Insights into the rice and Arabidopsis genomes: intron fates, paralogs, and lineage-specific genes

by

Haining Lin

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
C. Robin Buell, Co-major Professor
Xun Gu, Co-major Professor
Daniel F. Voytas
Hui-Hsien Chou
Zhijun Wu

Iowa State University

Ames, Iowa

2009

Copyright © Haining Lin, 2009. All rights reserved

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am sincerely grateful to many people who have helped me reach this milestone in my life. I am glad to take this opportunity to express my gratitude for them.

I am deeply indebted to my major professor, Dr. Robin Buell, without whom I would not have been able to complete this dissertation. Her enthusiasm and vision for science has inspired me throughout the course of my research and studies. I would also like to thank Dr. Xun Gu for his invaluable guidance, discussion, and support during my PhD studies. My special thanks go to Dr. Dan Voytas, Dr. Hui-Hsien Chou, and Dr. Zhijun Wu for their encouragement, advice, support, and help. I am very fortunate to have these five distinguished professors on my committee and will be deeply indebted to all of their efforts.

I would like to thank Dr. Wei Zhu, Dr. Joana Silva, Dr. Shu Ouyang, Amy Egan, Dr. Blake Meyers, Dr. Kan Nobuta, Brian Haas, Dr. Shin-han Shiu and Gaurav Moghe for their discussion and help. Also, I would like to thank members in Buell lab for their kindness, friendship, and encouragement.

I would like to thank Bioinformatics and Computational Biology program, especially Dr. Drena Dobbs, Dr. Vasant Honavar, Dr. Srinivas Aluru, and Dr. Christopher Tuggle for their support. I would like to thank Kathy Wiederin and Trish Stauble for helping with the administrative issues.

Finally, my deepest gratitude goes to my family. Words are not enough. I am sincerely grateful to my parents, Jiahan Lin and Liqin Wu, my parents-in-law, Qisheng Huang and Hanying Guo, my husband, Xuanning Huang, and my lovely kids, Amy and Ryan, for their love, encouragement, and support in my life.

CHAPTER 1. GENERAL INTRODUCTION

1. Introduction

Arabidopsis thaliana and *Oryza sativa*, model species for dicotyledonous and monocotyledonous, respectively, were the first two plant species in which a genome sequence was made available (Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005). As they represent two major divisions in the angiosperms, comparative genome analyses between Arabidopsis and rice, and to other plant species, can shed light on the evolution of angiosperms. The availability of a wealth of genomic, transcriptomic, and proteomic data, along with high quality gene annotation of both genomes, provides an opportunity to perform large-scale computational analyses to gain insights into the composition, arrangement, and evolution of plant genomes.

2. Dissertation organization

This dissertation consists of five chapters which are focused on genomic analyses of rice and Arabidopsis including the evolution of introns in the rice genome, paralogous protein families of rice and Arabidopsis, and lineage-specific genes in Arabidopsis.

Chapter 2 is a manuscript published in Genome Biology in 2006. In the manuscript, we first identified segmental duplication events in the rice genome. We then characterized intron gain and loss events in the rice genome with respect to segmental duplication events. Haining Lin designed the study, performed the computational analysis, and drafted the paper. Dr. Wei Zhu assisted in the design of the study. Dr. Joana Silva determined the number of synonymous substitutions per synonymous site for duplicated rice genes. Dr. Xun Gu

assisted in the design of the study. Dr. Robin Buell supervised the project and edited the paper.

Chapter 3 is a manuscript published in BMC Plant Biology in 2008. In the manuscript, we classified paralogous protein families in rice and Arabidopsis using a computational pipeline that utilized both Pfam and novel BLASTP-based domains. We characterized alternative splicing, functional classification of paralogous family proteins, expression patterns, and duplication age, and compared these data to those observed in single copy proteins. A parallel analysis of alternative splicing and functional domain composition of paralogous family proteins was performed with Arabidopsis to compare and contrast with the findings in rice. Haining Lin designed the study, performed the analyses, and drafted the paper. Dr. Shu Ouyang participated in the analysis of GOSlim and made Additional file 3. Drs. Kan Nobuta and Blake Meyers provided rice massively parallel signature sequencing data. Amy Egan and Dr. Joana Silva carried out the age analysis of paralogous families. Brian Haas identified alternative splicing isoforms in rice. Dr. Wei Zhu identified the high confidence gene set in rice. Dr. Xun Gu assisted in the analysis of alternative splicing. Dr. Robin Buell designed the study and drafted the manuscript.

Chapter 4 is a manuscript to be submitted to BMC Evolutionary Biology. In the manuscript, Arabidopsis lineage-specific genes and conserved Brassica-specific genes were identified and characterized. Haining Lin designed the study, conducted the majority of the computational analyses, and drafted the paper. Gaurav Moghe and Dr. Shin-han Shiu carried out the expression analysis. Dr. Shu Ouyang generated the Additional data file 1. Dr. Xun Gu supervised the analysis of single nucleotide polymorphisms and the study. Dr. Robin Buell designed the study, supervised the study, and drafted the paper.

Chapter 5 summarizes the findings in the dissertation and point out possible future research directions.

3. Literature review

This dissertation concentrates on genomic analysis of rice and Arabidopsis, which includes evolution of introns, paralogous protein families, and lineage-specific genes. The first part of the literature review introduces genomic resources of flowering plants; the second part focuses on rice genome organization and evolution; the last part focuses on lineage-specific genes.

3.1. Genomes of flowering plants

3.1.1. Arabidopsis as a model species

Arabidopsis thaliana, a member of the Brassicaceae family, is the premiere model organism for laboratory experiments in genetics and molecular biology of flowering plants. Although not an economically important plant, Arabidopsis has been intensively studied for over 50 years and has accumulated tremendous valuable genetic and molecular biology data and resources due to its small genome size, ease of growth, short life cycle, and facile transformation ability.

Arabidopsis was the first plant genome that was sequenced and it was completed in 2000 using a bacterial artificial chromosome (BAC) BAC-by-BAC sequencing strategy, covering 115.4 megabases (Mb) of the 125-Mb genome (Arabidopsis Genome Initiative 2000). Initially, a total of 25,498 protein-coding genes were predicted and analyzed in the Arabidopsis genome (Arabidopsis Genome Initiative 2000). Over the years, the annotation of the Arabidopsis genome has been updated by improvements of methods, tools, protocols (Haas, Delcher et al. 2003; Wortman, Haas et al. 2003; Haas, Wortman et al. 2005), and by

incorporation of community submitted annotations (The Arabidopsis Information Resource). The Arabidopsis Information Resource (TAIR) has provided high quality structural and functional annotation data along with metabolism, gene expression, seed stocks, and genetic and physical markers (The Arabidopsis Information Resource).

3.1.2. Rice as a model species

Oryza sativa, cultivated rice, is considered one of the most important crops in the world and is a major food and nutrition source in many countries, especially in the developing world. Rice serves as a model species for cereals and monocot species such as maize (*Zea mays*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and sorghum (*Sorghum bicolor*). As a monocotyledonous species within the angiosperms, rice provides a platform to perform comparative genomic analyses with the dicotyledonous model species, Arabidopsis, to shed light into the evolution of angiosperms.

Rice has a small genome whose sequence is available through multiple sequencing projects utilizing either a BAC-by-BAC (Sasaki and Burr 2000; Barry 2001) or whole genome shotgun sequencing approach (Goff, Ricke et al. 2002; Yu, Hu et al. 2002). In 2005, the map-based, finished quality sequence that covers 95% of the 389 Mb of a japonica subspecies of *O. sativa* was reported by the International Rice Genome Sequencing Project (International Rice Genome Sequencing Project 2005). Functional annotation of the rice genome is available through the Osa1 Rice Genome Annotation project which has provided high quality structural and functional annotation as well as comprehensive analyses on expression data, gene ontologies, flanking sequence tags, alternative spliced genes, repeats, and synteny (Ouyang, Zhu et al. 2007).

3.1.3. Sequenced plant genomes

In addition to the genomes of rice and Arabidopsis, the genomes of *Populus trichocarpa* (poplar), *Vitis vinifera* (grapevine), *Physcomitrella patens* (moss), *Carica papaya* (papaya), have been released. Poplar is a model species for forest trees, which cover more than one quarter of the earth's surface. The draft sequence of the genome of *P. trichocarpa* was released and 45,555 protein-coding genes were identified (Tuskan, Difazio et al. 2006). A high-quality draft sequence of the grapevine genome from a close to fully homozygous genotype, representing 8.4-fold coverage of the genome by a whole-genome shotgun sequencing strategy was recently released (Jaillon, Aury et al. 2007); a total of 30,434 protein-coding genes have been predicted in grapevine (Jaillon, Aury et al. 2007). A high-quality draft sequence of the genome of moss was determined by a whole-genome shotgun sequencing strategy with a depth of 8.6-fold coverage; a total of 35,938 predicted and annotated gene models were identified (Rensing, Lang et al. 2008). The draft sequence of the genome of 'SunUp' papaya, the virus-resistant transgenic tropical fruit tree, representing 3-fold of the genome by a whole-genome shotgun sequencing strategy, was released, and 28,629 gene models were identified (Ming, Hou et al. 2008).

3.2. Rice genome organization and evolution

3.2.1. Homologs, paralogs, and orthologs

Homologs refer to genes that are derived from a common ancestral sequence, usually identified by sequence similarity analysis. There are two fundamental types of homologs: paralogs and orthologs. Paralogs are genes within a species which are derived from duplication events while orthologs are genes in different species which are separated by a speciation event (Fitch 1970).

