. Comparison of *Populus* gene pairs with rice and *Arabidopsis* identified 58 and 267 divergent pairs conserved in rice and *Arabidopsis*, respectively. Among *Populus* convergent gene pairs, 114 were conserved in rice, while 421 were conserved in *Arabidopsis*. In each of these comparisons, the number and fraction of conserved convergent gene pairs were higher than those of conserved divergent pairs. These results suggest that the exact spatial arrangement of the gene pair is a necessary regulatory factor in only a small fraction of all such pairs.

Divergent and convergent gene pairs were found to be conserved in some species more frequently when both genes shared one or more Gene Ontology terms. Rice divergent and convergent gene pairs with shared GO terms were found to be more likely to be conserved in *Arabidopsis* or *Populus* compared to all divergent and convergent pairs (Table V). The fraction of *Populus* gene pairs with shared GO terms organized in divergent manner in rice and convergent manner in *Arabidopsis* increased significantly compared to all gene pairs. This trend was not observed in *Arabidopsis* gene pairs conserved in other two plant genomes. These results suggest that divergent and convergent genes with shared GO are more likely to be conserved compared to all conserved gene pairs.

Strongly correlated expression also raises the probability of a gene pair being conserved. While the increases in conservation frequency is seldom as great as those caused by shared GO terms, they are nonetheless quite significant, especially in case of rice divergent gene pairs conserved in *Arabidopsis* or *Populus*, where up to three-fold increases were observed (Table V). Similarly, a two-fold change was observed on the conservation of rice convergent genes in *Arabidopsis*, with correlated expression. This trend of few fold changes was not observed in other comparisons. This data indicates that the effect of strongly correlated expression on the conservation of divergent and convergent genes varies based on the organisms being examined. A partial list of conserved gene pairs, and conserved gene pairs displaying correlated expression, is given in Supplemental Tables S3 – S6.

Gene Ontology Classification of Divergent and Convergent Genes

Gene Ontology (GO) classification data was downloaded for all genes included in our analysis. While at least one GO classification was found for about 99% of *Arabidopsis* divergent and convergent genes, in rice only 41% of divergent genes and 45% of convergent genes had similar data available due to the ongoing process of GO

classification of the rice genome. A similar situation exists in *Populus*, where approximately 45% of both divergent and convergent genes have at least one GO classification. The full GO vocabulary was used in classifying the *Populus* genes, while the Plant GOslim vocabulary was used for rice and *Arabidopsis*.

Two analyses were performed on those pairs for which GO classification data was found for both genes. The first of these was a search for pairs in which both genes were grouped into the same GO class, as shared or related function could be a contributing factor in the coordinated expression of neighboring genes. Approximately 4.9% of rice divergent genes separated by <1 kb, have at least one GO class in common, and this percentage increased to 7.5% for genes separated by <250 bp. Among rice convergent genes, this percentage rose from 11.3% for genes separated by <1 kb to 15% for genes separated by <250 bp. A similar pattern is seen in *Populus*, but the effect of decreased intergenic distance is much weaker, with the fraction of pairs with shared GO classes increasing only from 2.9% to 4.3% among divergent pairs, and from 2.1% to 2.2% among convergent pairs. In *Arabidopsis*, the fraction of pairs with common GO classifications remained nearly constant at about 45% across different intergenic distances, for both divergent and convergent genes. These results suggest that the likelihood of the genes in a pair sharing the same GO class increases greatly in rice, to a lesser degree in *Populus* but not in *Arabidopsis*, if the genes are physically closer to each other.

The second analysis sought out GO classes that were disproportionately represented among divergent or convergent genes relative to the whole genome. Over- or under-representation was determined using the binomial test (normal approximation, $P < 0.0001$). In rice, two GO classes were found to be significantly under-represented among divergent genes. Genes whose protein products are involved in secondary metabolic and biosynthetic processes are under-represented in rice divergent pairs. Among others, several cytochrome P450 and gylcosyl hydrolase family proteins are part of GO class of secondary metabolic process. No over-represented GO classes were identified among rice divergent or convergent genes. Several over- and under-represented GO classes were found in *Arabidopsis* and *Populus* divergent and convergent gene pairs. GO class nucleic acid binding which includes zinc finger family proteins and translation initiation factors was found to be under-represented in both divergent and convergent gene pairs in *Arabidopsis*. However, genes belonging to this same GO class are over-represented in *Populus* divergent gene pairs. GO class signal transduction is over-represented in *Arabidopsis* divergent genes which includes several leucine-rich repeat family proteins and ethylene-responsive factors. Interestingly, GO classes apoptosis, defense response, and transmembrane receptor activity were under-represented in both divergent and convergent genes of *Populus*. Gene pairs over-represented in specific GO classes suggest

that they are more likely to be organized in a divergent or convergent manner. Similarly, under-represented GO classes suggest that genes belonging to these classes do not tend to be organized in a specific orientation (divergent or convergent). Although the reason for this bias is not known, it is possible that functional relationships exist among these genes. A full listing of these classes can be found in Supplemental Table S7, and the number of each species' correlated or conserved pairs in each GO class is given in Supplemental Tables S8, S9, and S10.

Regulatory Elements Over-represented in Intergenic Regions of Divergent Genes with Correlated Expression

The intergenic regions of all divergent and convergent pairs separated by 1 kb or less were analyzed for known regulatory elements using the Plant *Cis*-acting Regulatory DNA Element (PLACE) database (Higo et al., 1999; http://www.dna.affrc.go.jp/PLACE/index.html). In addition, 1 kb regions upstream of convergent genes were examined for the presence of regulatory elements. For each species and pair type, the gene pairs were divided into two subsets: those displaying strongly correlated expression and those with weak or no correlation. The fractions of sequences in these two sets containing each element were then compared, and their differences were tested for statistical significance using the binomial test ($P < 0.0001$).

In *Arabidopsis* and rice divergent gene pairs, several elements were found to be over-represented among strongly correlated pairs (Table VI). This differs significantly from the results obtained for convergent pairs, where none of the elements were found to be over-represented. These results suggest that correlated expression in divergent gene pairs is at least in part caused by the presence of specific regulatory elements in the intergenic region, where they can influence the expression of both genes in the pair. While similar numbers of elements were found in the intergenic regions of divergent and convergent pairs, we found no significant difference in the elements found between correlated and non-correlated convergent pairs. Therefore, it seems likely that correlated expression due to shared regulatory elements is a feature only of divergent gene pairs. A complete list of all over-represented regulatory elements identified can be found in Table S11.

The results for *Populus* were quite different from those for rice and *Arabidopsis*. Although many elements were identified in the *Populus* sequences, very few showed any significant difference in frequency between correlated and non-correlated pairs. This is most likely a reflection of the composition of the PLACE database, which contains regulatory elements gleaned from recent publications. As rice and *Arabidopsis* have been

more thoroughly studied than *Populus*, there may be many regulatory elements in the *Populus* genome involved in correlated expression, as we hypothesized there to be in rice and *Arabidopsis* that are not represented in the PLACE database.

Three of the regulatory elements identified were over-represented in the intergenic regions of coexpressed divergent gene pairs in both rice and *Arabidopsis*. These three elements are CGACG element required for the expression of rice alpha-amylase Amy3D gene (Hwang et al., 1998), E2F consensus sequence recognized by E2F transcription factors and present in promoters of target genes that regulate cell cycle, DNA replication, DNA repair, and chromatin structure (Vandepoele et al., 2005), and a sulfur-responsive element (SURE) core sequence present in the promoter region of a sulfate transporter gene of *Arabidopsis* (Maruyama-Nakashita et al., 2005). This last element overlaps largely with an auxin response element. Furthermore, one over-represented element, PRECONSCRHSP70A is shared by *Populus* and *Arabidopsis* promoters flanked by coexpressed divergent genes. This is the consensus sequence of a plastid response element (PRE) found in the promoter of nuclear gene *HSP70A* in *Chlamydomonas* and induced by a chlorophyll precursor, Mg-protoporphyrin and light (Von Gromoff et al., 2006). Furthermore, the most over-represented element in promoters of *Arabidopsis* correlated divergent gene pairs is UP2 motif which is found upstream of genes up-regulated on main stem decapitation (Tatematsu et al., 2005). GCC-box core found in many pathogen-responsive genes (Brown et al., 2003) was the most over-represented element in rice promoters with correlated divergent gene pairs. The occurrence of these elements between strongly correlated genes suggests that they play a role in regulating both genes in the pair, with either the elements being shared as part of a single bidirectional promoter or having a similar set of regulatory elements present in each gene's separate promoter.

2.4 DISCUSSION

With the recent completion of several plant genomes and the availability of genome-wide quantitative expression data, it is possible to unravel many of the unexplained aspects of the inner workings of complex organisms. Here, we investigated the organization of convergent and divergent genes in three plant genomes, their expression patterns and the degree of coexpression exhibited by them. Our study not only identified similar patterns with respect to the organization of divergent and convergent genes with decreasing intergenic distance in the three plant genomes but also cases where a pattern is unique to only one of the three plant genomes. It is very likely that some of these divergent trends are linked to the biology of the specific organism. This is further illustrated by over- and/or under-represented GO classes either shared by divergent and convergent genes or unique to them in one or more species, which is related to their function.

In rice, it was observed that the fraction of divergent and convergent gene pairs for which expression data exists for both genes increases as the distance between the two genes decreases. However, this phenomenon was not observed in either *Arabidopsis* or *Populus,* which may be caused by biological differences between monocots and dicots. This needs to be confirmed by the study of several other monocot and dicot genomes. Some of these differences can also be attributed to the source of the expression data based on different results obtained for *Arabidopsis* with the three data sets of fl-cDNA/EST, MPSS and microarray.

Our comparative analysis identified a number of divergent and convergent gene pairs in rice, *Arabidopsis*, and *Populus* that possess homologs in the same orientation in at least one other species. The fraction of conserved gene pairs ranges from 0.3% to 8.1% across species and pair types, which is in accord with the results of earlier studies. Seoighe and colleagues (2000), performing a similar analysis on two yeast species, found that only 9% of *S. cerevisiae* gene pairs remained adjacent, and of those 65% maintained the same orientations, leaving only 5.85% of all gene pairs conserved with regard to both gene order and relative orientation. Comparisons between rice and *Arabidopsis* (Liu et al., 2001) identified a rate of 5.5% for conservation of gene pair order, and a probability of only 0.005 for the pair to be conserved without additional genes being inserted between them. Ren and colleagues (2007) found no local coexpression domains in rice that were conserved in *Arabidopsis*. However, their criteria for coexpression ($R > 0.7$) were different than those used here, and a coexpression domain was only considered conserved if the homologous domain was also coexpressed. Therefore, there may be coexpressed divergent or convergent gene pairs which are conserved in our study but not in their study. Together with our findings, these results indicate that exact conservation of gene pair order and orientation between species is quite rare. This rarity, however,

would seem to imply that when divergent or convergent arrangement is conserved there is likely to be some regulatory aspect to that arrangement necessary for proper gene function, such as bidirectional promoters or enhancer sequences in the pair's intergenic region. Bidirectional promoters have been identified and characterized in relatively large numbers in the human genome (Adachi and Lieber 2002; Trinklein et al., 2004; Lin et al., 2007), yet have received little attention in plants. The over-representation of some regulatory elements in the intergenic regions of strongly correlated divergent gene pairs supports the hypothesis that shared elements are responsible, at least in part, for the coordinated expression observed in many divergent pairs. These elements may have novel mechanisms for regulating these genes. This explanation, however, does not apply to convergent gene pairs, despite the similar frequency of correlated expression observed among them. Other factors such as local chromatin organization may be responsible for the coexpression observed especially in case of convergent gene pairs. This study provides a foundation for more detailed studies of the regulatory elements involved in coordinating the expression of divergent and convergent gene pairs.

Two factors have been identified that affect the probability of a divergent or convergent gene pair being conserved in other species. Gene pairs that have one or more GO classifications in common are more likely to be conserved in another species. The second factor that increases the likelihood of a gene pair being conserved is strong coexpression. This association is most likely due to a shared or similar function of the genes in a pair.

The functional basis for the high level of coexpression observed in many divergent and convergent gene pairs can take on myriad forms. One of the most straightforward is involvement in the same biological process, a situation observed in numerous gene pairs based on the frequency of shared GO classifications. One such divergent gene pair found in rice consists of a phospho-2-dehydro-3-deoxyheptonate aldolase 1 and a cytokinin-O-glucosyltransferase 2 gene. Both genes are in the GO class "amino acid and derivative metabolic process", and the pair is strongly correlated (R = 0.62) and conserved in *Arabidopsis*. Another cause of gene pair coexpression is shared regulatory elements, which would induce the expression of both genes in response to a single stimulus. An example of such a gene pair is found on chromosome 1 in *Arabidopsis*. The genes in this divergent pair code for two auxin-responsive / indoleacetic acid-induced proteins, IAA3 and IAA17, which display correlated expression levels (R = 0.65) and are conserved in rice. One commonly observed trend is shared or similar molecular functions among genes in a divergent or convergent pair. The rice convergent gene pair consisting of a serine/threonine protein phosphatase PP2A catalytic subunit and a phosphatidic acid phosphatase family protein is an example of this. Both genes, in addition to being annotated as phosphatases, have the GO

classification “hydrolase activity” and are very strongly correlated (R = 0.78). Convergent arrangement of this pair is conserved in both *Arabidopsis* and *Populus*, which, along with all other such indicators, suggests very compellingly that these genes have some type of functional relationship and that their convergent arrangement is an essential part of their regulation. A similar set of circumstances surrounds the rice divergent gene pair consisting of a sugar transporter family protein and a protein kinase domain containing protein. According to their GO classifications, the products of both genes are located in the nuclear membrane. The pair is conserved in *Arabidopsis* and has a Pearson correlation of 0.73, which suggests that a functional relationship exists between the two genes. No such relationship is indicated in the available data, so the data relating to this gene pair generated in this study could serve as inspiration for further study of this and other similar pairs.

2.5 CONCLUSIONS

We identified patterns of expression and coexpression patterns of divergent and convergent gene pairs in rice, *Arabidopsis*, and *Populus*. Strongly correlated expression was observed in significant numbers of gene pairs in all three species, and at significantly higher levels than randomly paired genes. Cross-species conservation of divergent and convergent arrangement was found to be low, although the frequency of conservation was significantly higher among pairs with strongly correlated expression or shared Gene Ontology classifications. We identified several coexpressed gene pairs with shared GO terms suggesting functional correlation. Furthermore, we identified a few regulatory elements that may be involved in coordinating the expression of divergently arranged genes. In all, patterns of divergent and convergent gene pair coexpression and conservation were characterized, and several factors that influence these phenomena were identified, providing a foundation for more detailed study of the various mechanisms of regulating these genes.

2.6 METHODS

Identification of Divergent and Convergent Gene Pairs

Sequence and annotation data for the *Oryza sativa* ssp. *japonica* (cultivar Nipponbare) genome were downloaded from the Rice Genome Annotation Database at The Institute for Genomic Research (TIGR) (http://www.tigr.org/tdb/e2k1/osa1). Similar data for the *Arabidopsis thaliana* and *Populus trichocarpa* genomes was obtained from The Arabidopsis Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Genes/TIGR5_genome_release) and Joint Genome Institute (JGI) (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) websites, respectively. A Perl script was used to parse this data and identify pairs of adjacent genes on opposite strands, designating genes arranged head-to-head as divergent pairs and those arranged end-to-end as convergent pairs. Pairs containing genes annotated as hypothetical or transposon-related were excluded from all later analyses.

Analysis of Gene Pair Expression

EST data was downloaded for *Arabidopsis* (The Arabidopsis Information Resource [TAIR] EST FTP site: ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/), rice (Rice Full-length cDNA Consortium: http://cdna01.dna.affrc.go.jp/cDNA), and *Populus* (PopulusDB: http://poppel.fysbot.umu.se/proj_downl.php), and converted into BLAST databases. Genes in convergent and divergent pairs were aligned with the EST/fl-cDNA data using BLASTN. Hits with at least 95% identity were deemed significant and used, along with other types of expression data, to determine if annotated genes were actually expressed or false positives from gene prediction. In addition, *Arabidopsis* EST and fl-cDNA alignment data was downloaded from the Salk Institute Genomic Analysis Laboratory (http://signal.salk.edu/data) and was used to assign additional matches to *Arabidopsis* divergent and convergent genes.

Massively Parallel Signature Sequencing (MPSS) (Meyers et al., 2004) data was collected for rice (http://mpss.udel.edu/rice/) and *Arabidopsis* (http://mpss.udel.edu/at/) genes. Only 17 bp signatures of classes 1, 2, 5, and 7 that mapped to a single gene were used, and abundance values less than 5 were ignored as background interference. When multiple signatures had significant abundance values in the same library, the average abundance was used. Correlated expression between genes in divergent and convergent gene pairs was examined by calculating the Pearson correlation coefficient using each gene's average abundance values across multiple libraries (17 in *Arabidopsis*, 72 in rice).

Microarray data for all three species was compiled from several sources (Rice: Yale rice project [http://bioinformatics.med.yale.edu/rc/overview.jspx], *Arabidopsis*: Nottingham Arabidopsis stock centre microarray database [http://affymetrix.arabidopsis.info/narrays/help/usefulfiles.html] and Stanford Microarray Database [http://genome-www5.stanford.edu/], *Populus*: NCBI Gene Expression Omnibus [http://www.ncbi.nlm.nih.gov/geo/]). Expression data was collected for a total of 2829 hybridizations in *Arabidopsis*, 446 hybridizations in rice, and 150 hybridizations in *Populus*. Both the rice and *Arabidopsis* datasets included mappings of microarray spots to gene locus identifiers, while probe sequences on the *Populus* oligo arrays were aligned with the coding region sequences of *Populus* divergent and convergent genes using BLASTN. Oligos which aligned uniquely with 100% identity were inferred to be associated with individual genes. Correlated expression was again tested with the Pearson correlation coefficient, this time pairing data points from the same hybridization and channel.

Conservation of Gene Pair Arrangement

The protein sequences of all genes in divergent and convergent pairs from each species (rice, *Arabidopsis,* and *Populus*) were aligned with the full set of protein sequences from the two remaining species using BLASTP (Altschul et al., 1997) to identify homologs. If a divergent or convergent gene pair possessed homologs in the same arrangement, then that gene pair was considered conserved.

In an attempt to identify more distantly related homologs, an additional set of alignments was performed, this time aligning the protein sequences of each species with the translated genomes of the other two using TBLASTN (Altschul et al., 1997). When both genes in the original pair had hits with e-values no greater than 1E-20 within 50 kb of each other, in the same orientation (divergent or convergent) as the original, and with no other genes between them, then the pair was considered conserved.

Gene Ontology Classification

Gene Ontology (GO) classification data was downloaded for all rice, *Arabidopsis*, and *Populus* divergent and convergent genes (rice: TIGR Rice Database, *Arabidopsis*: TAIR GO FTP site: ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology, *Populus*: JGI Poplar Database FTP site: ftp://ftp.jgi-psf.org/pub/JGI_data/Poplar/annotation/v1.1/functional). Rice and *Arabidopsis* genes were classified using the higher-level Plant GOslim vocabulary, while only annotations

using the full GO vocabulary were available for *Populus*. GO class assignments for genes in divergent and convergent pairs were compared to identify pairs in which both genes were in the same class. In order to identify GO classes in which divergently or convergently arranged genes appeared significantly more or less frequently than genes in that species did overall, we compared the number of genes in each group (e.g. rice divergent genes) using the binomial test. The test statistic Z was computed using the following formula:

$$Z = \frac{(F_d - F_G)}{\sqrt{\frac{F_G \times (1 - F_G)}{N_d}}}$$

where F_d is the fraction of divergent or convergent genes in the GO class, F_G is the fraction of all genes in that class, and N_d is the total number of divergent or convergent genes in that species. A GO class was considered significantly over- or under-represented ($P < 0.0001$) when $|Z| > 3.719$.

Regulatory Motif Analysis

Intergenic regions were compiled for all divergent and convergent gene pairs separated by 1 kb or less. These sequences were then scanned for known regulatory elements using the Plant *Cis*-Acting Regulatory DNA Elements (PLACE) database (http://www.dna.affrc.go.jp/PLACE). For each element identified, we calculated the number of sequences in which it appeared. Elements represented in less than 30% of the intergenic regions of divergent and convergent genes were not considered for further analysis. We compared the frequency with which each element appeared in strongly correlated gene pairs with that of pairs showing little or no correlation. The normal approximation of the binomial test (cut-off value of $p < 0.0001$) was used to test for statistically significant differences in frequency of element occurrence between the two data sets.

2.7 SUPPLEMENTAL DATA

The following materials are available at this article's *Plant Physiology* website:

http://www.plantphysiol.org/cgi/content/full/pp.108.122416/DC1

Supplemental Table S1. Coexpressed divergent genes separated by <250bp with Pearson R >0.5.

Supplemental Table S2. Coexpressed convergent genes separated by <250bp with Pearson R >0.5.

Supplemental Table S3. Divergent genes separated by <250bp with conserved gene order and orientation.

Supplemental Table S4. Convergent genes separated by <250bp with conserved gene order and orientation

Supplemental Table S5. Conserved divergent gene pairs with high Pearson correlation R >0.5

Supplemental Table S6. Conserved convergent gene pairs with high Pearson correlation R >0.5

Supplemental Table S7. GO categories significantly under- or over-represented in different gene pair classes

Supplemental Table S8. Number of highly correlated or conserved rice genes in various Gene Ontology classes

Supplemental Table S9. Number of highly correlated or conserved *Arabidopsis* genes in various Gene Ontology classes

Supplemental Table S10. Number of highly correlated or conserved Populus genes in various Gene Ontology classes

Supplemental Table S11. Regulatory elements over-represented in intergenic regions of correlated gene pairs versus non-correlated pairs

2.8 ACKNOWLEDGEMENTS

The authors would like to thank Matthew McCormick for his invaluable assistance in the early stages of this project.

2.9 LITERATURE CITED

Adachi N, Lieber MR (2002) Bidirectional gene organization: A common architectural feature of the human genome. Cell 109: 807-809

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

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815

Brown RL, Kazan K, McGrath KC, Maclean DJ, Manners JM (2003) A role for the GCC-box in jasmonate-mediated activation of the PDF1.2 gene of Arabidopsis. Plant Physiol. 132: 1020-1032

Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. 58: 424-441

Chen N, Stein LD (2006) Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*. Genome Res. 16: 606-617

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. 26: 183–186

Devos KM, Beales J, Nagamura Y, Sasaki T (1999) *Arabidopsis*–rice: Will collinearity allow gene prediction across the eudicot–monocot divide? Genome Res. 9: 825–829

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database:1999. Nucleic Acids Res. 27: 297-300

Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. 5: 299–310

Hwang YS, Karrer EE, Thomas BR, Chen L, Rodriguez RL (1998) Three cis-elements required for rice alpha-amylase Amy3D expression during sugar starvation. Plant Mol. Biol. 36: 331-341

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436: 793-800

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al (2005) Antisense transcription in the mammalian transcriptome. Science 309: 1564–1566

Kruglyak S, Tang H (2000) Regulation of adjacent yeast genes. Trends Genet. 16: 109–111

Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, Myers RM, Weng Z (2007) Transcription factor binding and modified histones in human bidirectional promoters. Genome Res. 17: 818-827

Liu H, Sachidanandam R, Stein L (2001) Comparative genomics between rice and *Arabidopsis* shows scant collinearity in gene order. Genome Res. 11: 2020-2026

Maruyama-Nakashita A, Nakamura Y, Watanabe-Takahashi A, Inoue E, Yamaya T, Takahashi H (2005) Identification of a novel *cis*-acting element conferring sulfur deficiency response in Arabidopsis roots. Plant J. 42: 305-314Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD (2004) Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol. 135: 801-813

Mlynarova L, Loonen A, Mietkiewska E, Jansen RC, Nap J-P (2002) Assembly of two transgenes in an artificial chromatin domain gives highly coordinated expression in tobacco. Genetics 160: 727–740

Ren X-Y, Fiers M, Stiekema WJ, Nap J-P (2005) Local coexpression domains of two to four genes in the genome of Arabidopsis. Plant Physiol. 138: 923-934

Ren X-Y, Stiekema WJ, Nap J-P (2007) Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains. Plant Mol. Biol. 65: 205-217

Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Huizar L, Davis RW, Scherer S, Tait E, Shaw DJ, Harrisi D, Murphyi L, Oliveri K, Taylori K, Rajandreami MA, Barrelli BG, Wolfe KH (2000) Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. 97: 14433-14437

Tatematsu K, Ward S, Leyser O, Kamiya Y, Nambara E (2005) Identification of cis-elements that regulate gene expression during initiation of axillary bud outgrowth in Arabidopsis. Plant Physiol. 138: 757-766

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) An abundance of bidirectional promoters in the human genome. Genome Res. 14: 62-66

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596-1604

Von Gromoff ED, Schroda M, Oster U, Beck CF (2006) Identification of a plastid response element that acts as an enhancer within the Chlamydomonas HSP70A promoter. Nucleic Acids Res. 34: 4767-4779

Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res. 12: 1792-1801

Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GT, Gruissem W, Van de Peer Y, Inze D, De Veylder L (2005) Genome-wide identification of potential plant E2F target genes. Plant Physiol. 139: 316-328

Williams EJG, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. 14: 1060-1067

Wolfe KH, Gouy M, Yang Y-W, Sharp PM, Li W-H (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. 86: 6201–6205

Zhan S, Horrocks J, Lukens LN (2006) Islands of co-expressed neighbouring genes in *Arabidopsis thaliana* suggest higher-order chromosome domains. The Plant J. 45: 347-357

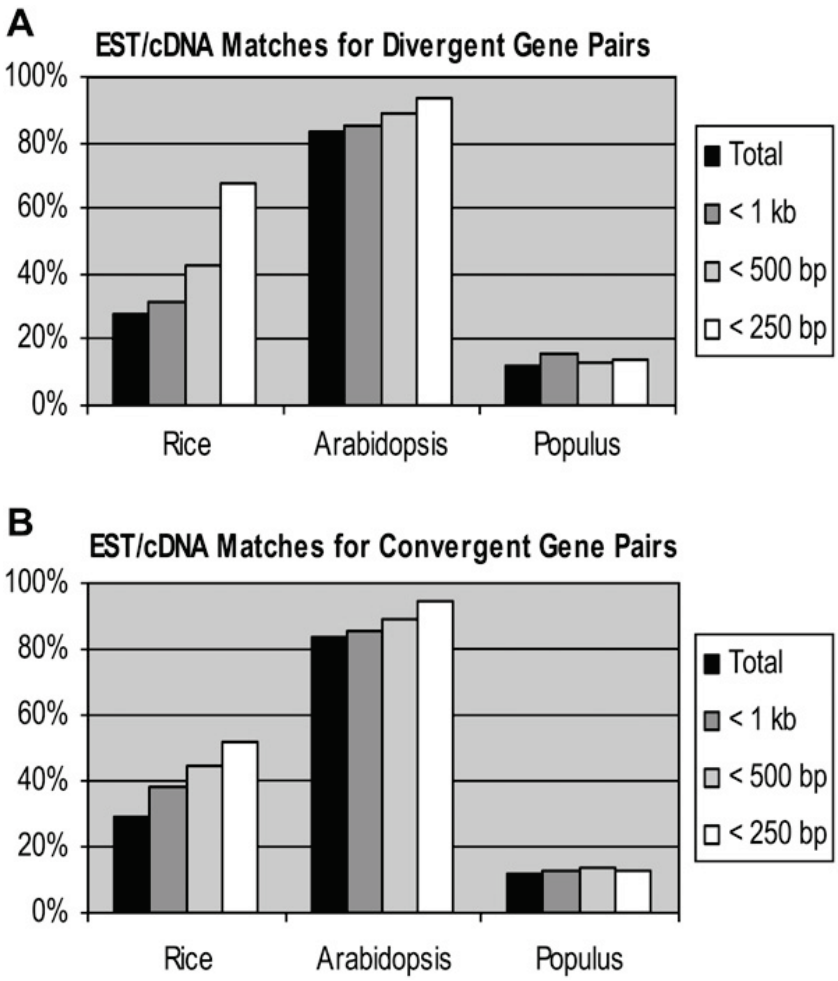


Figure 1 - A. Fractions of divergent gene pairs with matching EST or cDNA sequences for both genes in rice, *Arabidopsis*, and *Populus*. **B.** Fractions of convergent gene pairs with matching EST or cDNA sequences for both genes in rice, *Arabidopsis*, and *Populus*. 'Total' represents the entire population of divergent gene pairs in each species, while '<1 kb,' '<500 bp,' and '<250 bp' each represent a subset of the population with these maximum distances between the genes in a pair.

Table I. *Divergent and convergent gene pairs identified in rice, Arabidopsis, and Populus*

Total represents number of pairs among all genes analyzed. <1 kb represents number of gene pairs with <1,000 bp between transcription start sites (divergent pairs) or transcription stop sites (convergent pairs). <500 bp and <250 bp columns are similar.

	Divergent							Convergent						
	Total	<1 kb		<500 bp		250 bp		Total	<1 kb		<500 bp		<250 bp	
	No.	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Rice	8,742	1,242	14.2	532	6.1	212	2.4	8,772	3,060	34.9	1,626	18.5	734	8.4
Arabidopsis	5,763	2,106	36.5	1,145	19.9	462	8.0	4,949	3,528	71.3	2,650	53.5	1,758	35.5
Populus	8,823	613	6.9	276	3.1	141	1.6	8,967	2,212	24.7	824	9.2	320	3.6

Table II. *MPSS data for divergent and convergent gene pairs*

Columns list numbers and percentages of pairs with both genes showing expression with data from that species' MPSS database. $R > 0.5$ and $R < -0.5$ refer to gene pairs showing strong positive or negative Pearson correlation across all available libraries in the MPSS database. <1 kb, <500 bp, and <250 bp refer to the distance between transcription start (divergent pairs) or stop (convergent pairs) sites of the genes in each pair. *P*-value estimates: $|Z| > 3.0902$: $P < 0.001$, $|Z| > 2.3263$: $P < 0.01$, $|Z| > 1.6449$: $P < 0.05$. Negative *Z*, less than expected. Positive *Z*, more than expected. *Z* values in bold represent significantly different fractions compared to previous category at $P < 0.01$.

	Divergent											Convergent									
	Total		<1 kb			<500 bp			<250 bp			Total		<1 kb			<500 bp			<250 bp		
	No.	*%*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*Z*
Rice																					
Pairs with MPSS data	2,483	28.4	401	32.3	**3.0**	242	45.5	**6.5**	138	65.1	**5.7**	2,568	29.3	1311	42.8	**16.5**	813	50.0	**5.8**	429	58.4	**4.6**
$R > 0.5$	46	1.9	12	3.0	1.7	8	3.3	0.3	3	2.2	−0.7	50	1.9	20	1.5	−1.1	9	1.1	−1.0	5	1.2	0.1
$R < -0.5$	0	0.0	0	0.0		0	0.0		0	0.0		0	0.0	0	0.0		0	0.0		0	0.0	
Arabidopsis																					
Pairs with MPSS data	3,406	59.1	1,292	61.3	2.1	773	67.5	**4.3**	350	75.8	**3.8**	2,880	58.2	2157	61.1	**3.5**	1,736	65.5	**4.6**	1,226	69.7	**3.7**
$R > 0.5$	410	12.0	133	10.3	−1.9	78	10.1	−0.2	30	8.6	−0.9	298	10.3	211	9.8	−0.9	164	9.4	−0.5	117	9.5	0.1
$R < -0.5$	9	0.3	5	0.4	0.9	4	0.5	0.6	3	0.9	0.9	11	0.4	10	0.5	0.6	9	0.5	0.3	7	0.6	0.3

Table III. *Microarray expression data for divergent and convergent gene pairs*

Columns list numbers and percentages of pairs with both genes showing expression with our compiled microarray databases. $R > 0.5$ and $R < -0.5$ refer to gene pairs showing strong positive or negative Pearson correlation across all hybridizations. <1 kb, <500 bp, and <250 bp refer to the distance between transcription start (divergent pairs) or stop (convergent pairs) sites of the genes in each pair. *Z* values in bold represent significantly different fractions compared to previous category at $P < 0.01$.

	Divergent											Convergent									
	Total		<1kb			<500 bp			<250 bp			Total		<1kb			<500 bp			<250 bp		
	No.	*%*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*Z*
Rice																					
Pairs with microarray data	4,450	50.9	656	52.8	1.3	334	62.8	**4.6**	171	80.7	**5.4**	4,959	56.5	2,119	69.2	**14.2**	1,230	75.6	**5.6**	580	79.0	2.1
$R > 0.5$	1,148	25.8	164	25.0	−0.5	90	26.9	0.8	51	29.8	0.8	1,283	25.9	574	27.1	1.3	336	27.3	0.2	155	26.7	−0.3
$R < -0.5$	0	0.0	0	0.0		0	0.0		0	0.0		0	0.0	0	0.0		0	0.0		0	0.0	
Arabidopsis																					
Pairs with microarray data	5,111	88.7	1,883	89.4	1.1	1,038	90.7	1.4	421	91.1	0.3	4,295	86.8	3,075	87.2	0.7	2,341	88.3	1.8	1,576	89.6	1.7
$R > 0.5$	522	10.2	200	10.6	0.6	125	12.0	1.5	53	12.6	0.3	366	8.5	260	8.5	−0.1	205	8.8	0.5	143	9.1	0.4
$R < -0.5$	84	1.6	32	1.7	0.2	19	1.8	0.3	8	1.9	0.1	65	1.5	51	1.7	0.7	44	1.9	0.8	31	2.0	0.3
Populus																					
Pairs with microarray data	8,313	94.2	561	91.5	**−2.9**	246	89.1	−1.4	126	89.4	0.1	8,406	93.7	2,075	93.8	0.1	774	93.9	0.1	310	96.9	2.2
$R > 0.5$	4,036	48.6	295	52.6	1.9	145	58.9	2.0	76	60.3	0.3	4,063	48.3	1,006	48.5	0.1	381	49.2	0.4	161	51.9	1.0
$R < -0.5$	126	1.5	20	3.6	**4.0**	10	4.1	0.4	7	5.6	0.8	122	1.5	22	1.1	−1.5	6	0.8	−0.8	1	0.3	−0.9

Table IV. *Divergent and convergent gene pairs conserved in other species*

Gene pair conservation was determined using a combination of BLASTP and TBLASTN searches, aligning the protein and genomic sequences of divergent and convergent genes with all predicted protein sequences (BLASTP) or the entire genome of the other species. See "Materials and Methods" section for the criteria used to determine gene pair conservation.

	Divergent								Convergent							
	Total		<1 kb		<500 bp		<250 bp		Total		<1 kb		<500 bp		<250 bp	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Rice																
Conserved in Arabidopsis	26	0.3	4	0.3	2	0.4	1	0.5	42	0.5	23	0.8	13	0.8	6	0.8
Conserved in *Populus*	77	0.9	13	1.0	5	0.9	2	0.9	111	1.3	58	1.9	32	2.0	19	2.6
Arabidopsis																
Conserved in rice	29	0.5	5	0.2	1	0.1	1	0.2	52	1.1	42	1.2	34	1.3	23	1.3
Conserved in *Populus*	355	6.2	125	5.9	78	6.8	24	5.2	401	8.1	312	8.8	247	9.3	178	10.1
Populus																
Conserved in rice	58	0.7	7	1.1	4	1.4	1	0.7	114	1.3	39	1.8	14	1.7	7	2.2
Conserved in Arabidopsis	267	3.0	17	2.8	8	2.9	5	3.5	421	4.7	116	5.2	35	4.2	14	4.4

Table V. *Effect of shared GO terms and high correlation on gene pair conservation*

The likelihood of a divergent or convergent pair being conserved is increased significantly when the genes either share at least one GO classification or have strongly correlated expression levels (Pearson $R > 0.5$). Z values in bold are significantly different at $P < 0.01$.

	Divergent								Convergent							
	All Pairs		Shared GO			$\|R\| > 0.5$			All Pairs		Shared GO			$\|R\| > 0.5$		
	No.	*%*	*No.*	*%*	*Z*	*No.*	*%*	*Z*	*No.*	*%*	*No.*	*%*	*Z*	*No.*	*%*	*Z*
Rice																
Conserved in Arabidopsis	26	0.3	10	1.7	**6.3**	11	0.9	**4.0**	42	0.5	15	1.9	**5.9**	15	1.1	**3.4**
Conserved in *Populus*	77	0.9	31	5.3	**11.5**	28	2.4	**5.5**	111	1.3	25	3.2	**4.9**	26	2.0	2.3
Arabidopsis																
Conserved in rice	29	0.5	14	0.6	0.3	4	0.7	0.7	52	1.1	34	1.5	2.1	5	1.6	0.6
Conserved in *Populus*	355	6.2	176	6.9	1.5	45	8.1	1.9	401	8.1	171	7.5	−1.0	37	10.3	1.4
Populus																
Conserved in rice	58	0.7	6	2.2	**3.1**	32	0.8	0.9	114	1.3	7	3.0	2.3	51	1.2	−0.3
Conserved in Arabidopsis	267	3.0	10	3.6	0.6	131	3.2	0.5	421	4.7	24	10.1	**4.0**	213	5.3	1.2

Table VI. *Regulatory elements overrepresented in intergenic regions of correlated pairs*

Number of regulatory elements overrepresented among strongly correlated gene pairs. The composition of the intergenic regions of divergent gene pairs differs between those pairs that display strongly correlated expression levels and those that do not. Statistically significant variation was determined using the binomial test (normal approximation) with a cutoff value of $P < 0.0001$.

Divergent	No. of Elements	Overrepresented	Convergent	No. of Elements	Overrepresented
Arabidopsis	286	39	Arabidopsis	248	0
Rice	283	16	Rice	282	0
Populus	262	1	*Populus*	288	0

CHAPTER 3:

CONSERVATION, REARRANGEMENT, AND DELETION OF GENE PAIRS IN FOUR GRASS GENOMES

Nicholas Krom and Wusirika Ramakrishna

This work was supported by the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service, grant number 2007-35301-18036.

3.1 ABSTRACT

Gene order and content differ among homologous regions of closely related genomes. In addition, similarities in the expression profiles of some physically adjacent genes suggest that the proper functioning of these genes depends on maintaining a specific position relative to each other. In order to better understand the results of the interaction of these two genomic forces, we identified convergent, divergent, and tandem gene pairs in rice and sorghum, as well as their homologs in rice, sorghum, maize, and *Brachypodium*. Using this data, we determined the status of each pair in all four species: whether it was conserved, inverted, rearranged, or missing homologs. Several interesting trends were noted, including considerably lower rates of conservation among divergent gene pairs than convergent or tandem pairs, substantially higher rates of rearranged pairs and missing homologs in maize than in any other species, and evidence for the creation of significantly more genes in the ancestors of rice than in sorghum since the divergence of the two species 50-70 million years ago. In rice, gene pairs with strongly correlated expression levels were found to be conserved significantly more often than pairs with little or no correlation. By analyzing the status of each pair in all four species, we were able to assign each pair to one of fourteen putative evolutionary histories, leading to several significant observations regarding differences in the evolutionary dynamics of the three types of gene pair, as well as between the lineages of rice and sorghum.