Paralogs from recent duplications can have similar functions while paralogs from ancient duplications may have more divergent function due to relaxed selection pressure. Several models have been proposed for the diverging fates of duplicated genes. In the non/neo-functionalization, one of the duplicated genes retains the original function while the other becomes a functionless pseudogene through accumulation of deleterious mutations although on a rare occasion it may acquire a novel function under relaxed selection constraint (Ohno 1970). In the sub-functionalization model (Hughes 1994; Force, Lynch et al. 1999; Lynch and Force 2000), both duplicated genes adopt a subset of the function of the ancestral gene such that the duplicated genes together retain the original function. Alternatively, duplicated genes may retain redundant function to increase the robustness of biological systems (Gu, Steinmetz et al. 2003).

On the other hand, orthologs are likely to retain the same function over the evolutionary time and thus are used to facilitate the functional inference of genes within a genome (Eisen 1998). Identification of orthologous groups is critical for large scale automated functional annotation of newly sequenced genomes as orthologs can confirm the existence of genes with no known function within a genome thereby providing new and informative annotation on the validity of the gene and its potential function.

As the evolutionary relationships become complicated after several rounds of duplication and speciation, more specialized terms such as out-paralog, in-paralog, and co-ortholog have been introduced to accurately describe the relationship of paralogs and orthologs (Sonnhammer and Koonin 2002). Both out-paralog and in-paralogs are subtypes of paralog that separate before or after a given speciation, respectively (Sonnhammer and Koonin 2002). Out-paralogs are paralogs from a duplication event(s) before the speciation

event of interest. In contrast, in-paralogs are genes within one species that are derived from a species-specific duplication event(s) subsequent to the speciation of interest. Co-orthologs refers to in-paralogs in different species that are orthologs to each other (many-to-many or many-to-one) due to a speciation event followed by a lineage-specific duplication.

3.2.2. Gene duplication of rice

Rice is an ancient polyploidy that underwent two major rounds of large-scale segmental duplication events. Depending on the methods, parameters, and genome assemblies utilized, 15 ~ 62% of the rice genome (Vandepoele, Simillion et al. 2003; Guyot and Keller 2004; Paterson, Bowers et al. 2004; Simillion, Vandepoele et al. 2004; Wang, Shi et al. 2005) result from duplication events that occurred at ~70 (Vandepoele, Simillion et al. 2003; Paterson, Bowers et al. 2004; Wang, Shi et al. 2005) and 5 ~ 8 Million Years Ago (MYA) (The Rice Chromosomes 11 and 12 Sequencing Consortia 2005; Wang, Shi et al. 2005). Large-scale segmental duplication, coupled with other genetic events such as tandem duplication, has resulted in a substantial number of paralogous protein families in rice. Although paralogous protein families in rice have been studied extensively in rice, they mostly focus on a specific gene family such as stressed associated protein gene family (Vij and Tyagi 2006), IQD gene family (Abel, Savchenko et al. 2005), and protein kinase gene family (Ito, Takaya et al. 2005). The first systematic whole-genome analysis of paralogous protein families in rice was published by Horan and the colleagues (Horan, Lauricha et al. 2005), in which rice proteins were co-clustered with Arabidopsis proteins using either Pfam domain-based or BLASTP-based similarity clustering.

3.3. Intron evolution

Introns are DNA regions that are not translated into proteins, and thus are under less selection pressure and diverge faster than exons. Two controversial theories regarding the origin of introns have been competing each other since introns were discovered. The introns-early theory (IE) proposes that introns were extremely old and existed before any eukaryote-prokaryote divergence (Gilbert 1978; Gilbert 1987; Logsdon, Tyshenko et al. 1995; de Souza, Long et al. 1998; Fedorov, Cao et al. 2001; Roy and Gilbert 2005) and that intron loss has dominated intron evolution during which prokaryotes completely lost their introns (Roy and Gilbert 2005). The introns-late theory (IL) proposes that introns were inserted after eukaryote-prokaryote divergence (Cavalier-Smith 1985; Cavalier-Smith 1991; Sharp 1991; Logsdon, Tyshenko et al. 1995; Rzhetsky, Ayala et al. 1997), probably originated from type II self-splicing introns transferred from an organelle (Cavalier-Smith 1991) or from transposable elements (Giroux, Clancy et al. 1994; Iwamoto, Mackawwa et al. 1998).

Introns can be classified into three phases based on their location relative to the codon. Phase 0 introns do not interrupt the codon, i.e. they are located between the third base of the codon and the first base of the following codon. Phase 1 introns are located between the first and second bases of the codon while phase 2 introns are located between the second and third bases of the codon. Non-uniformity in intron phase distribution has been observed in animals, plants, fungi, insects, and protists, with phase 0 being the most abundant and phase 2 being the least abundant (Rogers 1990; Fedorov, Suboch et al. 1992; Long, de Souza et al. 1995; Tomita, Shimizu et al. 1996); on average a 5:3:2 ratio of phases 0, 1, 2 is observed although slight variation exists among taxonomical groups (Rogers 1990; Fedorov, Suboch et al. 1992; Tomita, Shimizu et al. 1996). This unequal distribution of intron phase has been a strong argument for the process of exon shuffling (Fedorov, Suboch et al. 1992;

Long, de Souza et al. 1995). More recent studies on ten protein families of eukaryotic protein-coding genes showed that intron phase bias is predominantly the result of a phase preference of intron gain (Qiu, Schisler et al. 2004).

Intron position with respect to the protein sequence is also relatively conserved. Conservation of the position of introns has been documented between distinct eukaryotic lineages such as between animal and fungal genes (Fedorov, Merican et al. 2002) and between *Plasmodium falciparum* and other eukaryotes (Rogozin, Wolf et al. 2003). A correlation of intron density and positional bias in coding region has been reported from studies on introns from seven eukaryotes (Sakurai, Fujimori et al. 2002). Introns have a tendency to be located towards the 5' end of the gene in intron-sparse organisms or intron-sparse genes. Typical examples of such genomes are *Saccharomyces cerevisiae* and *P. falciparum*. In principle, such biased pattern of intron distribution could be a result of a combination of intron gain and loss, two ongoing processes in intron evolution. Several studies have demonstrated the occurrence of intron gain and loss among species with various rates of intron gain and loss among species (Rogozin, Wolf et al. 2003; Roy, Fedorov et al. 2003; Coghlan and Wolfe 2004; Nielsen, Friedman et al. 2004; Roy and Gilbert 2005).

In the Plant Kingdom, large-scale genomic analyses of intron gain and loss have been focused on Arabidopsis (Sakurai, Fujimori et al. 2002; Fedorov, Roy et al. 2003; Rogozin, Wolf et al. 2003; Sverdlov, Babenko et al. 2004; Roy and Gilbert 2005; Roy and Gilbert 2005; Knowles and McLysaght 2006). With the availability of high quality sequence and gene annotation in rice (Ouyang, Zhu et al. 2007), it would be interesting to determine how introns evolve in monocot compared to a dicot.

3.4. Lineage-specific genes

Lineage-specific genes are genes that are restricted to a limited subset of related species or even a single species. It has been proposed that lineage-specific genes evolve rapidly, and as a consequence, no significant sequence similarity to genes from other lineages can be detected (Schmid and Aquadro 2001; Cai, Woo et al. 2006). Previous analyses showed that lineage-specific genes are on average smaller compared to genes with significant sequence similarity to a wide range of species (Charlebois, Clarke et al. 2003; Daubin and Ochman 2004; Campbell, Zhu et al. 2007).

Several hypotheses regarding the origin of lineage-specific genes have been proposed. One model suggests that lateral gene transfer plays an important role in generating lineage-specific genes that are not shared by closely related species (Daubin, Lerat et al. 2003; Striepen, Pruijssers et al. 2004). The second model proposes that lineage-specific genes may be generated by gene duplication followed by rapid sequence divergence (Domazet-Loso and Tautz 2003; Alba and Castresana 2005). Other models include *de novo* emergence from non-coding sequences which are more diverged between species (Levine, Jones et al. 2006), differential gene loss of duplicated genes (Blomme, Vandepoele et al. 2006), and possible artifacts from genome annotation (Schmid and Aquadro 2001). Although the evolutionary origin and functionary significance of lineage-specific genes remain unclear, the identification of putative lineage-specific genes provides insights into evolutionary processes such as adaptation and generation of diversity.

Identification of lineage-specific genes generally relies on the availability of complete or near-complete genome and/or transcribed sequences (Allen 2002). The percentage of lineage-specific genes within a species may reduce when more transcribed sequences and genomic sequences become available. For example, re-computing the number of lineage-

specific genes from eight microbial genomes revealed that 0 ~ 30% rather than the original reported 16 ~ 56% of the genes in the genome were lineage-specific genes (Fischer and Eisenberg 1999). Nevertheless, a systematic analysis from 60 fully sequenced microbial genomes showed that the number of singleton open reading frames which have no detectable sequence similarity to any other sequences grew with the increasing size of available sequence database although the percentage has diminished to 14% (Siew and Fischer 2003).

Early identification and characterization of lineage-specific genes focused on microbial species mainly for two reasons. Firstly, initially there were more genomic sequences available for microbial compared to other species. Secondly, studies on lineage-specific genes may result in discovery of putative targets for vaccine developments. With the availability of more complete genomes and transcriptomes across a wide range of taxa, lineage-specific genes have been extensively studied across a wide range of species (Boffelli, McAuliffe et al. 2003; Domazet-Loso and Tautz 2003; Nahon 2003; Fortna, Kim et al. 2004; Graham, Silverstein et al. 2004).