3.2 INTRODUCTION

One of the primary areas of investigation in comparative genomics is the identification and characterization of homologous regions in closely related genomes. The subjects of these investigations range in scale from multi-megabase syntenic regions covering most of a chromosome to small loci containing just a few genes. Studying syntenic regions can uncover large scale events in the evolutionary history of a genome, such as segmental duplications or polyploidization; however, these regions can still contain many significant differences between species due to the large number of genomic alterations that can take place over time while maintaining sufficient collinearity to define regions of synteny. Comparative analysis of small loci can produce detailed evolutionary histories of groups of neighboring genes and provide examples of the types of changes possible in a genome, but such studies are difficult to expand to a genome-wide scale due to the number of genes involved, the problem of generalizing types of changes to allow their quantification, and the subsequent difficulty of interpreting the results of such an analysis.

In this study, we conduct an intermediate form of comparative analysis. By examining pairs of adjacent genes, we are able to detect changes at the level of single genes, while maintaining the ability to observe relationships between genes. Due to the simplicity and small scale of our subjects, it is possible to assign all possible changes to a manageable number of classes, and therefore the results of a genome-wide study of this type can be easily interpreted. In contrast to our previous investigation of gene pairs (Krom and Ramakrishna, 2008), in which we compared three plant species (rice, *Arabidopsis*, and *Populus trichocarpa*) which diverged 130 to 200 million years ago (mya) (Wolfe et al., 1989; Chaw et al., 2004; Tuskan et al., 2006), the analysis presented here compares four members of the Poaceae family (rice, sorghum, maize, and *Brachypodium*) whose last common ancestor dates to 50-70 mya (Wolfe et al., 1989; Buell, 2009). The shorter evolutionary distances separating these species simplifies the interpretation of any observed genomic rearrangements, due to the reduced probability of multiple independent events affecting the same region. However, many small rearrangements have been identified in earlier comparative studies of Poaceae genomes (Ilic et al., 2003; Bennetzen and Ramakrishna, 2004), providing sufficient variation among genomes to identify any trends regarding selection for or against disruption of ancestral gene pairs. It has been hypothesized that gene order is not entirely random, but rather is connected to gene function and regulation (Hurst et al., 2004), and that genomic rearrangements can alter the function of genes or even lead to the creation of new gene families, and may therefore contribute to phenotypic differences between species, even when the individual genes are conserved (Ciccarelli, 2005).

We have previously observed (Krom and Ramakrishna, 2008) that their strand-wise arrangement has a significant influence on many characteristics of gene pairs. For this study, we have classified all pairs of adjacent genes as either convergent (→←), divergent (← →), or tandem (→ → or ← ←), identified homologous genes in other species, and determined the status of each pair (conserved, inverted, moved, and missing homologs). The effect of correlated expression on these types of gene pair rearrangements was also estimated. To gain an understanding of the evolutionary timing of the rearrangements we observed, a putative evolutionary history was created for each gene pair, based on its status in each of the four species. Overall, this study provides an overview of the frequencies and types of genomic rearrangements within a subset of the Poaceae, as well as many other properties of the genomes being studied.

3.3 RESULTS

Conservation and Rearrangements

Convergent, divergent, and tandem gene pairs in the rice and sorghum genomes were identified. Release 6 of the rice pseudomolecule annotation contained 57,840 genes in all, 40,821 of which were not annotated as either hypothetical or transposon-related. Among these genes, 4,800 convergent pairs, 3,711 divergent pairs, and 9,428 tandem pairs were identified. In sorghum, 32,245 out of 34,008 genes were analyzed, yielding 5,059 convergent, 4,913 divergent, and 11,847 tandem pairs.

The primary goal of this analysis was to determine how frequently the exact arrangement of a pair of adjacent genes is conserved in the genomes of other grass species, and what changes have taken place when the pair is not conserved. Out of rice, maize, sorghum, and *Brachypodium*, rice and sorghum were chosen as the starting points for these comparisons due to their sequence and annotation data sets being considerably better than those of maize and *Brachypodium*. Our comparative sequence analysis placed each rice or sorghum gene pair into one of six categories based on the presence or absence of homologous genes and their locations in the genome. A pair of genes adjacent in rice/sorghum may be conserved, with or without additional genes inserted between them. If a pair's homologs were found to still be adjacent or separated by a small number of insertions with a different strand-wise arrangement, then the pair was designated "inverted". Homologs falling on different contigs or separated by excessive distance were considered "moved". Finally, one or both genes in the original pair may be lacking homologs, either having been deleted in that species' lineage or having arisen in the ancestors of rice or sorghum since divergence from their last common ancestor. Together, these categories include all the major events of genomic evolution at this scale, and the relative frequencies of these events provide insight into the importance of proximity and strand-wise arrangement to proper gene function and regulation.

Rice gene pairs displayed very similar patterns of conservation and rearrangement in sorghum and *Brachypodium* (Table 1). In both species, divergent pairs were conserved least frequently (51.4% in sorghum, 50.0% in *Brachypodium*), while tandem pairs were conserved most often (59.8% in sorghum, 57.9% in *Brachypodium*). The fraction of pairs conserved exactly in sorghum, with no other genes inserted between the homologs, ranged from 29.6% of divergent pairs to 42.4% of convergent pairs. A similar pattern was seen in *Brachypodium*, with both divergent and tandem pairs conserved without insertions making up approximately two-thirds of all conserved pairs, compared to 77% of convergent pairs. Conservation of gene pair arrangement was found to be substantially less common in maize, with roughly one-third of convergent and tandem pairs conserved, while only 17.7% of divergent pairs were conserved. Among

those pairs that are conserved in maize, gene pair arrangement was conserved exactly in 11.9% of divergent pairs, 22.6% of tandem pairs, and 28.5% of convergent pairs. In all three species, divergent pairs were conserved significantly less often than convergent or tandem pairs, both of which were conserved at roughly equal rates. Conservation without insertions was consistently more common among convergent pairs than the other two types.

Inversion of one or both genes was found to be quite a rare phenomenon, ranging from 1.8% of rice tandem pairs inverted in sorghum to 5.2% of divergent pairs in maize. Tandem pairs were inverted least often in all three species, while convergent pairs were inverted most often in sorghum and *Brachypodium*, and divergent pairs were inverted most frequently in maize. A majority of inverted pairs were found to have additional genes inserted within them, with only 31-42% of paired genes remaining immediately adjacent after inversion. No clear connection between pair type and frequency of insertion was observed.

Approximately one-quarter of rice gene pairs were found to be separated by more than 50 kbp or located on different contigs in both sorghum and *Brachypodium*. While the proportion of such pairs was nearly identical in these two species, they were nearly twice as common in maize, where approximately 40% of convergent and tandem pairs displayed significantly different relative locations, as did over 54% of divergent pairs. In all three species, tandem pairs were the type least frequently rearranged in this manner, while divergent pairs were the most common.

Homologs for one or both genes were not found for a significant fraction of rice gene pairs. Sorghum and *Brachypodium* gave similar results, with 14-17% of rice pairs missing homologs. This frequency was somewhat higher in maize, ranging from 20.7% of convergent pairs to 23.7% of tandem pairs lacking homologs for one or both genes. Tandem pairs were consistently more likely to be missing homologs for both genes than convergent or divergent pairs, while divergent pairs were most likely to be missing single homologs.

Many of the trends in conservation observed in rice were also observed among sorghum gene pairs (Table 2). The fractions of sorghum gene pairs conserved (both with and without insertions) were nearly identical to those of rice pairs conserved in sorghum, as would be expected. The small differences observed are most likely due to variation in copy number of duplicated pairs between the two species, i.e. multiple rice pairs mapping to a single sorghum pair, and vice versa. Conservation rates for sorghum pairs in *Brachypodium* were consistently lower than those for rice pairs, but other trends noted in rice with regard to pair type and frequency of insertions were also observed in sorghum. Sorghum pairs of all types were conserved considerably more frequently in maize than

were rice pairs, with rates ranging from 24% of divergent pairs to 46% of convergent pairs. These data suggest that divergent pairs were far less likely to be conserved in maize than any other type.

Sorghum pairs were inverted in rice and *Brachypodium* with nearly the same frequencies as were rice pairs in sorghum and *Brachypodium*, while inversions in maize were slightly more common for all pair types. Inverted pairs without gene insertions varied only slightly across species and pair types, ranging generally from 1-2%.

When examining the results of this analysis, the primary area of variation between the sorghum-based and rice-based data was in the number of pairs whose homologs were either physically distant or missing. The fractions of sorghum pairs with physically distant homologs in rice were consistently higher, by about 10 percentage points, than the corresponding quantities of rice pairs' homologs in sorghum. A similar trend was noted in *Brachypodium*, where the homologs of sorghum pairs were distant 31-37% of the time, compared to 22-29% of rice pairs. In maize, the differences were much smaller but nonetheless continued the pattern.

The larger numbers of physically distant homologs among sorghum gene pairs compared to those of rice were accompanied by smaller numbers of pairs lacking homologs for one or both genes. The largest difference in the number of missing homologs was observed in maize, where less than half as many sorghum pairs lacked homologs as did rice pairs. Similarly, the numbers of sorghum pairs lacking homologs in rice were just over half of those of rice pairs in sorghum. While fewer sorghum pairs were missing homologs in *Brachypodium* than were rice pairs, the difference was far less pronounced, comprising between one and two percentage points.

Effects of Correlated Expression on Rearrangements

Using microarray and Massively Parallel Signature Sequencing (MPSS) expression data, the Pearson correlation coefficient of all rice genes pairs was calculated. Those pairs with coefficients of 0.5 or greater were considered significantly correlated as described earlier (Krom and Ramakrishna, 2008). The full set of rice gene pairs was then divided into correlated and uncorrelated sets, and difference in the frequencies of each type of rearrangement within these sets was tested for significance using the binomial test. The end result of this test was to determine whether gene pairs with correlated expression levels were subject to any of the various types of rearrangements at a significantly different rate than uncorrelated pairs.

With the single exception of rice convergent pairs in *Brachypodium*, all types of correlated gene pairs were significantly over-represented among conserved pairs (Table 3). The difference in conservation rates between correlated and uncorrelated pairs was highest for tandem pairs, followed closely by divergent pairs, while the effect of correlation on convergent pairs was considerably weaker.

When examining the effect of correlated expression on gene pair inversion, the effectiveness of the binomial test was reduced by the small sample sizes involved. Correlated convergent pairs are consistently over-represented among inverted pairs, although the difference is not always statistically significant. Correlated divergent pairs are weakly over-represented in sorghum, while in maize and *Brachypodium* they tend to be under-represented. In sorghum and maize, correlated tandem pairs were under-represented among inverted pairs, while in *Brachypodium* they were weakly over-represented. The general trends observed among correlated pairs regarding inversion tend to follow those seen among all pairs, with convergent pairs being inverted most frequently and tandem pairs least frequently, so it may be that correlated expression has little effect on gene inversion. However, due to the very small number of correlated inverted pairs observed, no definitive conclusion can be drawn.

In contrast, correlated expression appeared to select against disruption of a gene pair's physical arrangement, with divergent and tandem pairs being significantly under-represented in all three comparison species among pairs whose homologs are physically distant. Correlated convergent pairs displayed the same tendency in *Brachypodium*, while in maize they were weakly over-represented and showed no significant difference in sorghum.

No clear pattern emerged with regard to the effect of correlated expression on the absence of homologous genes. While convergent pairs were strongly under-represented among pairs lacking homologs in sorghum and maize, their under-representation in *Brachypodium* was quite weak. Correlated divergent pairs displayed little difference from their non-correlated brethren in maize and *Brachypodium*, yet were strongly under-represented in sorghum. In contrast, correlated tandem pairs showed no significant difference in sorghum and *Brachypodium*, and were actually over-represented among pairs lacking homologs in maize.

Evolutionary History of Gene Pairs

The status of each rice or sorghum gene pair in its three comparison species was examined in order to estimate its evolutionary history. For the purpose of this analysis, a pair could be in one of three states in each species: conserved (without insertions),

rearranged (physically distant homologs, any inversion, and "conserved" with insertions), or deleted (one or both homologs nonexistent). Based on the possible combinations of these states, fourteen categories of evolutionary history were devised. The putative evolutionary tree for the four species consisted of two branches, one made up of rice and *Brachypodium*, the other sorghum and maize. Similarities within branches, as well as differences between them, served as the basis for many of the fourteen categories.

The first category consisted of those pairs whose exact arrangement was shared in all four species (Table 4A, Figure 1A). Results varied little between the rice-based and sorghum-based analyses. Convergent pairs were the most common in this class, with over 18% of pairs falling into this category, followed by tandem pairs (~10%), and divergent pairs (~5%).

Pairs conserved in two of their three comparison species most likely underwent a single species-specific rearrangement or deletion (Table 4B, Figure 1B-C). The former event was by far the most common, comprising 16-19% of all pairs, compared to the ~1% or less of pairs with one or both homologs deleted in a single species. Rice and sorghum results differed by less than one percentage point across all pair types.

The six categories comprised of pairs conserved in only one other species were divided into two groups (Table 4C), those in which the pair was conserved within one branch of the evolutionary tree (i.e. a rice pair conserved in *Brachypodium*), referred to here as a "branch specific" pair (Figure 1D-F), and those in which the pair was conserved in one species in each branch, a state referred to as "cross-branch conservation" (Figure 1G-I). Among both branch-specific and cross-branch conserved pairs, it was far more common (9-12% of pairs in rice, 4-8% in sorghum) for the pair to be rearranged in the other two species than for it to be deleted in one (0.4-1.5%) or both species (0.1-1.5%). In rice, branch-specific pairs were slightly more common than cross-branch conserved pairs, while in sorghum the opposite was true. There were only two sets of genes in which rice and sorghum differed substantially. The first was branch-specific divergent pairs with the pair being rearranged in the other two species, which included 12% of rice divergent pairs but only 4.5% of sorghum pairs. The opposite situation was observed among cross-branch conserved divergent pairs, again with two rearrangements. These pairs made up 9.6% of rice divergent pairs, compared to 15% of sorghum pairs. Additionally, in sorghum cross-branch conserved pairs in general were much more common (39.6%) than branch-specific pairs (23%). In rice, branch-specific pairs were the more common type, but by a much smaller margin (34.6% vs. 31.9%). Overall, the number of pairs within each of these six categories appeared to be inversely proportional to the number of deletions, with pairs either conserved or rearranged in all species the most common, followed by those deleted in one species, and with pairs missing homologs in both species in which they are not conserved being least common.

However, the difference in frequency between the first type and the second is much greater than between the second and third.

The last five categories consist of those pairs that exist in only one species (Table 4D). Pairs whose genes exist in all four species, but whose pair-wise arrangement is found in a single species (Figure 1K), were the most common type in both species, although they were considerably more common in sorghum (28-38%) than in rice (22-30%). Divergent pairs fell into this category substantially more often than convergent or tandem pairs in both rice and sorghum. In rice, the second most common category is those pairs containing one or more genes unique to that species (Figure 1J), with 10-12% of all pairs, while in sorghum this category is approximately one-third as large, containing only 3.5-4.1% of all sorghum pairs. The differences between these first two categories were roughly complementary, with the sum of their frequencies being nearly equal for all three pair types. The distribution of pairs among the remaining three categories (Figure 1L-N) showed little variation between rice and sorghum or between pair types.

3.4 DISCUSSION

As the number of sequenced plant genomes increases, so does the range of potential discoveries by genome-wide comparative studies. With the genomes of several closely related grass species being available, it is now possible to classify and quantify the small scale genomic differences that arise across relatively small evolutionary distances. In this study, we examined three types of gene pairs in rice and sorghum, identified their homologs in maize, *Brachypodium*, sorghum, and rice, and studied their conservation, rearrangement and deletion. In addition, we studied the effect of correlated expression on gene pair conservation and rearrangement and produced a potential evolutionary history for each pair based on the status of its homologs among the other genomes being investigated.

Our study began with the identification of all gene pairs that did not contain hypothetical or transposon-related genes. Substantially more gene pairs meeting these criteria were identified in sorghum than in rice, despite a smaller pool of annotated genes. The sorghum genome is believed to contain a significantly larger proportion of transposons than the rice genome (Paterson et al, 2009), and one would therefore expect to find fewer acceptable gene pairs in sorghum than in rice, assuming similar distributions of transposons in both species. This discrepancy is caused by errors in the annotation of sorghum and / or rice genome. The first possibility is that many transposons in the sorghum genome are not annotated, and therefore many of the pairs of genes analyzed here are in fact separated by one or more transposons, rather than being directly adjacent as we had assumed. Another possibility is that many sorghum transposons are annotated inaccurately, assigned some other predicted function and with substantial amounts of non-transposon sequence included in the predicted transcript. In this situation both of our methods for excluding transposon-related genes from the analysis (filtering based on predicted function and transposon sequence content≥50% as determined by RepeatMasker) would fail to identify these transposons. The effect of these errors on our analyses would depend on both the effect of transposons within pairs of non-transposon genes on the pair's conservation and rearrangement, and the differences in how gene pairs including transposons tend to change on an evolutionary timescale.

Several factors were noted which appeared to influence the frequency with which gene pairs were conserved or subject to certain types of rearrangements. Divergent gene pairs were conserved least often among the three types in every comparison. While in comparisons between rice, sorghum, and *Brachypodium* divergent pairs were conserved 10-20% less often than convergent or tandem pairs, in maize this disparity grew to 50% and greater, especially when examining pairs conserved without insertions. This reduced frequency of conservation among divergent pairs is mirrored by an increase in the

frequency of physically distant homologs. Again, the degree to which divergent pairs diverge from the other pair types is considerably greater in maize. Divergent gene pairs also consistently lead the pack with regard to pairs missing a single homolog, albeit the magnitude of this difference is not nearly as large as in the previous two cases. Together, these observations suggest that divergent gene pairs are significantly more likely to be disrupted by the insertion of genes or transposons within them or by the relocation of one or both genes.

The dissimilarity in conservation and rearrangement rates between maize and the other comparison species most likely stems from three primary sources. First, the genomic sequence of maize used in this study is in the form of individual BAC sequences, most of which are less than ~250 kbp in length, rather than assembled sequences of near-chromosome length. Smaller sequences are, of course, less likely to contain complete gene pairs, especially if their intergenic regions have accumulated other genes or transposons over time. Second, transposons make up a much larger fraction of the maize genome than that of rice, sorghum, or *Brachypodium*. In addition to physically disrupting the region into which they insert themselves, transposons may also increase the likelihood of recombination, deletions, and other alterations in any area they inhabit. Third, the ancestors of maize quite likely suffered large scale gene loss (Lai et al., 2004). If the first gene in a pair is deleted from one copy and the second gene was deleted in the other copy of the pair, both genes would still exist in the genome, but would no longer be paired. This type of occurrence would explain why more rice and sorghum gene pairs were found to have physically distant homologs in maize than in any other comparison species. All of these factors would result in a general reduction in the frequency of gene pair conservation and the corresponding increase in rearranged pairs, as was observed. However, they do not explain the inordinately low conservation rates of divergent gene pairs observed in maize. Further investigation will be required to fully understand the cause of those observations.