Within the Plant Kingdom, lineage-specific genes have been identified and characterized in several species by comparative analyses of the wide availability of Expressed Sequence Tags (ESTs) and the finished genomes of rice and Arabidopsis. For example, a comparative approach screening six solanaceous transcriptomes against the predicted proteomes of rice and Arabidopsis and 21 plant gene indices revealed that approximately one fifth of the Solanaceae unique transcripts are likely Solanaceae-specific sequences (Rensink, Lee et al. 2005). A total of 2,525 putative legume-specific EST contigs were identified by BLAST comparisons of tentative consensus sequences from *Medicago truncatula*, *Lotus Japonicus*, *Glycine max*, and *Glycine soja* against various sequence data

types from non-legume species such as tentative consensus sequences, GenBank non-redundant database, genomes of rice and Arabidopsis (Graham, Silverstein et al. 2004). In a more recent analysis, which used the predicted proteome of rice, along with genomic sequences from Arabidopsis, Medicago, sorghum, maize, and Poplar, and clustered transcript assemblies from 184 plant species, a set of 861 rice genes that are evolutionarily conserved within, as well as specific to, the Poaceae was identified and characterized (Campbell, Zhu et al. 2007). Interestingly, despite the ongoing comparative studies in plants, only limited analysis on lineage-specific genes in Arabidopsis, the well-characterized dicotyledonous model species, has been performed to date, to the best of our knowledge. In a comparative genomic analysis of Arabidopsis and rice, 116 clusters that have at least two Arabidopsis sequences but no rice sequence were defined as Arabidopsis-specific clusters (Conte, Gaillard et al. 2008; Conte, Gaillard et al. 2008).

4. References

Abel, S., T. Savchenko, et al. (2005). "Genome-wide comparative analysis of the IQD gene families in Arabidopsis thaliana and Oryza sativa." BMC Evol Biol **5**: 72.

Alba, M. M. and J. Castresana (2005). "Inverse relationship between evolutionary rate and age of mammalian genes." Mol Biol Evol **22**(3): 598-606.

Allen, K. D. (2002). "Assaying gene content in Arabidopsis." Proc Natl Acad Sci U S A **99**(14): 9568-72.

Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796-815.

Barry, G. F. (2001). "The use of the Monsanto draft rice genome sequence in research." Plant Physiol **125**(3): 1164-5.

Blomme, T., K. Vandepoele, et al. (2006). "The gain and loss of genes during 600 million years of vertebrate evolution." Genome Biol **7**(5): R43.

Boffelli, D., J. McAuliffe, et al. (2003). "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." Science **299**(5611): 1391-4.

Cai, J. J., P. C. Woo, et al. (2006). "Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota." J Mol Evol **63**(1): 1-11.

Campbell, M. A., W. Zhu, et al. (2007). "Identification and characterization of lineage-specific genes within the Poaceae." Plant Physiol **145**(4): 1311-22.

Cavalier-Smith, T. (1985). "Selfish DNA and the origin of introns." Nature **315**(6017): 283-4.

Cavalier-Smith, T. (1991). "Intron phylogeny: a new hypothesis." Trends Genet **7**(5): 145-8.

Charlebois, R. L., G. D. Clarke, et al. (2003). "Characterization of species-specific genes using a flexible, web-based querying system." FEMS Microbiol Lett **225**(2): 213-20.

Coghlan, A. and K. H. Wolfe (2004). "Origins of recently gained introns in Caenorhabditis." Proc Natl Acad Sci U S A **101**(31): 11362-7.

Conte, M. G., S. Gaillard, et al. (2008). "Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants." BMC Genomics **9**: 183.

Conte, M. G., S. Gaillard, et al. (2008). "GreenPhylDB: a database for plant comparative genomics." Nucleic Acids Res **36**(Database issue): D991-8.

Daubin, V., E. Lerat, et al. (2003). "The source of laterally transferred genes in bacterial genomes." Genome Biol **4**(9): R57.

Daubin, V. and H. Ochman (2004). "Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli." Genome Res **14**(6): 1036-42.

de Souza, S. J., M. Long, et al. (1998). "Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins." Proc Natl Acad Sci U S A **95**(9): 5094-9.

Domazet-Loso, T. and D. Tautz (2003). "An evolutionary analysis of orphan genes in Drosophila." Genome Res **13**(10): 2213-9.

Eisen, J. A. (1998). "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis." Genome Res **8**(3): 163-7.

Fedorov, A., X. Cao, et al. (2001). "Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns." Proc Natl Acad Sci U S A **98**(23): 13177-82.

Fedorov, A., A. F. Merican, et al. (2002). "Large-scale comparison of intron positions among animal, plant, and fungal genes." Proc Natl Acad Sci U S A **99**(25): 16128-33.

Fedorov, A., S. Roy, et al. (2003). "Mystery of intron gain." Genome Res **13**(10): 2236-41.

Fedorov, A., G. Suboch, et al. (1992). "Analysis of nonuniformity in intron phase distribution." Nucleic Acids Res **20**(10): 2553-7.

Fischer, D. and D. Eisenberg (1999). "Finding families for genomic ORFans." Bioinformatics **15**(9): 759-62.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst Zool **19**(2): 99-113.

Force, A., M. Lynch, et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-45.

Fortna, A., Y. Kim, et al. (2004). "Lineage-specific gene duplication and loss in human and great ape evolution." PLoS Biol **2**(7): E207.

Gilbert, W. (1978). "Why genes in pieces?" Nature **271**(5645): 501.

Gilbert, W. (1987). "The exon theory of genes." Cold Spring Harb Symp Quant Biol **52**: 901-5.

Giroux, M. J., M. Clancy, et al. (1994). "De novo synthesis of an intron by the maize transposable element Dissociation." Proc Natl Acad Sci U S A **91**(25): 12150-4.

Goff, S. A., D. Ricke, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)." Science **296**(5565): 92-100.

Graham, M. A., K. A. Silverstein, et al. (2004). "Computational identification and characterization of novel genes from legumes." Plant Physiol **135**(3): 1179-97.

Gu, Z., L. M. Steinmetz, et al. (2003). "Role of duplicate genes in genetic robustness against null mutations." Nature **421**(6918): 63-6.

Guyot, R. and B. Keller (2004). "Ancestral genome duplication in rice." Genome **47**(3): 610-4.

Haas, B. J., A. L. Delcher, et al. (2003). "Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies." Nucleic Acids Res **31**(19): 5654-66.

Haas, B. J., J. R. Wortman, et al. (2005). "Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release." BMC Biol **3**: 7.

Horan, K., J. Lauricha, et al. (2005). "Genome cluster database. A sequence family analysis platform for Arabidopsis and rice." Plant Physiol **138**(1): 47-54.

Hughes, A. L. (1994). "The evolution of functionally novel proteins after gene duplication." Proc Biol Sci **256**(1346): 119-24.

International Rice Genome Sequencing Project (2005). "The map-based sequence of the rice genome." Nature **436**(7052): 793-800.

Ito, Y., K. Takaya, et al. (2005). "Expression of SERK family receptor-like protein kinase genes in rice." Biochim Biophys Acta **1730**(3): 253-8.

Iwamoto, M., M. Mackawwa, et al. (1998). "Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots." Theory and Applied Genetics **97**: 9-19.

Jaillon, O., J. M. Aury, et al. (2007). "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." Nature.

Knowles, D. G. and A. McLysaght (2006). "High rate of recent intron gain and loss in simultaneously duplicated Arabidopsis genes." Mol Biol Evol **23**(8): 1548-57.

Lee, Y., R. Sultana, et al. (2002). "Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA)." Genome Res **12**(3): 493-502.

Levine, M. T., C. D. Jones, et al. (2006). "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression." Proc Natl Acad Sci U S A **103**(26): 9935-9.

Logsdon, J. M., Jr., M. G. Tyshenko, et al. (1995). "Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory." Proc Natl Acad Sci U S A **92**(18): 8507-11.

Long, M., S. J. de Souza, et al. (1995). "Evolution of the intron-exon structure of eukaryotic genes." Curr Opin Genet Dev **5**(6): 774-8.

Lynch, M. and A. Force (2000). "The probability of duplicate gene preservation by subfunctionalization." Genetics **154**(1): 459-73.

Ming, R., S. Hou, et al. (2008). "The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)." Nature **452**(7190): 991-6.

Nahon, J. L. (2003). "Birth of 'human-specific' genes during primate evolution." Genetica **118**(2-3): 193-208.

Nielsen, C. B., B. Friedman, et al. (2004). "Patterns of intron gain and loss in fungi." PLoS Biol **2**(12): e422.

Ohno, S. (1970). Evolution by Gene Duplication, Springer-Verlag, New York.

Ouyang, S., W. Zhu, et al. (2007). "The TIGR Rice Genome Annotation Resource: improvements and new features." Nucleic Acids Res **35**(Database issue): D883-7.

Paterson, A. H., J. E. Bowers, et al. (2004). "Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics." Proc Natl Acad Sci U S A **101**(26): 9903-8.

Qiu, W. G., N. Schisler, et al. (2004). "The evolutionary gain of spliceosomal introns: sequence and phase preferences." Mol Biol Evol **21**(7): 1252-63.

Rensing, S. A., D. Lang, et al. (2008). "The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants." Science **319**(5859): 64-9.

Rensink, W. A., Y. Lee, et al. (2005). "Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts." BMC Genomics **6**: 124.

Rogers, J. H. (1990). "The role of introns in evolution." FEBS Lett **268**(2): 339-43.

Rogozin, I. B., Y. I. Wolf, et al. (2003). "Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution." Curr Biol **13**(17): 1512-7.

Roy, S. W., A. Fedorov, et al. (2003). "Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain." Proc Natl Acad Sci U S A **100**(12): 7158-62.

Roy, S. W. and W. Gilbert (2005). "Complex early genes." Proc Natl Acad Sci U S A **102**(6): 1986-91.

Roy, S. W. and W. Gilbert (2005). "Rates of intron loss and gain: implications for early eukaryotic evolution." Proc Natl Acad Sci U S A **102**(16): 5773-8.

Roy, S. W. and W. Gilbert (2005). "The pattern of intron loss." Proc Natl Acad Sci U S A **102**(3): 713-8.

Rzhetsky, A., F. J. Ayala, et al. (1997). "Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory." Proc Natl Acad Sci U S A **94**(13): 6820-5.

Sakurai, A., S. Fujimori, et al. (2002). "On biased distribution of introns in various eukaryotes." Gene **300**(1-2): 89-95.

Sasaki, T. and B. Burr (2000). "International Rice Genome Sequencing Project: the effort to completely sequence the rice genome." Curr Opin Plant Biol **3**(2): 138-41.

Schmid, K. J. and C. F. Aquadro (2001). "The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes." Genetics **159**(2): 589-98.

Sharp, P. A. (1991). ""Five easy pieces"." Science **254**(5032): 663.

Siew, N. and D. Fischer (2003). "Analysis of singleton ORFans in fully sequenced microbial genomes." Proteins **53**(2): 241-51.

Simillion, C., K. Vandepoele, et al. (2004). "Building genomic profiles for uncovering segmental homology in the twilight zone." Genome Res **14**(6): 1095-106.

Sonnhammer, E. L. and E. V. Koonin (2002). "Orthology, paralogy and proposed classification for paralog subtypes." Trends Genet **18**(12): 619-20.

Striepen, B., A. J. Pruijssers, et al. (2004). "Gene transfer in the evolution of parasite nucleotide biosynthesis." Proc Natl Acad Sci U S A **101**(9): 3154-9.

Sverdlov, A. V., V. N. Babenko, et al. (2004). "Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion." Gene **338**(1): 85-91.

The Arabidopsis Information Resource "http://www.arabidopsis.org/."

The Rice Chromosomes 11 and 12 Sequencing Consortia (2005). "The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications." BMC Biol **3**: 20.

Tomita, M., N. Shimizu, et al. (1996). "Introns and reading frames: correlation between splicing sites and their codon positions." Mol Biol Evol **13**(9): 1219-23.

Tuskan, G. A., S. Difazio, et al. (2006). "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)." Science **313**(5793): 1596-604.

Vandepoele, K., C. Simillion, et al. (2003). "Evidence that rice and other cereals are ancient aneuploids." Plant Cell **15**(9): 2192-202.

Vij, S. and A. K. Tyagi (2006). "Genome-wide analysis of the stress associated protein (SAP) gene family containing A20/AN1 zinc-finger(s) in rice and their phylogenetic relationship with Arabidopsis." Mol Genet Genomics.

Wang, X., X. Shi, et al. (2005). "Duplication and DNA segmental loss in the rice genome: implications for diploidization." New Phytol **165**(3): 937-46.

Wortman, J. R., B. J. Haas, et al. (2003). "Annotation of the Arabidopsis genome." Plant Physiol **132**(2): 461-8.

Yu, J., S. Hu, et al. (2002). "A draft sequence of the rice genome (Oryza sativa L. ssp. indica)." Science **296**(5565): 79-92.

CHAPTER 2. INTRON GAIN AND LOSS IN SEGMENTALLY DUPLICATED GENES IN RICE

A paper published in *Genome Biology* 2006, 7(5): R41

Haining Lin, Wei Zhu, Joana C Silva, Xun Gu, C Robin Buell

Abstract

Background

Introns are under less selection pressure than exons, and consequently, intronic sequences have a higher rate of gain and loss than exons. In a number of plant species, a large portion of the genome has been segmentally duplicated, giving rise to a large set of duplicated genes. The recent completion of the rice genome in which segmental duplication has been documented has allowed us to investigate intron evolution within rice, a diploid monocotyledonous species.

Results

Analysis of segmental duplication in rice revealed that 159 Mb of the 371 Mb genome and 21,570 of the 43,719 non-transposable element-related genes were contained within a duplicated region. In these duplicated regions, 3,101 collinearly paired genes were present. Using this set of segmentally duplicated genes, we investigated intron evolution from full-length cDNA-supported non-transposable element-related gene models of rice. Using gene pairs that have an ortholog in the dicotyledonous model species Arabidopsis thaliana, we identified more intron loss (49 introns within 35 gene pairs) than intron gain (5 introns within 5 gene pairs) following segmental duplication. We were unable to demonstrate preferential intron loss at the 3' end of genes as previously reported in mammalian genomes.

However, we did find that the four nucleotides of exons that flank lost introns had less frequently used 4-mers.

Conclusions

We observed that intron evolution within rice following segmental duplication is largely dominated by intron loss. In two of the five cases of intron gain within segmentally duplicated genes, the gained sequences were similar to transposable elements.

Background

Introns are under less selection pressure than exons, and consequently, their sequences diverge faster than exons. However, the position of the intron with respect to the protein sequence is relatively conserved and conservation of intron position has been observed between distinct eukaryotic lineages throughout ~1.5 billion years of evolution such as between animal and fungal genes [1] and between the malaria parasite *Plasmodium falciparum* and other eukaryotes [2]. With respect to intron position within genes, introns within intron-sparse species as well as single intron genes are preferentially located near the 5' end of the gene [3, 4], suggesting a biased pattern of intron distribution. Indeed, recent studies on 684 eukaryotic orthologous genes from eight eukaryotic species of animals, plants, fungi, and protists showed preferential intron loss [5, 6] and intron gain [6] in the 3' end of genes. This is in contrast to an analysis in fungal species in which no positional bias in intron loss was observed [7].

Introns can be classified into three categories based on location relative to the codon. Introns that do not interrupt the codons are termed phase 0, while phase 1 introns are located between the first and second bases of the codon and phase 2 introns are located between the second and third bases of the codon. It has been reported that eukaryotic genes have more

phase 0 introns than phase 1 or phase 2 introns; on average a 5:3:2 ratio of phase 0: phase 1: phase 2 introns is observed although the specific ratio of intron phase appears to be species specific [8-10]. Several explanations have been proposed for phase bias including legacy of gene formation in the intron early theory [11, 12], phase bias of intron insertion [13], and phase bias of intron loss or selection [5, 7].

Discovery of both intron loss and intron gain suggests that these two processes may be ongoing events in evolution. The rates of intron gain and loss seem to differ greatly among species [2, 7, 14-16] and the underlying mechanism(s) driving intron loss and gain are still unknown. With respect to plants, large-scale computational analyses of intron loss and gain have been focused on *Arabidopsis thaliana*, a model dicotyledonous plant [2, 4-6, 16-20]. With the availability of the near-complete, high quality rice genome sequence [21] and uniform, high quality gene annotation for the genome [22], we have the ability to examine intron loss and gain within a second plant species that represents the other major clade of angiosperms, monocotyledonous plants. Phylogenetic analysis indicates that date of divergence of Arabidopsis and rice is approximately 130-200 Million Years Ago (MYA) [23-25]. Interestingly, depending on the completeness and quality of the genome dataset, as well as the methods and parameters employed, the rice genome underwent a segmental duplication that involved 15-62% of the genome [25-29] and occurred approximate 70 MYA [25, 27], with the exception of the top arms of chromosomes 11 and 12 which underwent a more recent duplication estimated at 5 MYA [27].

Segmental duplication in rice provides the opportunity to study intron gain and loss within a subset of genes that have recently diverged. In this study, we report on the evolution of introns within Coding Sequences (CDS) after segmental duplication in rice. Through our

examination of segmentally duplicated genes, we anticipated that we would identify more intron gain or loss events than for non-duplicated genes due to the accelerated rate of intron loss or intron gain in duplicated versus orthologous genes as reported previously in two malaria parasites [30]. Other advantages of investigating segmentally duplicated genes are that the age of the duplication is approximately 70 MYA [25, 27] which is within the ~100 million years divergence limit for investigating recently gained introns [31, 32] and that segmentally duplicated blocks are more reliable than individually duplicated genes for this type of analysis. Furthermore, we could exploit the phylogeny of rice with *A. thaliana*, a model dicotyledous plant with a near-complete genome sequence, as the outgroup to readily classify "intron loss" and "intron gain" events between the two duplicated rice genes.

Results

Rice segmentally duplicated blocks

Previous analyses of segmental duplication in rice used sequence datasets that contained a substantial portion of unfinished genome sequence and lacked refined structural and functional annotation of the genes [25-29]. Thus, we repeated the analysis of segmental duplication using a set of pseudomolecules (~371 Mb total) that contain 98% finished sequence and had been annotated for genes both at the structural and functional level [22]. Depending on the maximum distance permitted between collinear gene pairs, 25.9-53.4% of the rice genome could be identified as segmentally duplicated (Table 1). Using a maximum distance of 200 Kb between collinear gene pairs, a total of 149 segmentally duplicated blocks were identified (Additional Data File 1). The largest block had 287 pairs of duplicated genes between chromosomes 11 and 12 consistent with the more recent duplicated reported between the top arms of these two chromosomes [27]. These 149 blocks covered 159 Mb

(42.8%) of the 371 Mb genome and contained 21,570 of the total 43,719 non-transposable element (TE) related genes (49.3%) in the rice genome. Of these 21,570 genes, 5,567 were retained within the blocks and corresponded to 3,101 pairs of segmentally duplicated genes distributed across all 12 chromosomes of rice (Additional Data File 2), with chromosomes 1 and 5 having the largest number of duplicated gene pairs (656 pairs).