This study began by identifying gene pairs in rice and sorghum, and both sets of gene pairs were used as the query sequences for comparisons with the other three genomes included in this study. Comparing the results of these two comparisons unveils a number of differences between the two species, some rather obvious, others less so. The conservation and inversion rates of rice pairs in sorghum and sorghum pairs in rice were nearly identical, as would be expected. However, the two species differed considerably in the number of pairs with homologs that were either physically distant or missing. More rice pairs were missing homologs than were sorghum pairs, while more sorghum pairs had physically distant homologs than did rice pairs. These results suggest the number of rice genes without homologs in sorghum is greater than the number of sorghum genes without homologs in sorghum. That is to say, more new genes have

arisen in the lineage of rice since its divergence from the ancestors of sorghum than have been created in the lineage of sorghum since that time. This situation would explain both the higher proportion of rice pairs missing homologs (due to the larger number of genes specific to rice and its close relatives) and the higher proportion of sorghum pairs with homologs that are present in rice but not as a pair of genes that are adjacent or in close proximity (due to the sorghum genome containing a higher proportion than rice of genes shared between the two species). Other investigations into shared and species-specific genes have identified large numbers of genes found only in rice or sorghum; however, no clear conclusion can be reached with regard to which species contains more unique genes. Campbell and colleagues (2007) identified 7427 rice genes not shared by any other species within the Poaceae, including sorghum, but did not investigate unique genes within the sorghum genome. Our analysis identified only approximately 4000 such genes in rice, although our exclusion of hypothetical genes (which number 11,721 in all) is a likely contributor to this difference. An investigation into gene families shared by rice and sorghum (Paterson et al., 2009) identified 2032 sorghum gene families not shared by rice, compared to 802 rice gene families not found in sorghum. However, considerably fewer sorghum genes were used in this analysis than in ours (~28,000 vs. ~34,000), and the number of rice genes analyzed in their study was far smaller than ours (~20,000 vs. ~40,000). These variations in data set size and methods make direct comparison of results difficult, although it can be agreed that considerable numbers of species-specific genes exist, and their numbers may vary greatly among even fairly closely related species such as rice and sorghum.

Another way in which rice and sorghum behaved differently can be seen in their conservation/rearrangement patterns in maize and *Brachypodium*. *Brachypodium* is more closely related to rice than to sorghum, and accordingly rice gene pairs are more frequently conserved in *Brachypodium* than are sorghum pairs. Likewise, maize is more closely related to sorghum than to rice, and thus sorghum pairs have higher rates of conservation in maize than do rice pairs. Additionally, larger numbers of rice pairs are missing homologs in maize than either sorghum in maize or sorghum in *Brachypodium*.

Rice gene pairs displaying strongly correlated expression levels were found to be significantly more likely to be conserved in sorghum, maize, and *Brachypodium*. Strongly correlated pairs of all types were significantly over-represented among conserved pairs when compared with non-correlated pairs, and correlated divergent and tandem pairs were under-represented among pairs with physically distant homologs. Correlated expression levels have also been found to increase the likelihood of conservation among fungi (Kensche et al., 2008). The largest increases in the frequency of conservation as a result of correlated expression was observed among divergent and tandem pairs, a pattern that has been observed before in a comparison of human, mouse,

and chicken gene clusters (Semon and Duret, 2006). These results lend further support to the hypothesis that the strand-wise arrangement of pairs of adjacent genes may be an essential part of the regulatory schemes of some strongly correlated gene pairs, such that rearrangements disturbing the pair would be selected against. Similar trends have been noted before in mammals (Singer et al., 2005; Semon and Duret, 2006). Therefore, these appear to be universal phenomena in eukaryotes including plant genomes.

In examining the results of the evolutionary history analysis, several informative differences were noted between the three types of gene pairs and between rice and sorghum. Divergent gene pairs were the type least likely to be conserved, with a majority (54% in rice, 57% in sorghum) being found in only one species, and approximately one-third and one-half as likely as convergent and tandem pairs, respectively, to be found in all four species examined. The difference in conservation frequency between convergent and tandem pairs was not nearly as great as between divergent pairs and either of the other types, suggesting that divergent pairs are somehow targeted for insertions, deletions, or other rearrangements.

The distribution of gene pairs among the categories of evolutionary history was, for the most part, similar between rice and sorghum. One area in which they differ greatly is in the fraction of pairs containing genes specific to that species (i.e. missing homologs in all other species). Such pairs are nearly three times as common in rice as in sorghum, further supporting the hypothesis, mentioned above, that considerably more new genes arose in the ancestors of rice than in those of sorghum since the divergence of the two lineages. This could also be due annotated genes in the rice genome which are not "real" genes. The proportion of pairs containing genes found in all species but only paired in sorghum is considerably higher than the analogous group in rice, a situation that would also result from a higher number of genes unique to rice than to sorghum.

Another interesting observation comparing rice and sorghum can be made when examining those gene pairs conserved in only one other species. In rice, these pairs are more likely to be conserved in *Brachypodium* than in sorghum or maize (except for tandem pairs, which are slightly less likely), as would be expected due to the shorter evolutionary distance between rice and *Brachypodium*. However, for sorghum gene pairs the situation is reversed, with all three pair types more frequently conserved in either rice or *Brachypodium* than in maize. While part of this discrepancy can be attributed to the unassembled state of the maize genome sequence used, the other probable cause is maize's far higher transposon content. Insertion of transposons within pairs of genes would disrupt the pair, thus a higher number of transposon insertions would lead to a lower number of conserved pairs. Another important factor that would contribute to this observation would be gene loss after whole genome duplication of the maize genome. This would be accomplished by the loss of one gene from the original gene pair and the

second gene from the duplicated copy of the gene pair. Additionally, the difference between the fraction of pairs conserved in maize versus those conserved in rice or *Brachypodium* is by far greatest for divergent pairs, which are over three times more likely to be conserved in the latter two species than in maize when they are conserved in only one species. Tandem pairs exhibit a similar tendency, although the difference in conservation frequency is somewhat less, while among convergent pairs the difference is a mere half a percentage point. These results further support the notion that divergent pairs are considerably more likely to be rearranged over time, while convergent pairs appear somewhat less susceptible to such changes.

3.5 CONCLUSIONS

We identified convergent, divergent, and tandem gene pairs in rice and sorghum, and determined the status of their homologs in rice, sorghum, maize, and *Brachypodium*. Significant differences in the frequency of the different types of genomic rearrangements were observed among the three types of gene pair, as well as between rice and sorghum, the two genomes used as starting points for our comparisons. We found evidence for the creation of significantly more new genes in the ancestors of rice than in those of sorghum since the divergence of the two species. Correlated expression was found to increase the frequency of gene pair conservation in rice. Based on the evolutionary organization of gene pairs in grass genomes, ancestral rearrangement patterns were inferred. Overall, our study provides information on conservation and rearrangement of gene pairs during the evolution of the grasses serving as basis for further investigations on functional interactions between adjacent genes.

3.6 METHODS

Identification of Gene Pairs

Genome sequence and annotation data were downloaded for rice (*Oryza sativa* subsp. *japonica*) (http://rice.plantbiology.msu.edu, rice pseudomolecules release 6), sorghum (*Sorghum bicolor*) (http://www.phytozome.net/sorghum, sequence assembly v1.0, gene set v1.4), maize (*Zea mays*) (http://www.maizesequence.org, release 3a.50), and *Brachypodium* (*Brachypodium distachyon*) (http://www.brachypodium.org, 4X coverage release). Annotated genes in the rice and sorghum genomes were sorted by chromosome and position and then, based on which strand the gene is transcribed from, all pairs of adjacent genes were classified as either convergent (→ ←), divergent (← →), or tandem (→ → or ← ←) pairs. Pairs containing transposon related genes, as determined by annotation and RepeatMasker (www.repeatmasker.org) (50% or greater transposon content of unspliced sequence), were excluded from all analyses, and pairs containing hypothetical genes were excluded from the main analysis.

Comparative Sequence Analysis

The coding region sequences of all rice and sorghum gene pairs were aligned with the genome assemblies of the other three species using BLASTN. For each gene, individual hits (presumably corresponding to single exons) with e-values of 1E-10 or less were grouped with other nearby hits on the same strand and contig to produce a putative homologous gene region. The locations of each pair's homologs were then used to determine the pair's status as conserved, inverted, or moved. Pairs were considered "conserved" if both genes had homologs in the original strand-wise arrangement (convergent, divergent, or tandem) within 50 kbp of each other. "Inverted" pairs also possessed homologs within the cutoff distance, but their strand-wise arrangement had been altered. Pairs were considered "rearranged" if homologs of both genes were identified but were either too far apart to be considered conserved or inverted or located on different contigs. An additional analysis was performed on the regions between the homologs of conserved and inverted pairs to identify those pairs within which other genes had been inserted. Those pairs in which one or both genes lacked a homolog in a given species were also identified.

Expression Analysis

Two types of quantitative expression data were collected for all rice genes: microarray and Massively Parallel Signature Sequencing (MPSS). MPSS (Meyers et al., 2004) data were downloaded from the Rice MPSS Database (http://mpss.udel.edu/rice/). Only 17-bp signatures of classes 1, 2, 5, and 7 that mapped to a single gene were used, and abundance values less than 5 were ignored as background interference. When multiple signatures had significant abundance values in the same library, their average abundance was used. Correlated expression between genes in convergent, divergent, and tandem pairs was examined by calculating the Pearson correlation coefficient using each gene's average abundance values across 72 libraries.

Microarray data was downloaded from the Yale rice project (http://bioinformatics.med.yale.edu/rc/overview.jspx). Expression data were collected for a total of 446 hybridizations. Correlated expression was again tested with the Pearson correlation coefficient, this time pairing data points from the same hybridization and channel.

Evolutionary Analysis of Gene Pairs

A maximum likelihood estimate of the evolutionary history of each gene pair was created by comparing the status of the pair in each of the four species in this study. The likelihood of a given scenario was based upon the number of gene rearrangements, deletions, and conservation to arrive at the present state starting from a common ancestor. Gene pairs were then assigned to one of fourteen groups based on their putative histories.

3.7 LITERATURE CITED

Bennetzen, J. L. and W. Ramakrishna. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol 48: 821-827.

Buell, C. R. 2009. Poaceae genomes: Going from unattainable to becoming a model clade for comparative plant genomics. Plant Physiol 149: 111–116.

Campbell, M. A., W. Zhu, N. Jiang, H. Lin, S. Ouyang, et al. 2007. Identification and characterization of lineage-specific genes within the Poaceae. Plant Physiol 145: 1311–1322.

Chaw, S.M., C.C. Chang, H.L. Chen, W.H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58: 424–441.

Ciccarelli, F. D., C. von Mering, M. Suyama, E. D. Harrington, E. Izaurralde, and P. Bork. 2005. Complex genomic rearrangements lead to novel primate gene function. Genome Res 15: 343-351.

Hurst, L. D., C. Pal, and M. J. Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299–310.

Ilic, K., P. J. SanMiguel, and J. L. Bennetzen. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes. Proc Natl Acad Sci 100: 12265–12270.

Kensche, P. R., M. Oti, B. E. Dutilh, and M. A. Huynen. 2008. Conservation of divergent transcription in fungi. Trends in Genet 24: 207-211.

Krom, N, and W. Ramakrishna. 2008. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and *Populus*. Plant Physiol 147: 1763-1773.

Lai, J., J. Ma, Z. Swigonova, W. Ramakrishna, et al. 2004. Gene loss and movement in the maize genome. Genome Res 14: 1924-1931.

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

Semon, M, and L. Duret. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol 23: 1715–1723.

Singer, G. A., A. T. Lloyd, L. B. Huminiecki, and K. H. Wolfe. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol 22: 767–775.

Tuskan, G.A., S. DiFazio, S. Jansson, J. Bohlmann, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596–1604.

Wolfe, K.H., M. Gouy, Y.W. Yang, P.M. Sharp, and W.H. Li. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86: 6201–6205.

Table 1: Rice Gene Pair Conservation and Rearrangement
All percentages are out of the total number of pairs.

Rice vs. Sorghum		Conserved				Inverted				Rearranged		Missing Homologs					
	Total Pairs	Total		No insertions		Total		No insertions				Any		Both		One	
Convergent	4800	2740	57.1%	2036	42.4%	163	3.4%	54	1.1%	1208	25.2%	689	14.4%	120	2.5%	569	11.9%
Divergent	3711	1908	51.4%	1100	29.6%	97	2.6%	35	0.9%	1101	29.7%	605	16.3%	112	3.0%	493	13.3%
Tandem	9428	5636	59.8%	3503	37.2%	166	1.8%	52	0.6%	2112	22.4%	1514	16.1%	409	4.3%	1105	11.7%

Rice vs. Maize		Conserved				Inverted				Rearranged		Missing Homologs					
	Total Pairs	Total		No insertions		Total		No insertions				Any		Both		One	
Convergent	4800	1686	35.1%	1369	28.5%	160	3.3%	58	1.2%	1962	40.9%	992	20.7%	261	5.4%	731	15.2%
Divergent	3711	656	17.7%	441	11.9%	194	5.2%	74	2.0%	2015	54.3%	846	22.8%	196	5.3%	650	17.5%
Tandem	9428	3146	33.4%	2132	22.6%	250	2.7%	106	1.1%	3799	40.3%	2233	23.7%	781	8.3%	1452	15.4%

Rice vs. *Brachypodium*		Conserved				Inverted				Rearranged		Missing Homologs					
	Total Pairs	Total		No insertions		Total		No insertions				Any		Both		One	
Convergent	4800	2670	55.6%	2063	43.0%	190	4.0%	73	1.5%	1226	25.5%	714	14.9%	160	3.3%	554	11.5%
Divergent	3711	1855	50.0%	1222	32.9%	138	3.7%	50	1.3%	1079	29.1%	639	17.2%	131	3.5%	508	13.7%
Tandem	9428	5456	57.9%	3497	37.1%	260	2.8%	98	1.0%	2148	22.8%	1564	16.6%	493	5.2%	1071	11.4%

Table 2: Sorghum Gene Pair Conservation and Rearrangement
All percentages are out of the total number of pairs.

Sorghum vs. Rice		Conserved				Inverted				Rearranged		Missing Homologs					
	Total Pairs	Total		No insertions		Total		No insertions				Any		Both		One	
Convergent	5059	2751	54.4%	2018	39.9%	162	3.2%	52	1.0%	1723	34.1%	423	8.4%	63	1.2%	360	7.1%
Divergent	4913	2327	47.4%	1306	26.6%	165	3.4%	51	1.0%	1916	39.0%	505	10.3%	63	1.3%	442	9.0%
Tandem	11847	6731	56.8%	4061	34.3%	191	1.6%	57	0.5%	3688	31.1%	1237	10.4%	281	2.4%	956	8.1%

Sorghum vs. Maize		Conserved				Inverted				Rearranged		Missing Homologs					
	Total Pairs	Total		No insertions		Total		No insertions				Any		Both		One	
Convergent	5059	2348	46.4%	1897	37.5%	208	4.1%	72	1.4%	2053	40.6%	450	8.9%	38	0.8%	412	8.1%
Divergent	4913	1193	24.3%	754	15.3%	350	7.1%	96	2.0%	2864	58.3%	506	10.3%	40	0.8%	466	9.5%
Tandem	11847	5204	43.9%	3442	29.1%	381	3.2%	121	1.0%	5039	42.5%	1223	10.3%	218	1.8%	1005	8.5%

Sorghum vs. *Brachypodium*		Conserved				Inverted				Rearranged		Missing Homologs					
	Total Pairs	Total		No insertions		Total		No insertions				Any		Both		One	
Convergent	5059	2461	48.6%	1797	35.5%	190	3.8%	65	1.3%	1758	34.7%	650	12.8%	105	2.1%	545	10.8%
Divergent	4913	2100	42.7%	1336	27.2%	204	4.2%	57	1.2%	1827	37.2%	782	15.9%	106	2.2%	676	13.8%
Tandem	11847	6029	50.9%	3788	32.0%	305	2.6%	102	0.9%	3696	31.2%	1817	15.3%	475	4.0%	1342	11.3%

Table 3: Rice gene pairs - correlated vs. uncorrelated

Figures in the "Z" column are test statistics of the binomial test. Figures in bold denote significant differences (P < 0.0001) in the frequency of strongly correlated pairs in each category compared to the frequency of uncorrelated pairs.

Rice vs. Sorghum		**Conserved**							**Inverted**					
			Total			**No Insertions**				**Total**			**No Insertions**	
		Total	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	Correlated	329	199	60.5%		153	46.5%		13	4.0%		7	2.1%	
	Uncorrelated	4471	2530	56.6%	***5.26***	1883	42.1%	***5.94***	161	3.6%	1.26	47	1.1%	***7.06***
Divergent	Correlated	296	170	57.4%		97	32.8%		10	3.4%		4	1.4%	
	Uncorrelated	3415	1728	50.6%	***7.99***	1003	29.4%	***4.36***	97	2.8%	1.89	31	0.9%	2.73
Tandem	Correlated	651	422	64.8%		276	42.4%		8	1.2%		4	0.6%	
	Uncorrelated	8777	5214	59.4%	***10.34***	3227	36.8%	***10.94***	158	1.8%	***-4.03***	48	0.5%	0.86

Rice vs. Maize		**Conserved**							**Inverted**					
			Total			**No Insertions**				**Total**			**No Insertions**	
		Total	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	Correlated	329	124	37.7%		101	30.7%		13	4.0%		5	1.5%	
	Uncorrelated	4471	1559	34.9%	***3.96***	1268	28.4%	3.47	150	3.4%	2.21	53	1.2%	2.07
Divergent	Correlated	296	66	22.3%		42	14.2%		13	4.4%		4	1.4%	
	Uncorrelated	3415	589	17.2%	***7.81***	399	11.7%	***4.56***	182	5.3%	-2.44	70	2.0%	-2.88
Tandem	Correlated	651	243	37.3%		166	25.5%		15	2.3%		6	0.9%	
	Uncorrelated	8777	2903	33.1%	***8.47***	1966	22.4%	***6.97***	235	2.7%	-2.17	100	1.1%	-1.92

Rice vs. *Brachypodium*		**Conserved**							**Inverted**					
			Total			**No Insertions**				**Total**			**No Insertions**	
		Total	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	Correlated	329	188	57.1%		141	42.9%		22	6.7%		7	2.1%	
	Uncorrelated	4471	2472	55.3%	2.49	1922	43.0%	-0.18	178	4.0%	***9.25***	66	1.5%	3.61
Divergent	Correlated	296	169	57.1%		116	39.2%		7	2.4%		2	0.7%	
	Uncorrelated	3415	1677	49.1%	***9.34***	1106	32.4%	***8.50***	140	4.1%	***-5.11***	48	1.4%	-3.62
Tandem	Correlated	651	416	63.9%		268	41.2%		18	2.8%		8	1.2%	
	Uncorrelated	8777	5040	57.4%	***12.28***	3229	36.8%	***8.51***	242	2.8%	0.04	90	1.0%	1.89

Table 3 (continued): Rice gene pairs - correlated vs. uncorrelated

Figures in the "Z" column are test statistics of the binomial test. Figures in bold denote significant differences (P < 0.0001) in the frequency of strongly correlated pairs in each category compared to the frequency of uncorrelated pairs.