An increase in genome coverage within the duplicated regions was observed if the maximum distance permitted between collinear gene pairs was expanded from 200 kb to 500kb, 1Mb, or 5 Mb whereas a much smaller percentage of the genome was covered if the maximum distance was limited to 100 kb (Table 1). Previous studies on segmental duplication in the rice genomes reported that 15-62% of the rice genome had undergone segmental duplication [25-29], consistent with our analyses of duplication within the rice genome. As we wished to examine intron evolution within segmentally duplicated genes and there was little difference in percent of the genome identified as duplicated using a maximum distance of 500 kb, 1 Mb, and 5 Mb between collinear gene pairs, we utilized the intermediate estimate of segmental duplication that we obtained using 200 kb as the maximum distance permitted between collinear gene pairs. Thus, our subsequent analyses report on duplicated genes with a maximum distance of 200 kb permitted between collinear gene pairs.

Conservation of exon-intron structure

Within the 43,719 non-transposable element-related gene models in rice, 140,827 introns within the CDS are present, with an average length of 385 bp (std 470) and an average GC content of 37.5%. Out of the 3,101 pairs of segmentally duplicated genes, 281 pairs had at least one intron that passed the manual review for fl-cDNA support and single

isoform. In total, 2,573 introns were present within these 281 gene pairs and had a similar length distribution (average 315 bp) and GC content (36.9% GC) to those found throughout the genome. We found that 197 of the 281 pairs (70%) had completely conserved exon-intron structure in the coding region (958 intron positions in the alignments), i.e., the intron number, position, and phase were identical among the duplicated genes (Fig. 1). The other 84 pairs (30%) had incongruent exon-intron structure. To eliminate the possibility that the incongruence was due to an aberrant alignment, these alignments were manually checked. Only introns surrounded by reliable alignments and only pairs with a putative orthologous gene from Arabidopsis were further investigated. Thus, 48 alignments were excluded and a total of 36 pairs of genes (137 intron positions within the alignments) which showed potential intron loss or intron gain were investigated further.

Abundance of intron loss after segmental duplication

To determine whether the incongruence was due to intron gain or loss, we used the putative orthologous gene from Arabidopsis for the gene pair. From our set of 36 gene pairs with validated alignments, we identified 31 gene pairs with an intron loss(es) (43 intron losses in total), one gene pair with a single gained intron, and four gene pairs in which both intron loss and gain were observed (6 intron losses and 4 intron gains). An example of intron loss is shown in Fig. 2. In this example, the third intron of LOC_Os07g49150.1 was lost as shown by the comparison to the duplicated rice gene model LOC_Os03g18690.1 and the putative ortholog from Arabidopsis At4g29040.1. Alignments of all of the 36 gene pairs with their orthologs from Arabidopsis are displayed in Additional Data File 3. The length of the lost introns (226 bp, std 206) was shorter than the average intron length in the rice genome (385 bp, std 470). The distribution of the length of the lost introns and gained introns and the

frequency of the length of the 33,011 fl-cDNA supported (FLS) rice introns (see Materials & Methods for detail) are shown in Fig. 3.

Intron loss showed no preference at the 3' end of genes

A single intron loss, termed an independent intron loss, was observed in 31 gene pairs as determined by alignment with the putative Arabidopsis ortholog. However, within these 31 gene pairs, 34 introns in total were lost as for three gene pairs, both rice genes underwent separate intron loss events. In these 31 gene pairs, we observed no bias in intron loss position at the 3' ends of genes (Fig 4). Neither was there a bias in the position of intron loss in our set of four gene pairs in which multiple intron losses were observed (data not shown). Interestingly, in one gene pair (LOC_Os05g02130.1 and LOC_Os01g74320.1), all seven introns were lost in LOC_Os01g74320.1, and in LOC_Os07g44140.1, multiple consecutive introns at the 3' end of the gene were lost (see Additional Data File 3).

Intron loss rate at phase 0, 1, 2

Previous reports on intron loss suggested a phase bias [5]. To investigate phase bias in intron loss, we first examined intron phase distribution within the rice genome using a set of introns (33,011 total) derived from the coding regions of 6,046 rice gene models that were supported with fl-cDNA evidence, had no alternative splicing isoform, and had at least one intron within the CDS. The phases of the coding introns were distributed as phase 0 (57.3%): phase 1 (21.5%): phase 2 (21.2%), comparable to the distribution reported previously in plants (62: 17: 21) [1].

To examine whether there was a bias in the phase of intron loss in segmentally duplicated genes in rice, we examined the 34 independently lost introns and excluded genes with multiple intron losses. The frequency of intron loss at phase 2 was higher but not

statistically significant than intron loss at phase 0 and 1 (Table 2, χ^2 test P-value = 0.155). Randomization tests showed that intron loss at phase 2 was unexpectedly high (P-value=0.06) and intron loss at phase 0 was unexpectedly low (P-value=0.08).

Rare 4-mers in the exonic sequence at the donor splice site of lost introns

Previous studies indicated sequence composition preferences surrounding splice sites [13, 33]. As our sample size was small, we restricted our analysis of nucleotide composition surrounding the splice site to the nearest four nucleotides (4-mers); a total of 31 gene pairs with an independent intron loss (34 total introns) were investigated to determine the exonic nucleotide composition flanking each pair of lost and retained intron (Fig. 5). We observed that the 4-mer usage flanking all rice introns was dependent on intron phase (Additional Data Files 4 and 5). For example, ACAA occurs at the exon donor splice site 70, 17 and 110 times at phase 0, phase 1 and phase 2, respectively. In order to determine if intron loss is independent of the nucleotide composition of the exon sequence flanking introns, we compared the 4-mers flanking lost introns with those flanking the corresponding retained introns, as well as with the 4-mers flanking all rice introns. To this end, the exonic 4-mers flanking the donor and acceptor splice sites of the lost and retained introns were each attributed a rank, with rank of 1 being the rarest, according to their frequency in the sample of all rice introns (Tables 3 and 4; see Methods).

The sum of the ranks (SoR) of the exonic 4-mers flanking the donor splice site of the lost introns (Observed SoR = 6,737) was very significantly lower than expected (Expected SoR = 7,647; $P \sim 0.0007$), while that at the acceptor site of the lost introns was within the average range (Table 5). These results reveal a preponderance of rare 4-mers flanking the 5'end of lost introns. This observation is further supported by the fact that the distribution of

ranks of 4-mers flanking the donor splice site in lost introns is significantly lower than that in the corresponding retained introns ($P < 0.013$; Wilcoxon's signed rank test). The rank distributions of 4-mers flanking the acceptor splice site did not differ significantly between lost and retained introns ($P \sim 0.069$).

Source of gained introns

Two out of the five gained introns showed several matches to known rice transposon sequences. The intron of LOC_Os12g02840.1 had a significant hit to a putative Ty1-copia subclass retrotransposon protein (82% identity over the entire intron). A large portion of the other gained intron (LOC_Os12g37660.1; 430 bp out of 741 bp) was highly similar (92% identity) to *Oryza sativa* transposon Rim2-M341 (BK000935) [34]. To ascertain if any of the five gained introns had inserted into other regions of the genome, we searched the five gained introns against our set of 12 pseudomolecules. Three of the gained introns did not match any sequence in the rice genome except itself. For the gained intron in LOC_Os12g02840.1, three high quality matches were detected: to the entire intron of LOC_Os11g03070 (98% identity, putative function of sodium/hydrogen exchanger family protein) which is another segmentally duplicated gene of LOC_Os12g02840.1 from the 5 MYA duplication event; 82% identity to the entire intron of LOC_Os10g05450 (annotated as a hypothetical protein); and 82% identity to the entire intron of LOC_Os06g36500 (annotated as retrotransposon protein, putative, Ty1-copia e subclass). For a second gained intron (LOC_Os12g37660.1), a large portion (~400bp) matched to numerous regions throughout the pseudomolecules. Of the 64 top alignments to the gained intron within LOC_Os12g37660.1 ($\geq$ 95% identity, $\geq$ 400bp in length), 54 were in intergenic regions and 10 were within introns of genes, all of which

lacked fl-cdna support (3 hypothetical proteins, 3 expressed proteins; 2 transposable-element related proteins, and 2 known proteins).

We examined these five cases of intron gain further by examining homologous genes from other plant species. With the exception of one case, the gained intron was clearly a straightforward insertion into one of the rice gene pairs (Additional Data File 6). For LOC_Os3g16960.1, the gained intron was observed in the maize and sorghum homologs, but absent in the Arabidopsis and poplar homologs. Thus, the most parsimonious explanation for the data is a single insertion into one of the rice duplicates prior to the divergence of rice, sorghum, and maize (Fig. 6).

Discussion

Intron loss and gain are two important processes in evolution. We observed more genes with intron loss than gain after segmental duplication in rice. We estimated the rates of intron loss and gain after the segmental genome duplication in rice. Allowing *p* to be the proportion of non-conserved introns between duplicated genes, we have p= 54/(137+958) = 0.0493, where 54 is the number of non-conserved introns, 137 is the total number of the aligned intron positions within the 36 gene pairs which have intron loss and gain, and 958 is the total number of aligned intron positions within the 197 conserved gene pairs. Given that intron loss and acquisition are rare events, the expected rate of intron loss and gain can be estimated under the simple Poisson model and calculated as $D_{int} = -\ln(1-p) = 0.0506$. If we estimate t=70 MYA for the rice genome duplication [21,23], we estimate that the rate of intron gain and loss is

$$\mu = D_{int}/2t = 0.0506/(2*70*10^6) = 3.61\times10^{-10} \text{ per intron per year.}$$

As a total of 49 lost introns and 5 gained introns were observed, we estimated the evolutionary rate of intron loss and intron gain after the genome duplication is

$\mu_{loss} = 3.61\times10^{-10} \times 49/(49+5) = 3.28 \times10^{-10}$ per intron per year

$\mu_{gain} = 3.61\times10^{-10}$ x $5(49+5) = 3.34 \times10^{-11}$ per intron per year

A previous study involving 684 groups of orthologous genes reported an intron loss rate in Arabidopsis of 2-3 $\times10^{-10}$ per year and an intron gain rate of 2.2-2.9 $\times10^{-12}$ per year [16]. Our study, which involved segmentally duplicated genes within rice, revealed a similar intron loss rate but a higher intron gain rate which may be reflective of the reduced evolutionary pressure on duplicated genes. The detection of transposon-related sequences in two of the five gained introns suggests that transposable elements may have a role in intron evolution and is consistent with the increased fraction of transposable elements in the rice genome compared to Arabidopsis [21].

It is possible that the rate of intron loss and gain differs within our set of segmentally duplicated genes as it has been previously reported that the segmental duplication between the top arms of chromosomes 11 and 12 is recent (within 5 MYA; [27]) in comparison to the bulk of the segmental duplication, estimated at 70 MYA. Thus, we determined the d_S for the 233 gene pairs that had a single isoform, were supported by a full-length cDNA, and had been manually validated (197 gene pairs with congruent intron structure and 36 gene pairs with intron loss and/or intron gain). The d_S values ranged from 0.03-24.86 with a clear peak between 0.6-1.4 (data not shown). Similar rates of intron loss (1.41 x 10^{-10} per intron per year) and intron gain (0.94 x 10^{-11} per intron per year) were obtained from the calculations performed using a subset of the 233 gene pairs in which the d_S between duplicates was

between 0.6 and 1.4 (117 pairs total with four gene pairs originating from the top arms of chromosomes 11 and 12).

A reverse transcriptase-mediated model in which a segment of the genomic copy of a gene can be replaced by a reverse-transcribed copy via homologous recombination was previously proposed to explain the pattern of intron loss [3, 35, 36] and has been further supported by recent genomic analysis of several species [5, 6, 15]. The 3' end bias of intron loss is important evidence for this model as reverse transcriptase is error-prone and, as a consequence, a high frequency of 5'-truncated cDNA fragments are generated. Although we did not observe a 3' end preference of intron loss, we did find examples of multiple consecutive intron loss at the 3' end of genes and even loss of all the introns, which is consistent with the reverse-transcriptase-mediated model. Lack of power due to a small sample size (34 lost introns) might be one explanation for the lack of evidence for a 3' bias of intron loss in rice. Another explanation may be the unusual intron distribution pattern which is similar to that of Arabidopsis (data not shown) in which there is no 5' bias in intron location within single-intron genes [4]. The other explanation is that the reverse-transcriptase-mediated model may not be the only mechanism for intron loss in rice and that intron loss may occur via genomic deletion as proposed by Cho *et al*. [37], who observed no intron loss bias at 3' end of genes in *Caenorhabditis*. However, according to the genomic deletion model we would expect some instances of imprecise deletion of introns, which is not the case in our sample. Therefore, an unknown recognition signal may exist that allows the exact deletion of introns in rice.

We did not observe any statistically significant differences in the frequency of intron losses in different phases. Nor did examination of nucleotide compositional patterns in the

exons surrounding the splicing site reveal an apparent pattern in the bi-nucleotide sequence of the exon at the boundary other than that shown by canonical splice site consensus sequence (AG|GT) in which '|' represents the intron position (data not shown). Yet conservation of the exon nucleotides adjacent to the exon-intron boundary has been reported to play an important role in correct splicing [38-40]. Within the four nucleotides at the donor splice site, we observed that the exon boundary of lost introns had less frequently used 4-mers than their corresponding retained introns, as well as relative to the sample of all ~33,000 introns. Thus, genes with less common exonic sequence at the donor site may experience splicing inaccuracy and inefficiency and, consequently, intron loss at these positions may be strongly favored by selection. Alternatively, it is possible that the less common 4-mers reflect exonic sequences more prone to direct intron loss, in the case of the genomic deletion model. Since we did not have a large sample for each intron phase, our data was insufficient to draw a correlation between intron loss rate at each phase and the nucleotide composition of the flanking exonic sequence.

Conclusions

We were able to document intron loss and gain in segmentally duplicated rice genes with a rate of loss and gain similar to that observed within orthologous genes across a range of eukaryotes. While we did not observe preferential intron loss at the 3' end of genes, we did observe a nucleotide bias within the exonic sequence flanking the lost introns.

Materials & Methods

Identification of segmentally duplicated genes

A total of 43,719 non-transposable element related rice protein sequences from release 3 of the TIGR Rice Genome Annotation [22] were used to identify segmental duplication in rice using an all versus all BLAST search (WU-BLASTP, parameters "V=5 B=5 E=1e-10 -filter seg) [41]. As alternative splicing occurs in rice and some genes have multiple splice forms, the largest peptide sequence was used whenever alternative isoforms existed. Repetitive matches were filtered using perl scripts to remove low scoring matches within multiple alignment regions that were defined by a High-scoring Segment Pair (HSP) within 50 Kb. Segmentally duplicated blocks were identified using DAGchainer [42] with parameters "-s -I -D 200000" which primarily includes self comparisons, ignores tandem duplication alignments, and sets the maximum distance allowed between two collinear gene pairs to 200 kb. A minimum of six gene pairs was used to define a block.

Identification of congruent and incongruent introns

Duplicated genes with at least one intron were checked to ensure that they were supported by a fl-cDNA and that no alternative isoforms existed. Intron positions and phases were retrieved from the TIGR Osa1 genome annotation database [22]. ClustalW [43, 44] with default parameter settings was run for each pair to obtain a global alignment. Intron positions and phases were then inserted into the ClustalW alignment using perl scripts. Alignments with incongruent exon-intron structure were manually checked to ensure the introns were supported by reliable alignments. For the 10 amino acids flanking the splice site (five amino acids on each side), we required that at a minimum, three amino acids had to be identical and that $\geq$60% similarity was observed.

Identification of intron loss and intron gain

Simple phylogeny analysis was used to determine if the incongruent exon-intron structure was attributable to loss or gain of an intron. We identified putative orthologous genes by searching the predicted Arabidopsis proteome (TIGR Release 5, [45]) with the predicted rice proteome using blastp (E-value <1e-10) and selecting the reciprocal best hit. In the event we did not identify an ortholog in Arabidopsis via the reciprocal top match method, we used the best Arabidopsis match. Using the Arabidopsis genes as the outgroup, we aligned the rice duplicated gene models to the orthologous Arabidopsis gene model. ClustalW with default parameter settings was run for each triplet (the two rice gene models and their putative Arabidopsis ortholog) and intron positions and phases were inserted into the ClustalW alignment (Additional Data File 3). Only loss or gain of introns after segmental duplication was examined further. An intron loss was defined if the intron was present at the same position in only a single rice gene and the putative Arabidopsis ortholog (referred as a retained intron). An intron gain was defined if the intron was present in single rice gene but absent in the other rice paralog and the putative Arabidopsis ortholog.

Randomization test for intron loss rate at phase 0, 1, 2

A total of 233 pairs of duplicated genes, among which 197 pairs have completely conserved introns and 36 pairs show putative loss and gain of introns, were used in our randomization test. The total number of conserved intron alignment positions at each phase was counted (P0: 580; P1: 236; P2: 225). The total number of independently lost introns at each phase was counted (P0: 15; P1: 7; P2: 12). A total of 10,000 iterations were simulated. A total of 34 phases were randomly generated in each iteration based on the frequencies of the conserved aligned intron positions at each phase from the 233 gene pairs. The number of lost introns at each phase was then compared with those generated by simulation.

Nucleotide composition of exonic sequences flanking lost introns, retained introns, and all introns

In order to determine whether lost introns in duplicated rice genes tend to be flanked by rare nucleotide combinations, we compared the frequency distribution of the four nucleotides (4-mers) in the exonic sequence that flanked lost introns, with the exonic 4-mers flanking the corresponding retained introns, as well as with the frequency distribution of the 4-mers flanking all introns in the genome. Comparisons were done independently for 4-mers flanking the donor and the acceptor ends of introns. The small number of lost introns, distributed over three intron phases (34 introns, of which 15, 7 and 12 were from phases 0, 1 and 2, respectively) relative to the total number of 4-mer classes ($4^4 = 256$) precludes effective use of standard tests, such as the chi-square test, to compare the distributions. Instead, tests based on rank distributions were used as described below.

i) Comparison of 4-mers flanking lost introns vs. all introns

A total of 33,011 introns within the coding regions from 6,046 rice gene models that were supported with fl-cDNA, had no alternative splicing isoform, and had at least one intron within the CDS were used to determine the 4-mer distribution in exonic sequences that flank the introns. The four nucleotides that flank the donor and acceptor splice sites of each intron were extracted and their frequency calculated. For each intron phase, each 4-mer was given a rank between 1 and 256, to cover all of the 4^4 nucleotide combinations, with the lowest frequency having the smallest rank (i.e., rank = 1). In this way, three rank distributions, one for each intron phase 0, 1 and 2, and their attached frequency distributions, were generated for each the donor and the acceptor flanking regions.