Rice vs. Sorghum			**Rearranged**			**Missing Homologs**								
						Any			**Both**			**One**		
		Total	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	Correlated	329	83	25.2%		34	10.3%		7	2.1%		27	8.2%	
	Uncorrelated	4471	1125	25.2%	0.10	655	14.6%	***-8.16***	113	2.5%	-1.70	542	12.1%	***-8.02***
Divergent	Correlated	296	74	25.0%		42	14.2%		7	2.4%		35	11.8%	
	Uncorrelated	3415	1027	30.1%	***-6.46***	563	16.5%	-3.62	105	3.1%	-2.40	458	13.4%	-2.72
Tandem	Correlated	651	113	17.4%		108	16.6%		31	4.8%		77	11.8%	
	Uncorrelated	8777	1999	22.8%	***-12.10***	1406	16.0%	1.46	378	4.3%	2.10	1028	11.7%	0.34

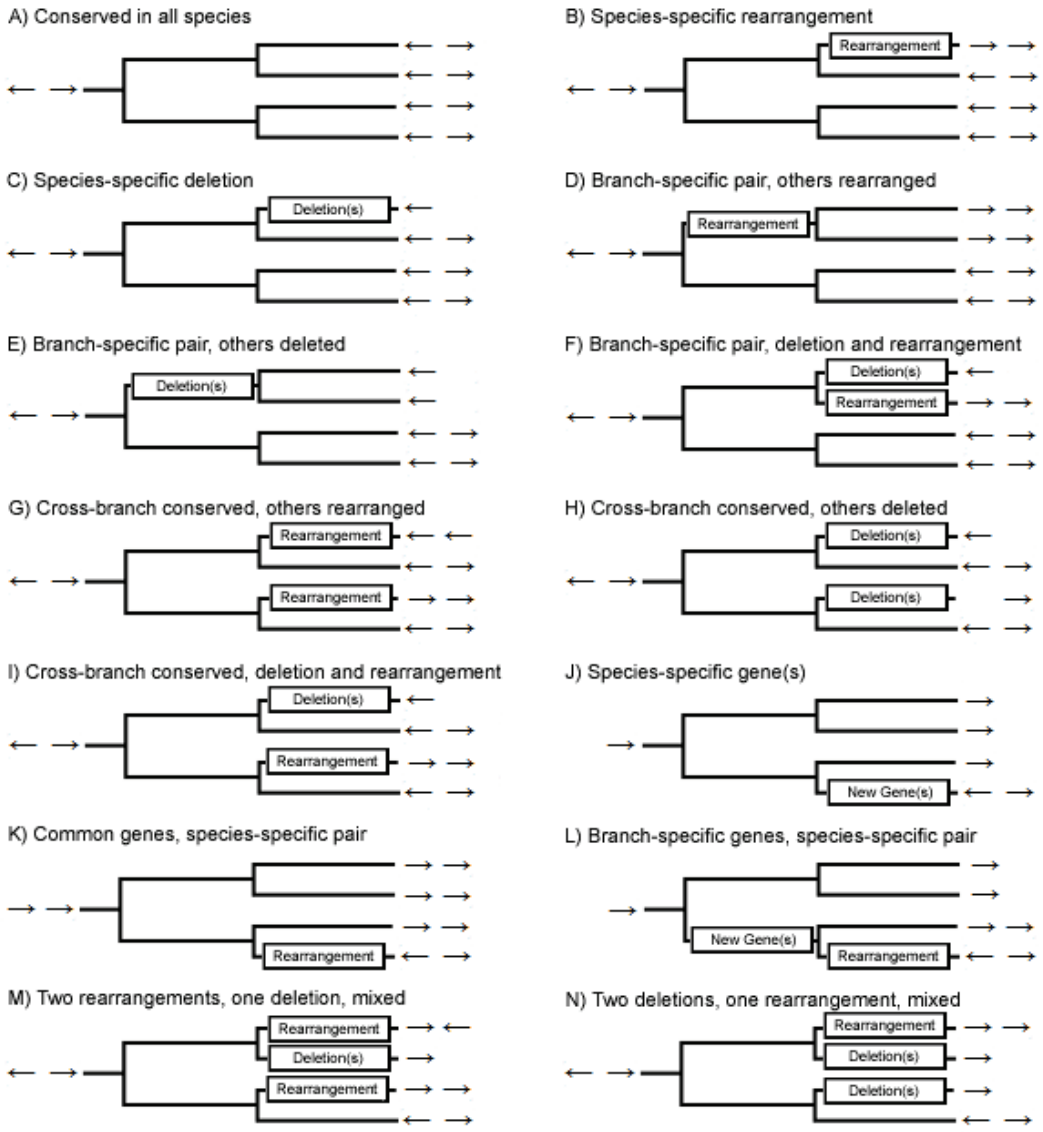
Rice vs. Maize			**Rearranged**			**Missing Homologs**								
						Any			**Both**			**One**		
		Total	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	Correlated	329	140	42.6%		52	15.8%		14	4.3%		38	11.6%	
	Uncorrelated	4471	1822	40.8%	2.45	940	21.0%	***-8.56***	247	5.5%	-3.71	693	15.5%	***-7.30***
Divergent	Correlated	296	148	50.0%		69	23.3%		18	6.1%		51	17.2%	
	Uncorrelated	3415	1867	54.7%	***-5.48***	777	22.8%	0.78	178	5.2%	2.28	599	17.5%	-0.48
Tandem	Correlated	651	230	35.3%		163	25.0%		68	10.4%		95	14.6%	
	Uncorrelated	8777	3569	40.7%	***-10.17***	2070	23.6%	3.21	713	8.1%	7.96	1357	15.5%	-2.25

Rice vs. *Brachypodium*			**Rearranged**			**Missing Homologs**								
						Any			**Both**			**One**		
		Total	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	Correlated	329	72	21.9%		47	14.3%		9	2.7%		38	11.6%	
	Uncorrelated	4471	1154	25.8%	***-6.00***	667	14.9%	-1.19	151	3.4%	-2.38	516	11.5%	0.02
Divergent	Correlated	296	68	23.0%		52	17.6%		12	4.1%		40	13.5%	
	Uncorrelated	3415	1011	29.6%	***-8.49***	587	17.2%	0.59	119	3.5%	1.81	468	13.7%	-0.32
Tandem	Correlated	651	110	16.9%		107	16.4%		38	5.8%		69	10.6%	
	Uncorrelated	8777	2038	23.2%	***-14.03***	1457	16.6%	-0.41	455	5.2%	2.76	1002	11.4%	-2.41

Table 4: Evolutionary History of Rice and Sorghum Gene Pairs

	Rice						Sorghum					
	Convergent		Divergent		Tandem		Convergent		Divergent		Tandem	
A) Pair with conserved orientation in four species	900	18.8%	215	5.8%	984	10.4%	929	18.4%	247	5.0%	1286	10.9%
B) Pair with conserved orientation in three species												
One rearrangement	885	18.4%	591	15.9%	1833	19.4%	939	18.6%	782	15.9%	2246	19.0%
One deletion	50	1.0%	25	0.7%	133	1.4%	46	0.9%	24	0.5%	138	1.2%
C) Pair with conserved orientation in two species												
Branch-specific pairs												
Others rearranged	427	8.9%	444	12.0%	916	9.7%	425	8.4%	222	4.5%	853	7.2%
Others deleted	21	0.4%	16	0.4%	65	0.7%	12	0.2%	9	0.2%	90	0.8%
One deletion, one rearrangement	37	0.8%	23	0.6%	97	1.0%	27	0.5%	20	0.4%	89	0.8%
Cross-branch conserved pairs												
Others rearranged	346	7.2%	356	9.6%	968	10.3%	438	8.7%	736	15.0%	1420	12.0%
Others deleted	14	0.3%	8	0.2%	61	0.6%	12	0.2%	5	0.1%	48	0.4%
One deletion, one rearrangement	53	1.1%	39	1.1%	141	1.5%	41	0.8%	51	1.0%	165	1.4%
D) Pair with unique orientation in one species												
Species-specific gene(s)	499	10.4%	449	12.1%	1114	11.8%	179	3.5%	189	3.8%	482	4.1%
Common genes, species-specific pair	1095	22.8%	1118	30.1%	2190	23.2%	1425	28.2%	1877	38.2%	3551	30.0%
Branch-specific genes, species-specific pair	111	2.3%	91	2.5%	250	2.7%	127	2.5%	197	4.0%	381	3.2%
Two rearrangements, one deletion, mixed	258	5.4%	246	6.6%	506	5.4%	348	6.9%	399	8.1%	795	6.7%
One rearrangement, two deletions, mixed	104	2.2%	90	2.4%	170	1.8%	111	2.2%	155	3.2%	303	2.6%

Figure 1 - Categories of Gene Pair Evolution: Each image is a representative of the many specific scenarios that may be found in that category. The bottom branch of each tree represents the species in which the pair was first identified (i.e. either rice or sorghum), and the two genes in question are shown in a divergent pair in these examples. Rearrangements are represented by the inversion of one gene, but may also involve inversion of both genes, insertions within the pair, or translocation to other regions or chromosomes. Likewise, deletions may involve one gene, as shown, or both genes in the pair. In some of the scenarios where the pair is conserved in two species (D, E, G, and H), the rearranged or deleted states are just as likely to be the ancestral state as the divergent pair shown. In scenario L, it is also possible that both genes existed in the common ancestor and a deletion took place in the top branch rather than new gene(s) being created.



CHAPTER 4:

RETROTRANSPOSON INSERTIONS ASSOCIATED WITH RICE GENE PAIR CONSERVATION AND REARRANGEMENT IN THREE GRASS GENOMES

Nicholas Krom and Wusirika Ramakrishna

4.1 ABSTRACT

Small-scale changes in gene order and orientation are common in plant genomes, even across relatively short evolutionary distances. We have previously characterized and quantified the types of genomic changes in pairs of adjacent genes in rice, sorghum, maize, and *Brachypodium*. In the present study, we investigate the correlation between retrotransposon association with rice gene pairs and the conservation and rearrangement of these gene pairs in three other grass genomes. Retrotransposons are a major component of most sequenced plant genomes, and may play a role in genomic rearrangements. We identified retrotransposon insertions (mostly fragmentary) within, between, and flanking pairs of rice genes in convergent, divergent, and tandem arrangement, and tested for significant differences in the frequency of gene pair conservation, inversion, rearrangement, and deletion among retrotransposon-associated pairs compared to the general populations of similar gene pairs. *Copia* and *Gypsy* LTR-retrotransposon insertions were found to be associated with reduced frequency of gene pair conservation and an increase in both gene pair rearrangement and gene deletions. LINEs and SINEs are also associated with reduced conservation, albeit to a lesser degree, and significantly increase gene deletions only. Convergent gene pairs were subject to these changes most often, while divergent pairs showed the least significant effects of retrotransposon insertion. Insertions within and between genes were associated with the greatest effects on gene pair arrangement, while insertions flanking the gene pair had significant effects only on divergent pairs. The observed effects were considerably weaker in maize than in sorghum or *Brachypodium*.

4.2 INTRODUCTION

A recurring theme in the field of plant comparative genomics is the tremendous amount of variation in genome size, gene order, and retrotransposon content among plant genomes. This variation is caused by a wide range of mechanisms, including gene and genome duplication, gene deletion, transposable element amplification, transposon mediated gene movement, polyploidization, and various types of recombination (Bennetzen and Chen, 2008). The combined action of these mechanisms can result in an astonishing degree of polymorphism within orthologous regions of closely related species (Bennetzen, 2007). For example, a detailed examination of the *Adh1* region in nine species within the genus *Oryza* (Ammiraju et al., 2008) identified deletions and duplications of genes and gene clusters, highly variable retrotransposon content, and segmental inversions and deletions of regions hundreds of kilobases long.

While there are many different forces that produce such changes in genome content, retrotransposons are one of the most influential, both through direct means, such as transposition into a new genomic locus, and through the various processes they promote, such as chromosome breakage. Differential retrotransposon activity between species is one of the primary contributors to the wide range of genome sizes observed among the grasses (Bennetzen, 2005), with rapid expansion of genome size occurring during bursts of amplification which are typically followed by rapid loss of retrotransposon sequence (Vitte et al., 2007). Much of this sequence loss is believed to occur through unequal homologous recombination or illegitimate recombination, which can at time delete sequence from the host genome in addition to retrotransposons (Devos et al., 2002; Ma et al., 2004).

Another common area of study involves functional interaction between neighboring genes. In plants, correlated expression has been observed in many pairs of adjacent genes (Krom and Ramakrishna, 2008; Williams and Bowles, 2004) and in local groups of two to four genes (Ren et al., 2005). A comparative analysis of convergent and divergent gene pairs in rice, *Arabidopsis,* and *Populus trichocarpa* (Krom and Ramakrishna, 2008) found that the arrangement of these gene pairs is conserved significantly more frequently when the paired genes displayed strongly correlated expression levels, and thus the genes' regulation may be dependent on maintaining a specific relative arrangement. Other studies have concluded that gene order is sometimes connected to gene function and regulation (Hurst et al., 2004; Ciccarelli, 2005), thus genomic rearrangements may result in changes in phenotype.

Due to the potential importance of gene pair order and orientation to proper function and regulation, and the role of retrotransposons in creating and promoting genomic rearrangements, we propose to investigate the possible correlation between the

presence of retrotransposons within gene pairs and the frequency of gene pair conservation or alteration. We have previously identified retrotransposon insertions inside or within 1000 bp upstream of approximately one-sixth of all rice genes (Krom et al., 2008). Other studies have observed frequent rearrangements in rice gene pairs in sorghum, maize, and *Brachypodium* (Krom and Ramakrishna, 2009) and in an orthologous region of the genomes of rice, sorghum, and maize (Ilic et al., 2003). Therefore, to determine whether a link exists between these two phenomena, we identified retrotransposon insertions in and near rice convergent, divergent, and tandem gene pairs. The frequencies of gene pair conservation, inversion, rearrangement, and deletion among the various classes of retrotransposon-associated gene pairs were then compared with the corresponding frequencies among all gene pairs of similar type, as determined by Krom and Ramakrishna (2009). Many significant differences were observed in the evolutionary behavior of the general population of rice gene pairs and those pairs containing retrotransposon insertions, which supports our hypothesis that retrotransposons promote several types of small-scale genomic rearrangements.

4.3 RESULTS

Retrotransposon Content of Gene Pairs

Our analysis began with the identification of all retrotransposons closely associated with the rice gene pairs described in our previous study (Krom and Ramakrishna, 2009). Retrotransposons were identified in three regions: within the genes themselves, in the intergenic region between the two genes, and in the intergenic regions flanking the gene pair. Four classes of retrotransposons were identified: *Copia* and *Gypsy* LTR-retrotransposons, Long Interspersed Nuclear Elements (LINEs), and Short Interspersed Nuclear Elements (SINEs).

A great deal of variation was observed among convergent, divergent, and tandem gene pairs with regard to the prevalence of retrotransposon insertions in these three regions (Table 1). Convergent pairs showed little preference for insertions in any particular region, with 8.2% of pairs being flanked by some type of retrotransposon, 11% of pairs having retrotransposons within one or both genes, and 13.6% having insertions between their genes. In contrast, retrotransposons were identified in the genes of only 11.4% of divergent pairs, but were found to flank 41.1% of such pairs. The most common positions for retrotransposons in and near tandem pairs also differed greatly, with flanking insertions being least common (6.7% of pairs) and intergenic insertions being most common by a significant margin (26.3% of pairs). One measure in which the three pair types varied little was the proportion of pairs with retrotransposon insertions within genes, which ranged from 11% of convergent pairs to 11.6% of tandem pairs. Far more variation was observed among intergenic insertions (from 13.6% of convergent pairs to 31.6% of divergent pairs) and flanking insertions (from 6.7% of tandem pairs to 41.1% of divergent pairs).

Considerable differences were also observed between the four types of retrotransposons. SINE insertions were the most common, followed closely by *Gypsy* insertions. LINE and *Copia* elements were both found considerably less often, usually being one-third to one-half as numerous as the other types. Among insertions within genes, in all three pair types *Copia* insertions are the least common, SINE insertions most common, and there exists an approximate two-fold difference in frequency between the two (e.g. 1.9% of tandem pairs with *Copia* insertions in genes, compared to 4.9% with SINE insertions). A similar distribution of frequencies can be seen among insertions flanking convergent and tandem gene pairs. In contrast, the frequency of intergenic retrotransposon insertions varies much more among retrotransposon types. The rates of *Copia* and LINE insertions are generally quite similar, as are those of *Gypsy* and SINE insertions; however, a three- to four-fold difference exists between these two groups.

Insertions flanking divergent pairs also follow this pattern, with 5.2% of such pairs being flanked by *Copia* elements, compared to 19.5% being flanked by *Gypsy* elements.

Retrotransposons and Gene Pair Evolution

Gene pairs that were found to contain retrotransposon insertions were compared as a group with the complete set of gene pairs of that type (convergent, divergent, or tandem) to identify any significant differences in the frequency of gene pair conservation, inversion, rearrangement, or gene deletion in three other grass genomes. The possible effects of retrotransposon insertion on these evolutionary events were analyzed for all retrotransposons in general, as well as for the four major classes of retrotransposons: *Copia* and *Gypsy* LTR-retrotransposons, LINEs, and SINEs. The statistical significance of the differences in the frequency of the various evolutionary events between all pairs of a given type and those pairs containing retrotransposons was measured using the normal approximation of the binomial test, with a P-value cutoff of $P < 0.01$ ($|Z| > 2.3267$).

There were several significant effects common to retrotransposon insertions of all types and locations (Tables 2-4), with some exceptions and variation between species and pair types. Rice gene pairs with retrotransposons within or near them are less likely to have their orientation conserved in other species, sometimes by a substantial margin. For example, 57.1% of all rice convergent pairs are conserved in sorghum, while only 36.6% of similar pairs with retrotransposons in their intergenic regions are conserved (Table 3A). Similarly, retrotransposon association makes gene pairs more likely to be rearranged, with both genes conserved but no longer near each other. The frequencies with which gene pairs are found to be missing homologs in other species also correlate with increased presence of retrotransposons. Although retrotransposons tend to increase the rate of both gene pair rearrangement and gene deletion, they appear to promote deletion to a greater degree. The effect of retrotransposons on the likelihood of one or both genes in a pair to be inverted varies considerably depending on the insertion type and location, such that no typical trend could be discerned.

In general, these effects are greatest (i.e. the largest difference between all pairs and those with retrotransposons) when the retrotransposon is inserted between the two genes (Table 3). Insertions within the genes (Table 2) show similar effects to those between genes, although the magnitude of the difference is usually less. Retrotransposons in the intergenic regions flanking a gene pair (Table 4) tend to have the weakest effects, displaying a great deal of variation among the different species, insertion types, and gene pair types.

Convergent pairs are the type most frequently disrupted by retrotransposon insertions in and between genes, usually having the largest decreases in conservation and increases in rearrangements and gene deletions. Divergent pairs show the least effect from these types of insertions, but are the pair type most affected by insertions flanking the gene pair, showing the greatest reductions in conservation of any pair type in response to that class of insertion.

These overall trends were followed quite closely in the results of the analyses of rice gene pairs in sorghum and *Brachypodium* (Tables 2A, 2C, 3A, 3C, 4A, 4C), although some differences between the two species were observed. Insertions in the genes of tandem pairs are associated with modest, but not statistically significant, reduction in the frequency of inversions in *Brachypodium*, while these insertions had no significant effect on inversions in sorghum. In contrast, insertions flanking tandem pairs tend to increase the frequency of inversions in sorghum, yet had no such effect in *Brachypodium*. Retrotransposons flanking tandem pairs also appeared to be significantly associated with an increase in the number of pairs missing homologs in *Brachypodium* but not in sorghum.

The maize homologs of retrotransposon-associated rice gene pairs differed in several ways from their counterparts in sorghum and *Brachypodium*. Retrotransposons within genes (Table 2B) are coupled to a reduced frequency of gene pair conservation considerably less in maize than in sorghum or *Brachypodium*, with no significant effect on rearrangements, and they increased the number of missing homologs. Insertions between genes (Table 3B) are linked to a smaller reduction in the conservation frequency of tandem pairs and a smaller increase in the rearrangement frequency of tandem and convergent pairs. Unlike sorghum and *Brachypodium*, retrotransposons flanking gene pairs (Table 4B) had no significant effect on gene pair conservation in maize. Only in maize were convergent pairs flanked by retrotransposons rearranged significantly more often and missing homologs significantly less often.

The observations above describe the effect of retrotransposons as a whole on gene pair evolution. To determine what differences may exist among the various types of retrotransposons with regard to their influence on gene pairs, this data will serve as a baseline for comparison with similar analyses focusing only on gene pairs containing a specific type of retrotransposon.