We devised a statistic that we call “sum of ranks”, SoR, to determine if the 4-mers flanking lost introns are less common than expected by chance. This statistic SoR corresponds to the sum of the ranks of all introns in a sample, as determined by their nucleotide composition and phase. The test was conducted as follows: 10,000 pseudo-replicates were generated by randomly sampling the three rank distribution obtained for all introns, according to their frequency distribution (i.e., each rank was selected with probability equal to its frequency). Each pseudo-replicate consisted of 34 sampled introns, 15, 7 and 12 of which were sampled from the rank distribution of phase 0, 1, and 2 introns, respectively, to preserve the characteristics of the observed distribution of lost introns. A SoR value was obtained for each pseudo-replicate to generate the distribution of expected “sum of ranks”. The SoR for the 34 lost introns was compared against this distribution to determine the probability *P* of obtaining this value by chance. *P* is approximately equal to the fraction of pseudo-replicated with a smaller or equal SoR value.

ii) Comparision of 4-mers flanking lost introns vs. retained introns, in the corresponding duplicate gene

A rank was attributed to each lost intron, based on the composition of its 4-mer and its intron phase, according to the rank distributions obtained for all 33,011 introns (see above), to obtain a distribution of ranks for the set of lost introns. A distribution of ranks for the set of retained introns was obtained in a similar way. The two distributions were compared using a Wilcoxon’s signed rank test. This procedure was done for both donor and acceptor flanking sequences.

Identification of the source elements of gained introns

Sequences of the five gained introns were searched against NCBI non-redundant database and were further searched against all the 12 rice pseudomolecules [22]. Significant hits were manually checked. For each case of a gained intron, we examined homologous proteins from three plant species with substantial genome sequence: maize, sorghum, and poplar. Using the protein sequences of the 10 rice genes with gained introns, we searched the TIGR Assembled Zea Mays (AZMs) sequences which are assemblies of gene enrichment sequences [46] [47], TIGR Assembled Sorghum Bicolor (ASBs) which are assemblies of gene enrichment reads from sorghum, and contigs from the poplar genome project [48]. All of the top hits from maize and sorghum had > 70% similarity at the protein level with the rice proteins. Gene models were predicted by running the *ab initio* gene finder FGENESH [49] on the maize, sorghum and poplar genomic sequences. We used ClustalW with default parameter settings to align the six proteins (two rice proteins and the homologous proteins from Arabidopsis, maize, sorghum and poplar) and inserted the intron positions/phases into the ClustalW alignment.

Determination of substitutions per site

The number of synonymous substitutions per synonymous site (d_S) between each of the two rice duplicates was estimated by maximum likelihood, using the codon-based substitution model of Yang et al. [50] as implemented in codeml of PAML, version 3.15 [50, 51]. Codeml was run using in pairwise mode (runmode = –2), with codon equilibrium frequencies estimated from average nucleotide frequencies at each codon position (codonFreq = 2). Given the estimated age of ~70 MYA for the polyploidization event in rice [25], and the estimated substitution rate in synonymous sites of ~$6.5X10^{-9}$/site/year [52], rice

paralogs resulting from this polyploidization event are expected to differ on average by ~0.9 synonymous substitution per site.

List of Abbreviations

CDS: Coding Sequence

MYA: Million Years Ago

HSP: High-scoring Segment Pair

FLS: fl-cDNA supported

Acknowledgements

This work was supported by a National Science Foundation Plant Genome Research Program grant to C.R.B. (DBI-0321538).

References

1. Fedorov A, Merican AF, Gilbert W: **Large-scale comparison of intron positions among animal, plant, and fungal genes**. *Proc Natl Acad Sci U S A* 2002, **99**(25):16128-16133.

2. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution**. *Curr Biol* 2003, **13**(17):1512-1517.

3. Fink GR: **Pseudogenes in yeast?** *Cell* 1987, **49**(1):5-6.

4. Sakurai A, Fujimori S, Kochiwa H, Kitamura-Abe S, Washio T, Saito R, Carninci P, Hayashizaki Y, Tomita M: **On biased distribution of introns in various eukaryotes**. *Gene* 2002, **300**(1-2):89-95.

5. Roy SW, Gilbert W: **Complex early genes**. *Proc Natl Acad Sci U S A* 2005, **102**(6):1986-1991.

6. Sverdlov AV, Babenko VN, Rogozin IB, Koonin EV: **Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion**. *Gene* 2004, **338**(1):85-91.

7. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE: **Patterns of intron gain and loss in fungi**. *PLoS Biol* 2004, **2**(12):e422.

8. Fedorov A, Suboch G, Bujakov M, Fedorova L: **Analysis of nonuniformity in intron phase distribution**. *Nucleic Acids Res* 1992, **20**(10):2553-2557.

9. Long M, de Souza SJ, Gilbert W: **Evolution of the intron-exon structure of eukaryotic genes**. *Curr Opin Genet Dev* 1995, **5**(6):774-778.

10. Tomita M, Shimizu N, Brutlag DL: **Introns and reading frames: correlation between splicing sites and their codon positions**. *Mol Biol Evol* 1996, **13**(9):1219-1223.

11. Gilbert W: **The exon theory of genes**. *Cold Spring Harb Symp Quant Biol* 1987, **52**:901-905.

12. Gilbert W, Glynias M: **On the ancient nature of introns**. *Gene* 1993, **135**(1-2):137-144.

13. Qiu WG, Schisler N, Stoltzfus A: **The evolutionary gain of spliceosomal introns: sequence and phase preferences**. *Mol Biol Evol* 2004, **21**(7):1252-1263.

14. Coghlan A, Wolfe KH: **Origins of recently gained introns in Caenorhabditis**. *Proc Natl Acad Sci U S A* 2004, **101**(31):11362-11367.

15. Roy SW, Fedorov A, Gilbert W: **Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain**. *Proc Natl Acad Sci U S A* 2003, **100**(12):7158-7162.

16. Roy SW, Gilbert W: **Rates of intron loss and gain: implications for early eukaryotic evolution**. *Proc Natl Acad Sci U S A* 2005, **102**(16):5773-5778.

17. Roy SW, Gilbert W: **The pattern of intron loss**. *Proc Natl Acad Sci U S A* 2005, **102**(3):713-718.

18. Fedorov A, Roy S, Fedorova L, Gilbert W: **Mystery of intron gain**. *Genome Res* 2003, **13**(10):2236-2241.

19. Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV: **Prevalence of intron gain over intron loss in the evolution of paralogous gene families**. *Nucleic Acids Res* 2004, **32**(12):3724-3733.

20. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: **Conservation versus parallel gains in intron evolution**. *Nucleic Acids Res* 2005, **33**(6):1741-1748.

21. **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.

22. Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, Hamilton J, Haas B, Sultana R, Cheung F *et al*: **The institute for genomic research Osa1 rice genome annotation database**. *Plant Physiol* 2005, **138**(1):18-26.

23. Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH: **Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data**. *Proc Natl Acad Sci U S A* 1989, **86**(16):6201-6205.

24. Crane PR, Friis, E. M.,Pedersen, K. R.: **The origin and early diversification of angiosperms**. *Nature* 2002, **374**:27 - 33

25. Paterson AH, Bowers JE, Chapman BA: **Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics**. *Proc Natl Acad Sci U S A* 2004, **101**(26):9903-9908.

26. Vandepoele K, Simillion C, Van de Peer Y: **Evidence that rice and other cereals are ancient aneuploids**. *Plant Cell* 2003, **15**(9):2192-2202.

27. Wang X, Shi X, Hao B, Ge S, Luo J: **Duplication and DNA segmental loss in the rice genome: implications for diploidization**. *New Phytol* 2005, **165**(3):937-946.

28. Simillion C, Vandepoele K, Saeys Y, Van de Peer Y: **Building genomic profiles for uncovering segmental homology in the twilight zone**. *Genome Res* 2004, **14**(6):1095-1106.

29. Guyot R, Keller B: **Ancestral genome duplication in rice**. *Genome* 2004, **47**(3):610-614.

30. Castillo-Davis CI, Bedford TB, Hartl DL: **Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites**. *Mol Biol Evol* 2004, **21**(7):1422-1427.

31. Logsdon JM, Jr.: **The recent origins of spliceosomal introns revisited**. *Curr Opin Genet Dev* 1998, **8**(6):637-648.

32. Logsdon JM, Jr.: **Worm genomes hold the smoking guns of intron gain**. *Proc Natl Acad Sci U S A* 2004, **101**(31):11195-11196.

33. Long M, Deutsch M: **Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns**. *Mol Biol Evol* 1999, **16**(11):1528-1534.

34. Wang GD, Tian PF, Cheng ZK, Wu G, Jiang JM, Li DB, Li Q, He ZH: **Genomic characterization of Rim2/Hipa elements reveals a CACTA-like transposon superfamily with unique features in the rice genome**. *Mol Genet Genomics* 2003, **270**(3):234-242.

35. Bernstein LB, Mount SM, Weiner AM: **Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites**. *Cell* 1983, **32**(2):461-472.

36. Lewin R: **How mammalian RNA returns to its genome**. *Science* 1983, **219**(4588):1052-1054.

37. Cho S, Jin SW, Cohen A, Ellis RE: **A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution**. *Genome Res* 2004, **14**(7):1207-1220.