Copia LTR-retrotransposons

The most striking feature of the *Copia*-only data is the marked decrease in conservation rates among nearly all species and insertion locations. Thus *Copia* LTR-

retrotransposon association with genes appears to disrupt gene pairs substantially more than the average rice retrotransposon. These reductions in conservation are accompanied by widespread growth in the fractions of rearranged pairs and large, nearly universal increases in the fractions of pairs missing homologs. While there are some sizeable changes in the frequencies of inversions in *Copia*-associated pairs, the extremely small number of pairs involved (no more than 14 in any set) makes it impossible to reliably state that these changes are due to different behavior of *Copia* elements compared to other types of retrotransposons rather than mathematical artifacts.

Copia insertions within genes (Table 2) are associated with larger increases in the fractions of convergent and tandem pairs rearranged in sorghum and *Brachypodium*, as well as divergent pairs in sorghum. However, *Copia* LTR-retrotransposons appear to have no significant effect on the fraction of pairs rearranged in maize, as do retrotransposons in general. The proportions of rice gene pairs missing homologs in other grass genomes are also greater in the presence of *Copia* elements inside the genes of convergent and tandem pairs, except in *Brachypodium*, where the magnitude of the insertion effect is nearly the same as that of all retrotransposons. In divergent pairs, *Copia* insertions behave much like any other retrotransposon with regard to missing homologs, aside from a modest increase in sorghum.

The presence of *Copia* elements between genes (Table 3) has some effects that differ from the overall average for retrotransposons. In sorghum and *Brachypodium*, rearrangement is more common among all convergent pairs and sorghum tandem pairs and less common (but still not significantly different) among all divergent pairs and tandem pairs in maize and *Brachypodium*. *Copia* LTR-retrotransposons between genes are also linked to significantly higher frequencies of missing homologs in all types of gene pairs, and to a higher degree than retrotransposons in general, especially among divergent pairs.

Copia elements flanking gene pairs (Table 4) greatly increase the frequency of missing homologs in all pair types except convergent pairs in maize. They are also correlated with increased rearrangement rates in divergent pairs in sorghum and *Brachypodium* and in tandem pairs in sorghum. In addition, these insertions appear to lower the frequency of rearrangement of tandem pairs in maize and *Brachypodium* and have no significant effect on rearrangement in convergent pairs in maize.

Overall, *Copia* LTR-retrotransposon insertions are correlated with reduced rates of gene pair conservation. They appear to disrupt gene pair arrangement primarily through the loss of homologous genes, and to a lesser extent the physical relocation of previously paired genes. Convergent pairs are most likely to be disturbed by the presence

of *Copia* elements, while divergent pairs are affected least often, except by *Copia* elements flanking the gene pair.

Gypsy LTR-retrotransposons:

Gypsy LTR-retrotransposons appear to influence gene pair evolution in many of the same ways as their relatives in the *Copia* family. However, while sizeable differences remain among the three types of gene pair, all types tend to be significantly affected, with a few exceptions. Gene pairs associated with nearly all varieties of *Gypsy* elements are conserved significantly less often. Insertions in and between genes appear to have the greatest effect, especially among convergent pairs, where the rate of conservation is at times cut in half compared to such pairs as a whole. *Gypsy* elements flanking gene pairs also tend to significantly reduce the rate of conservation, with the exception of convergent pairs in sorghum and tandem pairs in maize. In addition, the possible effect of flanking *Gypsy* retrotransposons is overall less dramatic than those within the pair and greatest on divergent pairs, two tendencies that were also observed with *Copia* retrotransposons. Inversions are slightly less common among most classes of *Gypsy*-associated gene pairs, although the differences are rather small and as before the small number of pairs involved makes the tests less reliable.

While *Gypsy* insertions of all types tend to reduce conservation, some variation among them does exist. Insertions within genes (Table 2) appear to correlate with an increase in the frequency of gene pair rearrangement in both sorghum and *Brachypodium*, while rearrangements are significantly less common in divergent pairs and unaffected by insertions in convergent and tandem pairs in maize. Having *Gypsy* insertions in one or both genes in rice makes gene pairs more likely to be missing one or both homologs in sorghum, maize, and *Brachypodium*. *Gypsy* elements promote this type of event to a similar degree as *Copia* elements, and to a far greater degree than do retrotransposon insertions in general.

Among rice gene pairs with *Gypsy* insertions between their genes (Table 3), rearrangements are significantly more common in sorghum and *Brachypodium*, although only slightly more so than among those containing any type of retrotransposon. In maize, rearrangement is significantly more common in convergent pairs, and less common in tandem pairs, while divergent pairs show no significant difference. The frequency of gene deletion is increased across the board, although the effect is relatively weak among divergent pairs. Compared with *Copia* insertions, *Gypsy* insertions are associated with considerably more deletions among convergent pairs and the same or slightly fewer in divergent and tandem pairs.

Gypsy insertions in the regions flanking rice gene pairs (Table 4) were found to correlate with an increase in the frequency of rearrangement only in divergent pairs in sorghum and *Brachypodium*, while rearrangements in maize were less common in the presence of these insertions. Convergent pairs with this type of retrotransposon in rice were considerably more likely to be rearranged in maize, as well. Rice divergent pairs flanked by *Gypsy* elements were significantly more likely to have one or both genes deleted in all three species. This increase was also observed in convergent and tandem pairs with deletions in sorghum and *Brachypodium*, while deletions in maize were not affected in these types of pairs.

Gypsy LTR-retrotransposons are powerful agents of gene pair disruption. With some exceptions, their association with rice gene pairs correlates with an increase in the frequency of gene pair rearrangement and homolog deletion. Compared to their *Copia* brethren, presence of *Gypsy* elements is somewhat more likely to rearrange any gene pair with which they are closely associated in rice and to delete genes in convergent pairs. In other pair types, the presence of *Gypsy* elements appears to delete genes slightly less often than *Copias*, while both types are substantially more likely to do so than the general population of retrotransposon insertions.

LINEs:

Long Interspersed Nuclear Elements (LINEs) differ greatly from the two classes of LTR-retrotransposons described above, in both their structure and their effects on gene pair evolution. Conservation rates of rice gene pairs whose genes contain LINE insertions (Table 2), while still lower than those of all gene pairs, are generally similar to or higher than those of all retrotransposon-associated gene pairs. When present between genes (Table 3), LINEs slightly decrease the frequency of gene pair conservation, especially among convergent pairs. For both types of LINE insertion, the magnitude of their effects on conservation is rather small compared to that of *Copia* or *Gypsy* LTR-retrotransposons. LINE insertions that flank gene pairs (Table 4) have rather variable effects on conservation. Conservation of convergent pairs is considerably lower in sorghum and *Brachypodium*, and tandem pairs are slight less frequently conserved in these species as well. Conservation of divergent gene pairs with flanking LINE insertions is higher than that of divergent pairs flanked by retrotransposons in general, although the rates are still slightly lower than those of all gene pairs. The presence of flanking LINEs significantly increases the fraction of conserved tandem pairs in maize, while convergent and divergent pairs are not significantly affected.

Inversion of one or both genes in a pair is typically unaffected or slightly reduced by the presence of LINEs within genes. LINEs between genes are associated with an increase in the frequency of inversion in convergent pairs in all species and divergent pairs in maize and *Brachypodium*, and have no noticeable effect on tandem pairs. The fraction of inverted pairs is increased significantly among convergent and tandem pairs in all species, and is reduced slightly among divergent pairs, when LINEs are found in the flanking intergenic regions of a pair.

LINE insertions within genes do not appear to affect the frequency of gene pair rearrangement in either sorghum or *Brachypodium*, while in maize there is a small but significant decrease in rearrangement among convergent and divergent pairs. Convergent pairs with LINEs in their intergenic regions are more likely to be rearranged, especially in *Brachypodium*. Divergent and tandem pairs also show slight increases in rearrangement frequency. LINEs flanking the outside of gene pairs are weakly associated with decreased gene pair rearrangements, although their influence is not statistically significant.

Gene pairs of all types with LINE insertions within one or both genes are more likely to be missing homologs in sorghum, maize, and *Brachypodium*. LINEs between genes have a similar effect, although convergent pairs generally show the most significant increases. This trend is continued among pairs flanked by LINE insertions, but the fractions of pairs missing homologs are only slightly higher among gene pairs with this type of retrotransposon insertion.

Overall, gene pairs associated with LINE insertions are less likely to be conserved. LINEs are, however, less likely to interfere with conservation than LTR-retrotransposons. In addition, their association with rice gene pairs appears to break up gene pairs in other grass genomes almost entirely by promoting deletion of homologs, and have little effect on gene pair rearrangements.

SINEs:

As has been the case with the other types of retrotransposon insertions described here, SINE insertions associated with rice gene pairs appear to reduce the frequency with which the orientation of gene pairs is conserved in other species. SINEs within genes (Table 2) are linked to a significant reduction in the conservation frequency of convergent pairs in sorghum, maize, and *Brachypodium*, and of tandem pairs in *Brachypodium*. When the SINEs are found between the two genes (Table 3), the greatest reductions in conservation are again found in convergent and tandem pairs, while divergent pairs show no significant change in any species. Convergent gene pairs flanked

by SINE insertions (Table 4) are more likely to be conserved in all three species, with statistically significant increases in sorghum and *Brachypodium*, while conservation of divergent and tandem pairs are slightly less likely in the presence of such insertions. In general, the probable impact of SINEs on conservation frequency is greater than that of LINEs and considerably less than that of both types of LTR-retrotransposons.

Inversion of one or both genes in a pair is not promoted or prevented by SINE insertions within genes or flanking gene pairs. SINEs between genes in convergent pairs are twice as likely to contain an inverted gene in maize as are convergent pairs in general. Gene inversion is somewhat more frequent among all pairs with this type of insertion.

Gene pairs tend to be more frequently rearranged when they are closely associated with SINE insertions. Among gene pairs with SINEs inside genes, the largest increases in rearrangement are seen in convergent and divergent pairs in sorghum and *Brachypodium*. Tandem pairs in sorghum and *Brachypodium*, and all pairs in maize are not significantly affected by these insertions. Rice gene pairs with SINEs between their genes are more likely to be rearranged in other grass species. The increase is statistically significant among convergent pairs in sorghum and maize, and tandem pairs in maize and *Brachypodium*. Divergent gene pairs flanked by SINEs are rearranged significantly more often in sorghum and *Brachypodium*.

Rice genes that contain SINE insertions and are part of convergent or tandem pairs are somewhat more likely to be deleted in sorghum, maize, and *Brachypodium*. SINEs between genes have little effect on homolog deletions, except for those in convergent pairs in *Brachypodium* and maize. Divergent pairs show slight decreases in deletions, while convergent and tandem pairs are slightly more likely to have missing homologs in all three species. SINE insertions flanking divergent and tandem gene pairs have little effect on homolog deletions, while deletions are considerably more rare among convergent pairs.

To summarize, the presence of SINEs in close proximity to a rice gene pair typically correlates with a modest decrease in the probability of that pair being conserved. Among the non-conserved pairs, rearrangement of homologs is somewhat more common than their deletion. Gene inversions also appear to be more commonly associated with SINEs than with other types of retrotransposons, but remain quite rare overall.

4.4 DISCUSSION

It is clear that rice gene pairs closely associated with retrotransposon insertions are less likely to be conserved in other grass species. Considerable variation exists among different families of retrotransposons with regard to the types of structural changes they are associated with, as well as the magnitude of the influence they exert on gene pairs. The location of the retrotransposon relative to the gene pair also has a significant influence on their interaction over time. Another major factor in the interaction between gene pairs and their associated retrotransposons is the strand-wise arrangement of the paired genes, either convergent, divergent, or tandem. The results of the analyses presented here also vary among the three genomes with which the retrotransposon-associated gene pairs were compared.

Considerable variation was observed in the frequencies of intergenic and flanking retrotransposon insertions among the three gene pair types. Intergenic insertions were most commonly found within divergent pairs (31.6%), slightly less common in tandem pairs (26.3%), and least common in convergent pairs (13.6%). These results may appear counterintuitive if one considers the likelihood of the retrotransposon insertion interfering with the genes' promoters (since both promoters are in the intergenic region of divergent pairs while neither promoter is there in convergent pairs). However, the fraction of pairs with intergenic insertions correlates quite well with the mean intergenic distance of each pair type, which are 4371 bp, 3734 bp, and 2562 bp for divergent, tandem, and convergent pairs, respectively. Thus there appears to be little selective pressure for or against intergenic retrotransposon insertions based on pair type, and insertion frequency may simply be determined by the space available. The variation in flanking insertion frequency cannot be explained by differences in the size of the intergenic regions flanking each pair type. These regions are more consistent in size among pair types, ranging from 3211 bp on average for divergent pairs to 4114 bp for convergent pairs, while insertion frequency varies greatly, from 6.7% of tandem pairs to 41.1% of divergent pairs. The frequency of this type of insertion may be influenced by the possibility of disrupting regulatory elements, as only divergent pairs have no promoters in the pair's flanking region.

Copia and *Gypsy* elements both belong to the Long Terminal Repeat (LTR) family of retrotransposons, and thus have many structural similarities. These similarities carry over into their possible interactions with gene pairs. We observed substantial reductions in conservation frequency among gene pairs associated with both types of retrotransposons, accompanied by increases in gene pair rearrangement and missing homologs. Due to the strength of the statistical correlation between the presence of *Copia* or *Gypsy* insertions and gene relocation or deletion, it can be assumed that these types of retrotransposons frequently cause or promote the observed changes in gene pair

structure. Of the two, *Gypsy* elements are the more potent agents of change, being associated with more rearranged or deleted genes. This difference between the two classes is especially evident among divergent gene pairs, which are less affected by the presence of *Copia* elements than other pair types, while the effects of *Gypsy* insertions are more evenly distributed among all pairs. One possible explanation for lower conservation rates in LTR-retrotransposons than in non-LTR elements is the role played by LTR sequences in illegitimate recombination. Most models of illegitimate recombination involving LTR-retrotransposons depend upon alignment of the LTR sequences within an element or between LTRs in nearby similar elements (Devos et al., 2002). Illegitimate recombination between elements can also result in the deletion of any sequence between the two retrotransposons as well, providing a mechanism for retrotransposon-mediated deletion of genes. Another possible mechanism of retrotransposon-related gene pair rearrangement is the repair of double-stranded DNA breakage, which can be induced by the presence of transposable elements (Bennetzen, 2005). Depending on the repair mechanism used, these breaks can result in the duplication or deletion of sequence near the break, or the insertion of seemingly unrelated genomic sequence at the breakage point (Puchta, 2004; Salomon and Puchta, 1998). Retrotransposon cDNA sequences have also been found to be inserted during such repairs, so in some cases retrotransposon insertions may be the result of double-stranded break repair, rather than a cause (Puchta, 2005; Moore and Haber, 1996).

LINEs and SINEs differ both from each other and from the LTR-retrotransposons with regard to their correlation with particular events in gene pair evolution. While all four types of insertions reduce the frequency of gene pair conservation, the reductions associated with LINEs and SINEs are much smaller than those of *Copia* and *Gypsy* elements. The effect of LINEs is the weaker of the two, generally only achieving statistical significance when inserted between paired genes. Neither type significantly effects conservation when found in the regions flanking the pair. Rice gene pairs with LINE elements are more likely to be rearranged than deleted (one or both genes), while the opposite is true for SINEs. Frequency of gene inversion is also weakly related to LINE and SINE insertions, while LTR-retrotransposon insertions have little or no effect. Both LINEs and SINEs have been found to cause several types of genomic rearrangements via recombination, although it appears the great majority of such studies have been in animal genomes. Homologous recombination between LINEs has produced deletions in the human genome (Burwinkel and Kilimann, 1998) and segmental duplications in the pig genome (Giuffra et al., 2002). Segmental duplications have also been attributed to SINE-SINE recombination in the human (Jurka et al., 2003) and mouse (Jurka et al., 2005) genomes. While LINEs and SINEs appear to be involved in recombination events similar to those caused by LTR-retrotransposons, the data produced in this study suggests that non-LTR retrotransposons do so less frequently (at least, in

ways involving relocation or deletion of nearby genes) than LTR-retrotransposons. However, without further study it cannot be determined exactly what differences between the LTR and non-LTR retrotransposons identified in this study are responsible for the differential evolutionary behavior of their associated gene pairs.

When comparing the effects of retrotransposon insertions in various locations relative to their associated gene pair, those of insertions flanking the gene pair stand out the most. These insertions are unique in that they appear to significantly affect only divergent gene pairs. This discrepancy is most likely due to flanking retrotransposons being far more common among divergent pairs (found in 41.1% of pairs) than convergent (8.2%) or tandem (6.7%) pairs. Flanking insertions may be most common near divergent pairs because they are always downstream of the nearest gene, and thus less likely to interfere with transcription. Both gene pair rearrangement and gene deletion are more common in divergent pairs flanked by *Copia*, *Gypsy*, and LINE elements. SINEs flanking gene pairs give results quite different than other retrotransposon types, with elevated rates of conservation and lower fractions of pairs with missing homologs among convergent pairs, and slightly higher rates of rearrangement among divergent pairs. The effects of retrotransposon insertions within and between genes are both more profound and widespread than those flanking gene pairs. If we assume that recombination between retrotransposons is responsible for the majority of retrotransposon-mediated gene pair alterations, as described above, then it follows that retrotransposon insertions within the gene pair would be associated with more deletions and rearrangements, as the recombined region between the insertion in the pair and the outside retrotransposon would always include all or part of at least one gene. While they are largely similar, producing substantial decreases in conservation and increases in rearrangement and deletions, intergenic and intragenic insertions vary somewhat in how they affect the different types of gene pair. Generally, retrotransposons between genes are more likely to be associated with disruption of the structure of convergent gene pairs than those within genes, while insertions inside genes have a greater effect on divergent pairs than those between genes. Tandem pairs are affected roughly equally by these two types of insertion.

The three types of gene pair vary somewhat with regard to their overall susceptibility to disruption by retrotransposon insertions. Retrotransposons of all types, when found within or between genes, have the largest negative effect on convergent gene pair conservation and a slightly smaller effect on that of tandem pairs. Divergent pairs are least affected by these types of insertions, usually by a sizeable margin, even though divergent pairs have retrotransposons in their intergenic regions more frequently than any other type. This may be due to differences in intergenic distances among pair types, as divergent pairs generally have larger intergenic regions than convergent pairs (Krom et

al., 2008), allowing for more retrotransposon activity (insertion, deletion, recombination, etc.) without disturbing the surrounding genes. With the exception of SINEs, retrotransposons flanking gene pairs in rice are linked to disrupted divergent pairs in other grass genomes, while convergent and tandem pairs are generally not affected to any significant degree.

4.5 CONCLUSIONS

A sizeable fraction of rice gene pairs are closely associated with retrotransposons, and these pairs are significantly less likely to be conserved in other grass genomes than are rice gene pairs in general. While all types of retrotransposon insertions reduce the probability of conservation, *Copia* and *Gypsy* LTR-retrotransposons do so to a greater degree than LINEs and SINEs, and strongly correlate with both gene pair rearrangement and gene deletion, while the non-LTR types show little association with rearrangement frequency, but are associated with higher rates of gene deletions. Despite being more frequently associated with retrotransposons than convergent or tandem pairs, divergent gene pairs typically show the weakest evolutionary effects from that association, showing the smallest changes in conservation, rearrangement, and deletion rates. In contrast, convergent pairs are the type least frequently associated with retrotransposons yet show the greatest effects of their presence. Insertions between genes in a pair have the most significant effects on gene pair evolution, while flanking insertions significantly affect only divergent pairs. Overall, this study provides valuable insight into how the evolution of gene pair arrangement correlates with the presence of nearby retrotransposons.

4.6 METHODS

Details of gene pair identification and comparative analysis are given in Krom and Ramakrishna (2009). All rice gene pairs employed in that study were scanned for retrotransposon insertions using RepeatMasker (www.repeatmasker.org). For each pair, five sequences were analyzed: the two genes' unspliced genomic sequence, the intergenic region between them, and the two intergenic regions flanking the pair. The evolutionary status (conserved, inverted, rearranged, or deleted) of those pairs containing *Copia* or *Gypsy* LTR-retrotransposons, LINEs, or SINEs within, between, or flanking their genes was then determined via cross-reference with the results of our previous study (Krom and Ramakrishna, 2009). The normal approximation of the binomial test was used to test the statistical significance of the differences in conservation, inversion, rearrangement, or deletion frequency between the complete sets of gene pairs and the sets of retrotransposon-associated pairs. Differences with a p-value less than 0.01 were considered significant.

4.7 LITERATURE CITED

Ammiraju, J.S.S., Lu F., Sanyal A., Yu Y., et al. 2008. Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. Plant Cell 20:3191-3209.

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

Bennetzen, J.L., Chen M. 2008. Grass genomic synteny illuminates plant genome function and evolution. Rice 1:109-118.

Burwinkel, B., and M.W. Kilimann. 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. J Mol Biol 277: 513-517.

Ciccarelli, F. D., C. von Mering, M. Suyama, E. D. Harrington, E. Izaurralde, and P. Bork. 2005. Complex genomic rearrangements lead to novel primate gene function. Genome Res 15: 343-351.

Devos, K.M., J.K.M. Brown, J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12: 1075-1079.

Giuffra, E., A. Tornsten, S. Marklund, E. Boncam-Rudloff, et al. 2002. A large duplication associated with dominant white color in pigs originated by homolgous recombination between LINE elements flanking *KIT*. Mammalian Genome 13: 569-577.

Hurst, L. D., C. Pal, and M. J. Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299–310.

Ilic, K., P. J. SanMiguel, and J. L. Bennetzen. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes. Proc Natl Acad Sci 100: 12265–12270.

Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2003. Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci 101: 1268-1272.

Jurka, J., O. Kohany, A Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2005. Clustering, duplication, and chromosomal distribution of mouse SINE retrotransposons. Cytogenet Genome Res 110: 117-123.

Krom, N., J. Recla, and W. Ramakrishna. 2008. Analysis of genes associated with retrotransposons in the rice genome. Genetica 134: 297-310.

Krom, N., and W. Ramakrishna. 2008. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and *Populus*. Plant Physiol 147: 1763-1773.

Krom, N., and W. Ramakrishna. 2009. Conservation, rearrangement, and deletion of gene pairs in four grass genomes. Manuscript in preparation.

Lai, J., J. Ma, Z. Swigonova, W. Ramakrishna, E. Linton, V. Llaca, B. Tanyolac, Y-J. Park, O. Y. Jeong, J. L. Bennetzen, and J. Messing. 2004. Gene loss and movement in the maize genome. Genome Res 14: 1924-1931.

Ma, J., K.M. Devos, J.L. Bennetzen. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860-869.

Moore, J.K., and J.E. Haber. 1996. Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383: 644-646.

Puchta, H. 2005. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. Journal of Experimental Botany 56: 1-14.

Ren, X.Y., M. Fiers, W.J. Stiekema, and J.P. Nap. 2005. Local coexpression domains of two to four genes in the genome of Arabidopsis. Plant Physiol 138: 923–934.

Salomon, S., and H. Puchta. 1998. Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells.

Vitte, C., O. Panaud, and H. Quesneville. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8: 218.

Williams, E.J.G., and D.J. Bowles. 2004. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res 14: 1060–1067.

Table 1: Retrotransposon Insertions in Rice Gene Pairs							
		Genic		Intergenic		Flanking	
Convergent	*Any Retro*	526	11.0%	655	13.6%	393	8.2%
	Copia LTR	92	1.9%	113	2.4%	113	2.4%
	Gypsy LTR	116	2.4%	242	5.0%	147	3.1%
	LINE	161	3.4%	150	3.1%	72	1.5%
	SINE	210	4.4%	259	5.4%	195	4.1%
Divergent	*Any Retro*	423	11.4%	1174	31.6%	1527	41.1%
	Copia LTR	72	1.9%	192	5.2%	192	5.2%
	Gypsy LTR	101	2.7%	495	13.3%	723	19.5%
	LINE	131	3.5%	223	6.0%	325	8.8%
	SINE	162	4.4%	552	14.9%	637	17.2%
Tandem	*Any Retro*	1089	11.6%	2477	26.3%	627	6.7%
	Copia LTR	183	1.9%	380	4.0%	380	4.0%
	Gypsy LTR	246	2.6%	961	10.2%	242	2.6%
	LINE	329	3.5%	534	5.7%	135	1.4%
	SINE	461	4.9%	1180	12.5%	279	3.0%

Table 2: Gene Pair Rearrangements and Retrotransposons within Genes

Numbers in the columns labelled "Z" are test statistics from the binomial test, comparing the fraction of the various types of retrotransposon-associated gene pairs in each conservation/rearrangement class with the fraction of all gene pairs in the same class. Bold, italic numbers denote a statistically significant difference (P < 0.01).

A) Rice vs. Sorghum			**Conserved**			**Inverted**			**Rearranged**			**Missing Homologs**		
		Total Pairs	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	All Pairs	4800	2740	57.1%		163	3.4%		1208	25.2%		689	14.4%	
	Any Retro.	526	214	40.7%	***-7.60***	16	3.0%	-0.45	172	32.7%	***3.98***	124	23.6%	***6.03***
	Copia LTR	92	30	32.6%	***-4.74***	1	1.1%	-1.22	36	39.1%	***3.09***	25	27.2%	***3.51***
	Gypsy LTR	116	30	25.9%	***-6.79***	3	2.6%	-0.48	48	41.4%	***4.02***	35	30.2%	***4.86***
	LINE	161	74	46.0%	***-2.85***	6	3.7%	0.23	40	24.8%	-0.09	41	25.5%	***4.02***
	SINE	210	96	45.7%	***-3.33***	7	3.3%	-0.05	71	33.8%	***2.89***	36	17.1%	1.15
Divergent	All Pairs	3711	1908	51.4%		97	2.6%		1101	29.7%		605	16.3%	
	Any Retro.	423	173	40.9%	***-4.33***	9	2.1%	-0.63	147	34.8%	2.29	94	22.2%	***3.30***
	Copia LTR	72	26	36.1%	***-2.60***	1	1.4%	-0.65	26	36.1%	1.20	19	26.4%	2.32
	Gypsy LTR	101	31	30.7%	***-4.17***	1	1.0%	-1.02	37	36.6%	1.53	32	31.7%	***4.18***
	LINE	131	56	42.7%	-1.98	4	3.1%	0.32	38	29.0%	-0.17	33	25.2%	***2.75***
	SINE	162	77	47.5%	-0.99	3	1.9%	-0.61	58	35.8%	1.71	24	14.8%	-0.51
Tandem	All Pairs	9428	5636	59.8%		166	1.8%		2112	22.4%		1514	16.1%	
	Any Retro.	1089	564	51.8%	***-5.38***	18	1.7%	-0.27	275	25.3%	2.26	232	21.3%	***4.71***
	Copia LTR	183	80	43.7%	***-4.43***	1	0.5%	-1.25	60	32.8%	***3.37***	42	23.0%	***2.54***
	Gypsy LTR	246	103	41.9%	***-5.73***	5	2.0%	0.32	77	31.3%	***3.35***	61	24.8%	***3.73***
	LINE	329	185	56.2%	-1.31	5	1.5%	-0.33	74	22.5%	0.04	65	19.8%	1.83
	SINE	461	253	54.9%	-2.15	9	2.0%	0.31	110	23.9%	0.75	89	19.3%	1.90

B) Rice vs. Maize			**Conserved**			**Inverted**			**Rearranged**			**Missing Homologs**		
		Total Pairs	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	All Pairs	4800	1686	35.1%		160	3.3%		1962	40.9%		992	20.7%	
	Any Retro.	526	122	23.2%	***-5.73***	18	3.4%	0.11	220	41.8%	0.44	166	31.6%	***6.17***
	Copia LTR	92	20	21.7%	***-2.69***	0	0.0%	-1.78	39	42.4%	0.30	33	35.9%	***3.60***
	Gypsy LTR	116	17	14.7%	***-4.62***	5	4.3%	0.59	50	43.1%	0.49	44	37.9%	***4.59***
	LINE	161	48	29.8%	-1.41	4	2.5%	-0.60	60	37.3%	-0.93	49	30.4%	***3.06***
	SINE	210	48	22.9%	***-3.72***	10	4.8%	1.15	97	46.2%	1.57	55	26.2%	1.98
Divergent	All Pairs	3711	656	17.7%		194	5.2%		2015	54.3%		846	22.8%	
	Any Retro.	423	53	12.5%	***-2.78***	19	4.5%	-0.68	223	52.7%	-0.65	128	30.3%	***3.66***
	Copia LTR	72	7	9.7%	-1.77	6	8.3%	1.18	39	54.2%	-0.02	20	27.8%	1.01
	Gypsy LTR	101	12	11.9%	-1.53	1	1.0%	-1.91	47	46.5%	-1.57	41	40.6%	***4.26***
	LINE	131	15	11.5%	-1.87	5	3.8%	-0.73	65	49.6%	-1.08	46	35.1%	***3.36***
	SINE	162	26	16.0%	-0.54	8	4.9%	-0.17	90	55.6%	0.32	38	23.5%	0.20
Tandem	All Pairs	9428	3146	33.4%		250	2.7%		3799	40.3%		2233	23.7%	
	Any Retro.	1089	323	29.7%	***-2.60***	22	2.0%	-1.30	428	39.3%	-0.67	316	29.0%	***4.14***
	Copia LTR	183	46	25.1%	***-2.36***	3	1.6%	-0.85	71	38.8%	-0.41	63	34.4%	***3.42***
	Gypsy LTR	246	61	24.8%	***-2.85***	4	1.6%	-1.00	99	40.2%	-0.02	82	33.3%	***3.56***
	LINE	329	101	30.7%	-1.03	6	1.8%	-0.93	134	40.7%	0.16	88	26.7%	1.31
	SINE	461	150	32.5%	-0.38	12	2.6%	-0.06	177	38.4%	-0.83	122	26.5%	1.40

C) Rice vs. Brachypodium			**Conserved**			**Inverted**			**Rearranged**			**Missing Homologs**		
		Total Pairs	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**	**#**	**%**	**Z**
Convergent	All Pairs	4800	2670	55.6%		190	4.0%		1226	25.5%		714	14.9%	
	Any Retro.	526	216	41.1%	***-6.72***	18	3.4%	-0.63	166	31.6%	***3.16***	126	24.0%	***5.85***
	Copia LTR	92	31	33.7%	***-4.23***	2	2.2%	-0.88	33	35.9%	2.27	26	28.3%	***3.61***
	Gypsy LTR	116	30	25.9%	***-6.45***	3	2.6%	-0.76	49	42.2%	***4.12***	34	29.3%	***4.37***
	LINE	161	84	52.2%	-0.88	5	3.1%	-0.55	41	25.5%	-0.02	31	19.3%	1.56
	SINE	210	89	42.4%	***-3.86***	8	3.8%	-0.11	68	32.4%	2.27	45	21.4%	***2.67***
Divergent	All Pairs	3711	1855	50.0%		138	3.7%		1079	29.1%		639	17.2%	
	Any Retro.	423	174	41.1%	***-3.64***	11	2.6%	-1.22	139	32.9%	1.71	99	23.4%	***3.37***
	Copia LTR	72	34	47.2%	-0.47	2	2.8%	-0.42	20	27.8%	-0.24	16	22.2%	1.12
	Gypsy LTR	101	26	25.7%	***-4.87***	3	3.0%	-0.40	37	36.6%	1.67	35	34.7%	***4.64***
	LINE	131	55	42.0%	-1.83	2	1.5%	-1.33	39	29.8%	0.18	35	26.7%	***2.88***
	SINE	162	75	46.3%	-0.94	4	2.5%	-0.84	58	35.8%	1.89	25	15.4%	-0.60
Tandem	All Pairs	9428	5456	57.9%		260	2.8%		2148	22.8%		1564	16.6%	
	Any Retro.	1089	546	50.1%	***-5.17***	18	1.7%	-2.23	287	26.4%	***2.81***	238	21.9%	***4.67***
	Copia LTR	183	85	46.4%	***-3.13***	3	1.6%	-0.92	55	30.1%	***2.35***	40	21.9%	1.92
	Gypsy LTR	246	111	45.1%	***-4.05***	4	1.6%	-1.08	72	29.3%	***2.43***	59	24.0%	***3.12***
	LINE	329	172	52.3%	-2.05	5	1.5%	-1.37	78	23.7%	0.40	74	22.5%	***2.88***
	SINE	461	240	52.1%	***-2.53***	8	1.7%	-1.34	120	26.0%	1.66	93	20.2%	2.07

Table 3: Gene Pair Rearrangements and Retrotransposons Between Genes

Numbers in the columns labelled "Z" are test statistics from the binomial test, comparing the fraction of the various types of retrotransposon-associated gene pairs in each conservation/rearrangement class with the fraction of all gene pairs in the same class. Bold, italic numbers denote a statistically significant difference (P < 0.01).

A) Rice vs. Sorghum			Conserved			Inverted			Rearranged			Missing Homologs		
		Total Pairs	#	%	Z	#	%	Z	#	%	Z	#	%	Z
Convergent	All Pairs	4800	2740	57.1%		163	3.4%		1208	25.2%		689	14.4%	
	Any Retro.	655	240	36.6%	***-10.57***	32	4.9%	2.10	230	35.1%	***5.87***	153	23.4%	***6.57***
	Copia LTR	113	38	33.6%	***-5.04***	4	3.5%	0.08	40	35.4%	***2.51***	31	27.4%	***3.97***
	Gypsy LTR	242	66	27.3%	***-9.37***	12	5.0%	1.34	91	37.6%	***4.46***	73	30.2%	***7.02***
	LINE	150	61	40.7%	***-4.06***	8	5.3%	1.31	44	29.3%	1.18	37	24.7%	***3.60***
	SINE	259	111	42.9%	***-4.63***	14	5.4%	1.79	90	34.7%	***3.55***	44	17.0%	1.21
Divergent	All Pairs	3711	1908	51.4%		97	2.6%		1101	29.7%		605	16.3%	
	Any Retro.	1174	572	48.7%	-1.85	36	3.1%	0.97	362	30.8%	0.87	204	17.4%	1.00
	Copia LTR	192	82	42.7%	***-2.41***	7	3.6%	0.90	56	29.2%	-0.15	47	24.5%	***3.07***
	Gypsy LTR	495	217	43.8%	***-3.37***	13	2.6%	0.02	164	33.1%	1.69	101	20.4%	***2.47***
	LINE	223	100	44.8%	-1.96	5	2.2%	-0.35	76	34.1%	1.44	42	18.8%	1.02
	SINE	552	282	51.1%	-0.15	22	4.0%	2.02	169	30.6%	0.49	79	14.3%	-1.27
Tandem	All Pairs	9428	5636	59.8%		166	1.8%		2112	22.4%		1514	16.1%	
	Any Retro.	2477	1342	54.2%	***-5.68***	45	1.8%	0.21	606	24.5%	***2.46***	484	19.5%	***4.72***
	Copia LTR	380	187	49.2%	***-4.20***	5	1.3%	-0.66	106	27.9%	***2.57***	82	21.6%	***2.93***
	Gypsy LTR	961	477	49.6%	***-6.41***	11	1.1%	-1.45	253	26.3%	***2.92***	220	22.9%	***5.77***
	LINE	534	285	53.4%	***-3.02***	12	2.2%	0.85	126	23.6%	0.66	111	20.8%	***2.98***
	SINE	1180	669	56.7%	-2.16	22	1.9%	0.27	286	24.2%	1.51	203	17.2%	1.07

B) Rice vs. Maize			Conserved			Inverted			Rearranged			Missing Homologs		
		Total Pairs	#	%	Z	#	%	Z	#	%	Z	#	%	Z
Convergent	All Pairs	4800	1686	35.1%		160	3.3%		1962	40.9%		992	20.7%	
	Any Retro.	655	112	17.1%	***-9.66***	30	4.6%	1.78	309	47.2%	***3.28***	204	31.1%	***6.62***
	Copia LTR	113	19	16.8%	***-4.08***	3	2.7%	-0.40	53	46.9%	1.30	38	33.6%	***3.40***
	Gypsy LTR	242	26	10.7%	***-7.95***	8	3.3%	-0.02	115	47.5%	2.10	93	38.4%	***6.82***
	LINE	150	30	20.0%	***-3.88***	9	6.0%	1.82	66	44.0%	0.78	45	30.0%	***2.82***
	SINE	259	45	17.4%	***-5.98***	17	6.6%	***2.90***	130	50.2%	***3.05***	67	25.9%	2.07
Divergent	All Pairs	3711	656	17.7%		194	5.2%		2015	54.3%		846	22.8%	
	Any Retro.	1174	168	14.3%	***-3.02***	72	6.1%	1.39	654	55.7%	0.97	280	23.9%	0.86
	Copia LTR	192	21	10.9%	***-2.45***	10	5.2%	-0.01	101	52.6%	-0.47	60	31.3%	***2.79***
	Gypsy LTR	495	56	11.3%	***-3.71***	26	5.3%	0.02	282	57.0%	1.19	131	26.5%	1.95
	LINE	223	27	12.1%	-2.18	20	9.0%	***2.51***	120	53.8%	-0.15	56	25.1%	0.82
	SINE	552	83	15.0%	-1.63	35	6.3%	1.17	323	58.5%	1.99	111	20.1%	-1.51
Tandem	All Pairs	9428	3146	33.4%		250	2.7%		3799	40.3%		2233	23.7%	
	Any Retro.	2477	717	28.9%	***-4.67***	65	2.6%	-0.09	1019	41.1%	0.86	676	27.3%	***4.22***
	Copia LTR	380	104	27.4%	***-2.48***	4	1.1%	-1.94	145	38.2%	-0.85	127	33.4%	***4.46***
	Gypsy LTR	961	274	28.5%	***-3.19***	27	2.8%	0.30	357	37.1%	-1.99	303	31.5%	***5.72***
	LINE	534	144	27.0%	***-3.14***	15	2.8%	0.23	227	42.5%	1.04	148	27.7%	2.19
	SINE	1180	347	29.4%	***-2.89***	32	2.7%	0.13	516	43.7%	***2.40***	285	24.2%	0.38

C) Rice vs. Brachypodium			Conserved			Inverted			Rearranged			Missing Homologs		
		Total Pairs	#	%	Z	#	%	Z	#	%	Z	#	%	Z
Convergent	All Pairs	4800	2670	55.6%		190	4.0%		1226	25.5%		714	14.9%	
	Any Retro.	655	234	35.7%	***-10.25***	34	5.2%	1.62	227	34.7%	***5.35***	160	24.4%	***6.87***
	Copia LTR	113	37	32.7%	***-4.90***	3	2.7%	-0.71	43	38.1%	***3.05***	30	26.5%	***3.49***
	Gypsy LTR	242	66	27.3%	***-8.88***	13	5.4%	1.13	81	33.5%	***2.83***	82	33.9%	***8.31***
	LINE	150	48	32.0%	***-5.82***	11	7.3%	2.12	59	39.3%	***3.87***	32	21.3%	2.22
	SINE	259	112	43.2%	***-4.01***	14	5.4%	1.19	76	29.3%	1.40	57	22.0%	***3.23***
Divergent	All Pairs	3711	1855	50.0%		138	3.7%		1079	29.1%		639	17.2%	
	Any Retro.	1174	537	45.7%	***-2.91***	53	4.5%	1.44	366	31.2%	1.58	218	18.6%	1.23
	Copia LTR	192	78	40.6%	***-2.59***	14	7.3%	***2.62***	52	27.1%	-0.61	48	25.0%	***2.86***
	Gypsy LTR	495	202	40.8%	***-4.08***	21	4.2%	0.62	172	34.7%	***2.78***	100	20.2%	1.76
	LINE	223	93	41.7%	***-2.47***	11	4.9%	0.96	72	32.3%	1.06	47	21.1%	1.53
	SINE	552	273	49.5%	-0.25	23	4.2%	0.56	166	30.1%	0.52	90	16.3%	-0.57
Tandem	All Pairs	9428	5456	57.9%		260	2.8%		2148	22.8%		1564	16.6%	
	Any Retro.	2477	1279	51.6%	***-6.28***	62	2.5%	-0.77	621	25.1%	***2.71***	515	20.8%	***5.62***
	Copia LTR	380	175	46.1%	***-4.67***	8	2.1%	-0.78	93	24.5%	0.79	104	27.4%	***5.65***
	Gypsy LTR	961	470	48.9%	***-5.63***	13	1.4%	***-2.66***	240	25.0%	1.62	238	24.8%	***6.81***
	LINE	534	275	51.5%	***-2.98***	13	2.4%	-0.46	138	25.8%	1.69	108	20.2%	2.26
	SINE	1180	633	53.6%	***-2.94***	35	3.0%	0.44	308	26.1%	***2.72***	204	17.3%	0.65

Table 4: Gene Pair Rearrangements and Retrotransposons Flanking Gene Pairs

Numbers in the columns labelled "Z" are test statistics from the binomial test, comparing the fraction of the various types of retrotransposon-associated gene pairs in each conservation/rearrangement class with the fraction of all gene pairs in the same class. Bold, italic numbers denote a statistically significant difference (P < 0.01).

A) Rice vs. Sorghum			Conserved			Inverted			Rearranged			Missing Homologs		
		Total Pairs	#	%	Z	#	%	Z	#	%	Z	#	%	Z
Convergent	All Pairs	4800	2740	57.1%		163	3.4%		1208	25.2%		689	14.4%	
	Any Retro.	393	236	60.1%	1.19	14	3.6%	0.18	89	22.6%	-1.15	54	13.7%	-0.35
	Copia LTR	113	38	33.6%	***-5.04***	4	3.5%	0.08	40	35.4%	***2.51***	31	27.4%	***3.97***
	Gypsy LTR	147	84	57.1%	0.01	3	2.0%	-0.91	36	24.5%	-0.19	24	16.3%	0.68
	LINE	72	35	48.6%	-1.45	7	9.7%	***2.96***	18	25.0%	-0.03	12	16.7%	0.56
	SINE	195	128	65.6%	***2.41***	6	3.1%	-0.25	42	21.5%	-1.17	19	9.7%	-1.84
Divergent	All Pairs	3711	1908	51.4%		97	2.6%		1101	29.7%		605	16.3%	
	Any Retro.	1527	665	43.5%	***-6.15***	42	2.8%	0.33	515	33.7%	***3.47***	305	20.0%	***3.88***
	Copia LTR	192	82	42.7%	***-2.41***	7	3.6%	0.90	56	29.2%	-0.15	47	24.5%	***3.07***
	Gypsy LTR	723	279	38.6%	***-6.90***	23	3.2%	0.96	245	33.9%	***2.48***	176	24.3%	***5.85***
	LINE	325	152	46.8%	-1.68	6	1.8%	-0.87	95	29.2%	-0.17	72	22.2%	***2.86***
	SINE	637	306	48.0%	-1.71	18	2.8%	0.34	218	34.2%	***2.52***	95	14.9%	-0.95
Tandem	All Pairs	9428	5636	59.8%		166	1.8%		2112	22.4%		1514	16.1%	
	Any Retro.	627	364	58.1%	-0.88	15	2.4%	1.20	143	22.8%	0.24	105	16.7%	0.47
	Copia LTR	380	187	49.2%	***-4.20***	5	1.3%	-0.66	106	27.9%	***2.57***	82	21.6%	***2.93***
	Gypsy LTR	242	136	56.2%	-1.14	4	1.7%	-0.13	57	23.6%	0.43	45	18.6%	1.07
	LINE	135	75	55.6%	-1.00	6	4.4%	***2.37***	31	23.0%	0.16	23	17.0%	0.31
	SINE	279	164	58.8%	-0.34	5	1.8%	0.04	67	24.0%	0.65	43	15.4%	-0.29

B) Rice vs. Maize			Conserved			Inverted			Rearranged			Missing Homologs		
		Total Pairs	#	%	Z	#	%	Z	#	%	Z	#	%	Z
Convergent	All Pairs	4800	1686	35.1%		160	3.3%		1962	40.9%		992	20.7%	
	Any Retro.	393	142	36.1%	0.42	14	3.6%	0.25	171	43.5%	1.06	66	16.8%	-1.90
	Copia LTR	113	19	16.8%	***-4.08***	3	2.7%	-0.40	53	46.9%	1.30	38	33.6%	***3.40***
	Gypsy LTR	147	44	29.9%	-1.32	2	1.4%	-1.33	71	48.3%	1.83	30	20.4%	-0.08
	LINE	72	27	37.5%	0.42	4	5.6%	1.05	26	36.1%	-0.82	15	20.8%	0.03
	SINE	195	80	41.0%	1.73	7	3.6%	0.20	83	42.6%	0.48	25	12.8%	***-2.71***
Divergent	All Pairs	3711	656	17.7%		194	5.2%		2015	54.3%		846	22.8%	
	Any Retro.	1527	235	15.4%	***-2.34***	63	4.1%	-1.93	821	53.8%	-0.42	408	26.7%	***3.65***
	Copia LTR	192	21	10.9%	***-2.45***	10	5.2%	-0.01	101	52.6%	-0.47	60	31.3%	***2.79***
	Gypsy LTR	723	105	14.5%	-2.22	26	3.6%	-1.97	361	49.9%	***-2.36***	231	32.0%	***5.87***
	LINE	325	58	17.8%	0.08	10	3.1%	-1.74	175	53.8%	-0.16	82	25.2%	1.05
	SINE	637	102	16.0%	-1.10	38	6.0%	0.84	359	56.4%	1.04	138	21.7%	-0.68
Tandem	All Pairs	9428	3146	33.4%		250	2.7%		3799	40.3%		2233	23.7%	
	Any Retro.	627	212	33.8%	0.24	24	3.8%	1.83	246	39.2%	-0.54	145	23.1%	-0.33
	Copia LTR	380	104	27.4%	***-2.48***	4	1.1%	-1.94	145	38.2%	-0.85	127	33.4%	***4.46***
	Gypsy LTR	242	79	32.6%	-0.24	6	2.5%	-0.17	101	41.7%	0.46	56	23.1%	-0.20
	LINE	135	52	38.5%	1.27	4	3.0%	0.23	49	36.3%	-0.95	30	22.2%	-0.40
	SINE	279	89	31.9%	-0.52	12	4.3%	1.71	110	39.4%	-0.30	68	24.4%	0.27

C) Rice vs. Brachypodium			Conserved			Inverted			Rearranged			Missing Homologs		
		Total Pairs	#	%	Z	#	%	Z	#	%	Z	#	%	Z
Convergent	All Pairs	4800	2670	55.6%		190	4.0%		1226	25.5%		714	14.9%	
	Any Retro.	393	232	59.0%	1.36	14	3.6%	-0.40	92	23.4%	-0.97	55	14.0%	-0.49
	Copia LTR	113	37	32.7%	***-4.90***	3	2.7%	-0.71	43	38.1%	***3.05***	30	26.5%	***3.49***
	Gypsy LTR	147	74	50.3%	-1.29	5	3.4%	-0.35	39	26.5%	0.27	29	19.7%	1.65
	LINE	72	34	47.2%	-1.44	5	6.9%	1.30	21	29.2%	0.71	12	16.7%	0.43
	SINE	195	129	66.2%	***2.96***	5	2.6%	-1.00	42	21.5%	-1.28	19	9.7%	-2.01
Divergent	All Pairs	3711	1855	50.0%		138	3.7%		1079	29.1%		639	17.2%	
	Any Retro.	1527	657	43.0%	***-5.44***	58	3.8%	0.16	496	32.5%	***2.93***	316	20.7%	***3.60***
	Copia LTR	192	78	40.6%	***-2.59***	14	7.3%	***2.62***	52	27.1%	-0.61	48	25.0%	***2.86***
	Gypsy LTR	723	272	37.6%	***-6.65***	27	3.7%	0.02	236	32.6%	2.11	188	26.0%	***6.26***
	LINE	325	154	47.4%	-0.94	10	3.1%	-0.61	85	26.2%	-1.16	76	23.4%	***2.94***
	SINE	637	296	46.5%	-1.78	22	3.5%	-0.35	223	35.0%	***3.30***	96	15.1%	-1.44
Tandem	All Pairs	9428	5456	57.9%		260	2.8%		2148	22.8%		1564	16.6%	
	Any Retro.	627	352	56.1%	-0.88	14	2.2%	-0.80	140	22.3%	-0.27	121	19.3%	1.82
	Copia LTR	380	175	46.1%	***-4.67***	8	2.1%	-0.78	93	24.5%	0.79	104	27.4%	***5.65***
	Gypsy LTR	242	131	54.1%	-1.18	5	2.1%	-0.66	56	23.1%	0.13	50	20.7%	1.70
	LINE	135	75	55.6%	-0.54	5	3.7%	0.67	29	21.5%	-0.36	26	19.3%	0.83
	SINE	279	159	57.0%	-0.30	6	2.2%	-0.62	65	23.3%	0.20	49	17.6%	0.44

CONCLUSION

In the preceding studies, we described various properties of gene pairs, gene pair rearrangement, and retrotransposon insertions in several plant genomes. We also determined several ways in which they interact, and the significance of that interaction. Each of these studies uncovered a number of mechanisms through which their studies impact the overall scheme of gene expression, genome organization, and genome evolution.

Given their prevalence in plant genomes, it is unsurprising that retrotransposons were found to affect genome function and structure in many ways. Insertion upstream of genes tends to preclude expression, presumably due to disruption of regulatory elements in the promoter, and this effect weakens with greater distance from the gene's translation start site. When found inside genes retrotransposon insertions increase the probability of that gene having alternative splicing models through the introduction of novel splice sites and recruitment of transposon sequence as exons. Retrotransposon insertions do not appear to be randomly distributed in the rice genome, with the distribution of retrotransposon-associated genes across chromosomes significantly diverging from that of genes in general. Some Gene Ontology classes are over- and under-represented among retrotransposon-associated genes. There is also significant variation in retrotransposon insertion in different gene pair types, being least commonly associated with convergent pairs and found substantially more often in and around divergent pairs. The presence of retrotransposons within gene pairs reduces the probability of conserved pair arrangement in other species by promoting gene relocation and deletion. We also noted some differences in behavior among different families of retrotransposons. *Copia* and *Gypsy* LTR-retrotransposons appear to interfere most with their surroundings, more frequently precluding expression of nearby genes and reducing gene pair conservation to a greater extent than non-LTR retrotransposons. LINE and SINE insertions in genes increase the probability of alternative splicing, while *Copia* and *Gypsy* insertions reduce it.

The importance of gene pair arrangement with regard to gene regulation and expression appears to be fairly modest. While strongly correlated expression between adjacent genes is quite common, it is not significantly more common in any one type of pair, despite the major differences among them for the potential sharing of regulatory elements, such as bidirectional promoters controlling both genes of a divergent pair. Both strongly correlated expression and shared Gene Ontology classification increase the probability of conservation, but conservation across long evolutionary distances (such as between rice and Arabidopsis) remains quite rare. Conservation within the grasses is substantially higher overall, and correlated expression tends to increase conservation rates. Overall, correlated expression due to regulatory mechanisms dependent on a specific arrangement, rather than simply being adjacent, appears to be quite rare. Gene

pair conservation and rearrangement varies somewhat among pair types. Divergent pairs are consistently the type least often conserved, although the margin in usually fairly small. The exception to this is in maize, where divergent pairs are conserved almost half as frequently as convergent and tandem pairs. The effect of retrotransposon insertions on conservation and rearrangement also affects pair types differently. Convergent pairs show the greatest decrease in conservation when they contain retrotransposons, while the decrease among divergent pairs is considerably smaller. Divergent pairs are also the only pair type that is significantly less often conserved when flanked by retrotransposons.

Overall, the investigations described here made several significant discoveries about some of the forces at work within plant genomes. In addition, the data produced in these analyses can serve as useful guides in further studies, both experimental and bioinformatic, of these phenomena. It is our hope that these humble writings may be judged a valuable addition to the body of human knowledge.

FUTURE WORK

The studies described previously were all genome-wide surveys of specific features and phenomena. As such, there is a practical limit on the number of topics each study can address and the detail in which they may be investigated while remaining focused and relatively straightforward to interpret. Therefore, there remain a large number of potential investigations, both bioinformatic and wet-lab, that would serve to further elucidate the significance and inner workings of many of the subjects described here.

Retrotransposon insertions in and near genes are quite common in rice, and appear to affect gene expression in several ways. In order to be certain that the observed differences in expression among retrotransposon-associated genes are caused by the retrotransposon insertions, transgenic rice plants could be created, with the retrotransposon sequence in a specific gene deleted. Expression of the modified gene would be monitored to detect any changes, such as reactivation of a previously inactive gene. In addition, this approach could identify quantitative changes in expression levels due to retrotransposon insertions, rather than being limited to the binary expressed/non-expressed results described earlier. Promoters containing retrotransposon insertions could be used to drive expression of reporter genes, in both their original forms and with the retrotransposons deleted, in order to determine if such genes that are inactive in rice were inactivated by the retrotransposon insertion.

Identification of retrotransposon-associated genes in other species, in a manner similar to the study described in chapter 1, could also prove illuminating by highlighting interspecies variation in the prevalence of such genes. Such a study could also involve the dating of retrotransposon insertions, to see if older insertions, which will presumably remain as smaller fragments than more recent ones, have similar effects on gene expression. An investigation into retrotransposon-associated genes in maize would be of particular interest, due to the massive amplification of retrotransposons that took place quite recently (~5 mya) in its evolutionary history.

Our studies of gene pair expression and conservation may also be extended in several ways. To further investigate trends in small-scale changes in gene order, the status of genes flanking previously identified gene pairs could be determined. This would identify the types of rearrangements that prevent gene pair conservation, such as how often they involve either single or multiple genes. This data could also be used to examine interaction between larger groups of genes, both in terms of gene order and coexpression. Also of interest would be an updated expression analysis, especially in *Populus*, which had far less expression data available than rice or Arabidopsis at the time that analysis was performed. In addition to more complete expression data, an expanded

study would also benefit from more consistent data across the species being compared. If all the expression data were derived from the same platform, tissues, procedures, etc., comparisons of the frequency of coexpression among different species would be more valid and informative than is possible with the data currently presented.

Another potential area of investigation is the different mechanisms involved in producing the widespread coexpression observed among all types of gene pairs. Regulatory elements have been identified that are more frequently found in coexpressed divergent gene pairs, but no such elements were found to be associated with coexpressed convergent pairs, and coexpressed tandem pairs have not yet been examined for over-represented elements. As specific regulatory elements do not appear to be more commonly found in coexpressed convergent gene pairs, other potential causes of coexpression must be investigated, such as over-represented combinations of regulatory elements, DNA or histone methylation, and chromatin states.

More detailed studies of the over-represented regulatory elements in the intergenic regions of coexpressed divergent gene pairs could also be performed in order to determine if they are present in greater numbers due to the presence of two separate but very similar promoters or if they are unique to bidirectionally active promoters that control both genes. Such studies could be either bioinformatic or experimental. A bioinformatic analysis would analyze the distribution of regulatory elements within the intergenic region and thus determine the most likely promoter scenario. An experimental investigation could involve the use of the intergenic region to drive expression of two different reporter genes and the creation of deletion constructs missing various segments of the intergenic region to determine if there exists a single segment upon which the expression of both genes depends, suggesting that the pair being examined is controlled by a single bidirectional promoter.

There are several aspects of the association between retrotransposon insertions and gene pair rearrangements and deletions that would benefit from further investigation. To better support the hypothesis that retrotransposons are the cause of gene pair rearrangements, the regions near homologs of rearranged pairs could also be analyzed for retrotransposons, as could the surrounding regions of the lone identified homolog from pairs believed to have undergone deletion of a single gene. This would ensure that the retrotransposons identified here are in fact shared by both genomes being compared. This approach could be combined with estimating the date of insertion for the LTR retrotransposons in question, which would determine if the retrotransposon was present when the two species diverged.

Recombination between similar retrotransposons may be a cause of gene pair rearrangement and deletion. To determine if such recombination is possible for a given

gene pair, the presence of retrotransposons in multiple regions could be investigated. For instance, a retrotransposon flanking a gene pair could recombine with one in the intergenic region to delete or relocate the gene that they surround. Likewise, if a gene pair has retrotransposons in both flanking regions, both genes could conceivably be deleted via recombination. Such an investigation would help identify the mechanism by which retrotransposon-associated gene pairs are rearranged or deleted.

The research projects described above include both genome-wide studies of additional features and in-depth studies of specific genes or gene pairs. The fact that so many widely varied projects could be performed based on the data produced in the four studies described here highlights the value of bioinformatics-based genome-wide studies. We believe that any such follow-up projects could easily have as great an impact as the studies they were based on, and would comprise a worthy legacy for our work.

APPENDIX

Chapter 1 of this dissertation has been previously published separately in the journal Genetica. The author has obtained permission from the publisher to reprint this material, as shown below:

SPRINGER LICENSE
TERMS AND CONDITIONS

Aug 20, 2009

This is a License Agreement between Nicholas D Krom ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2158940567499
License date	Mar 30, 2009
Licensed content publisher	Springer
Licensed content publication	Genetica
Licensed content title	Analysis of genes associated with retrotransposons in the rice genome
Licensed content author	Nicholas Krom
Licensed content date	Nov 1, 2008
Volume number	134
Issue number	3
Pages	297 - 310
Type of Use	Thesis / Dissertation
Details of use	Print
Requestor Type	Individual
Portion of the article	Full text
Title of your thesis / dissertation	Conservation and Arrangement of Gene Pairs and Regulating Gene Expression in Plant Genomes
Expected completion date	Aug 2009
Billing Type	Invoice
Company	Nicholas D Krom
Billing Address	MTU Department of Biological Sciences 1400 Townsend Dr. Houghton, MI 49931 United States
Customer reference info	
Total	0.00 USD

Terms and Conditions

Introduction
The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License
With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry. Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided it is password protected or on the university's intranet, destined to microfilming by UMI and University repository. For any other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, provided permission is also obtained from the (co) author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well). Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Altering/Modifying Material: Not Permitted
However figures and illustrations may be altered minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

Reservation of Rights
Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

Chapter 2 of this dissertation has been previously published separately in the journal Plant Physiology. The journal's copyright policy allows reuse of published materials by the authors as stated below:

To Our Authors:

ASPB grants to authors whose work has been published in *Plant Physiology*® or *The Plant Cell* the royalty-free right to reuse images, portions of an article, or full articles in any book, book chapter, or journal article of which the author is the author or editor. Reproductions must bear the full citation, the journal URL (www.plantphysiol.org or www.plantcell.org), and the following notice: "Copyright American Society of Plant Biologists." ASPB further grants to authors the permission to make digital or hard copies of part or all of a work published in *Plant Physiology*® or *The Plant Cell* without fee for personal or classroom use.