38. Seraphin B, Rosbash M: **Exon mutations uncouple 5' splice site selection from U1 snRNA pairing**. *Cell* 1990, **63**(3):619-629.

39. Treisman R, Proudfoot NJ, Shander M, Maniatis T: **A single-base change at a splice site in a beta 0-thalassemic gene causes abnormal RNA splicing**. *Cell* 1982, **29**(3):903-911.

40. Jacobsen SE, Binkowski KA, Olszewski NE: **SPINDLY, a tetratricopeptide repeat protein involved in gibberellin signal transduction in Arabidopsis**. *Proc Natl Acad Sci U S A* 1996, **93**(17):9292-9296.

41. **http://blast.wustl.edu**. In.

42. Haas BJ, Delcher AL, Wortman JR, Salzberg SL: **DAGchainer: a tool for mining segmental genome duplications and synteny**. *Bioinformatics* 2004, **20**(18):3643-3646.

43. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**(13):3497-3500.

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

45. **http://www.tigr.org/tdb/e2k1/ath1/**

46. Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL *et al*: **Enrichment of gene-coding sequences in maize by genome filtration**. *Science* 2003, **302**(5653):2118-2120.

47. **http://maize.tigr.org/**

48. **http://genome.jgi-psf.org/Poptr1/Poptr1.home.html**

49. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA**. *Genome Res* 2000, **10**(4):516-522.

50. Yang Z, Nielsen R, Hasegawa M: **Models of amino acid substitution and applications to mitochondrial protein evolution**. *Mol Biol Evol* 1998, **15**(12):1600-1611.

51. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput Appl Biosci* 1997, **13**(5):555-556.

52. Gaut BS, Morton BR, McCaig BC, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL**. *Proc Natl Acad Sci U S A* 1996, **93**(19):10274-10279.

Additional data files

The following additional data are available with the online version of the paper "Intron gain and loss in segmentally duplicated genes in rice", *Genome Biology* 2006, 7(5): R41. They can be downloaded from http://genomebiology.com/2006/7/5/R41/additional/.

Additional data file 1 lists the segmentally duplicated blocks within the rice genome.

Additional data file 2 lists 3,101 pairs of segmentally duplicated genes along with their pairings and their sequence.

Additional data file 3 shows the ClustalW alignment of the two rice duplicated genes and their orthologous gene from Arabidopsis.

Additional data file 4 lists the occurrence of background exonic 4-mers at the donor splice sites of different intron phase.

Additional data file 5 lists the occurrence of background exonic 4-mer at the acceptor splice sites of different intron phase.

Additional data file 6 shows the ClustalW alignment of the two rice duplicated proteins with putative orthologous proteins from Arabidopsis, poplar, maize and sorghum.

Table 1. Statistics of genome, genes, and regions within segmentally duplicated blocks of the rice genome.

	Maximum distance between collinear gene pairs				
Statistics	100 kb	200Kb	500Kb	1Mb	5Mb
Region covered by duplicated blocks (Mb)	96.04	158.9	193.25	196.35	197.96
Region covered by multiple duplicated blocks (Mb)	7.16	30.6	45.2	45.31	45.74
No. of duplicated blocks	151	149	101	98	96
Genome coverage (%)	25.9	42.8	52.1	52.9	53.4
Non-TE-gene coverage (%)	30.3	49.3	59.1	59.7	60
Total no. of non-TE-genes retained within duplicated blocks	4,377	5,567	5,879	5,894	5,894
No. gene pairs retained within duplicated blocks	2,277	3,101	3,346	3,355	3,355
Total no. non-TE genes within duplicated blocks	13,250	21,570	25,819	26,114	26,248

Table 2. Distribution of phase of intron loss in segmentally duplicated rice genes

	Phase 0	Phase 1	Phase 2
Intron loss[a]	15	7	12
Conserved introns[b]	580	236	225
Intron loss rate[c]	2.5%	2.8%	5.1%

a: Multiple consecutively lost introns were excluded from this analysis.

b: Conserved aligned intron positions within all 235 duplicate gene pairs.

c: Intron loss rate was calculated by (intron loss / (intron loss + conserved introns)) * 100.

Table 3. 4-mer usage of exonic sequence at donor splice site of lost and corresponding retained introns

Intron lost			Intron retained			
Locus name[a]	4-mer[b]	Rank[c]	Locus name[d]	Phase	4-mer	Rank
LOC_Os05g48520.1	CAAG	256	LOC_Os01g48540.1	0	CAAG	256
LOC_Os06g44300.1	CGAG	245	LOC_Os02g08230.1	0	CGAG	245
LOC_Os06g11920.1	CAAG	256	LOC_Os02g51600.1	0	CAAG	256
LOC_Os06g10850.1	GAGG	219	LOC_Os02g52830.1	0	CCAT	211
LOC_Os07g02440.1	CGAC	130	LOC_Os03g55420.1	0	CGAG	245
LOC_Os07g12340.1	CAGG	234	LOC_Os03g60080.1	0	CAGG	234
LOC_Os01g13130.1	CGCC	154	LOC_Os05g14240.1	0	CATG	244
LOC_Os11g01820.1	GCTC	103	LOC_Os05g39600.1	0	CATG	244
LOC_Os12g02840.1	CCTC	172	LOC_Os05g40650.1	0	CCTC	172
LOC_Os02g14430.1	CCAG	251	LOC_Os06g35480.1	0	CAAC	193
LOC_Os09g39720.1	GGAG	246	LOC_Os08g44590.1	0	GGAG	246
LOC_Os02g54640.1	GTTC	28	LOC_Os09g26160.1	0	TTTT	133
LOC_Os08g39370.1	CAAC	193	LOC_Os09g31130.1	0	CAAC	193
LOC_Os08g41880.1	CGAG	245	LOC_Os09g32840.1	0	TGAG	253
LOC_Os03g01820.1	GAGG	219	LOC_Os10g39810.1	0	CAAG	256
LOC_Os05g38420.1	TTCG	225	LOC_Os01g62490.1	1	TTCG	225
LOC_Os06g12960.1	GACG	228	LOC_Os02g50810.1	1	CACG	222
LOC_Os09g26160.1	CATC	54	LOC_Os02g54640.1	1	CACA	171
LOC_Os06g51050.1	ACCG	223	LOC_Os03g04060.1	1	ACAG	250
LOC_Os02g46780.1	GCCG	227	LOC_Os04g50770.1	1	GCAG	251
LOC_Os01g50760.1	GGAG	247	LOC_Os05g46580.1	1	GGAG	247
LOC_Os11g09020.1	GTCG	216	LOC_Os12g08090.1	1	ATCT	194
LOC_Os05g04690.1	CGTG	88	LOC_Os01g18400.1	2	CATG	237
LOC_Os05g48700.1	TGAG	246	LOC_Os01g55240.1	2	TCCG	222
LOC_Os05g39720.1	GGTG	115	LOC_Os01g61080.1	2	GATG	217
LOC_Os07g49280.1	CAAG	254	LOC_Os03g18140.1	2	CCCG	142
LOC_Os07g49150.1	AGAG	251	LOC_Os03g18690.1	2	AGAG	251
LOC_Os07g49000.1	GGAG	245	LOC_Os03g19200.1	2	GGAG	245
LOC_Os09g26360.1	GAAG	249	LOC_Os08g34910.1	2	GAAG	249
LOC_Os08g41730.1	GCGG	208	LOC_Os09g32800.1	2	GCGG	208
LOC_Os12g08090.1	TGCG	115	LOC_Os11g09020.1	2	TGCT	163
LOC_Os01g09540.1	TCGG	225	LOC_Os05g10210.1	2	ATGG	238
LOC_Os05g10210.1	TCCA	175	LOC_Os01g09540.1	2	TAAG	248
LOC_Os03g21820.1	GCCG	195	LOC_Os05g39990.1	2	GCAG	252

a: Locus name of the rice gene model with intron loss.
b: The exonic 4-mer at the donor splice site of the lost intron was inferred from the pair-wise alignment of the coding sequences as illustrated in Fig 5.
c: Each 4-mer is associated with an intron phase-dependent rank ranging from 1 to 256 based on the frequency of occurrence calculated from exonic 4-mers at the exon-intron boundary of all 33,011 FLS introns.
d: The corresponding rice duplicated gene with retained intron.

Table 4. 4-mer usage of exonic sequence at acceptor splice site of lost and corresponding retained introns

Intron lost			Intron retained			
Locus name[a]	4-mer[b]	Rank[c]	Locus name[d]	Phase	4-mer	Rank
LOC_Os05g48520.1	ACCG	53	LOC_Os01g48540.1	0	ATCG	186
LOC_Os06g44300.1	TACA	136	LOC_Os02g08230.1	0	TACA	136
LOC_Os06g11920.1	GGCT	183	LOC_Os02g51600.1	0	GGTT	222
LOC_Os06g10850.1	GCCA	206	LOC_Os02g52830.1	0	GTGA	255
LOC_Os07g02440.1	GGCT	183	LOC_Os03g55420.1	0	GGAT	201
LOC_Os07g12340.1	CTGG	176	LOC_Os03g60080.1	0	TTGG	169
LOC_Os01g13130.1	GCCA	206	LOC_Os05g14240.1	0	GCGA	178
LOC_Os11g01820.1	GTCG	204	LOC_Os05g39600.1	0	GGCG	152
LOC_Os12g02840.1	GCCG	143	LOC_Os05g40650.1	0	GCTG	251
LOC_Os02g14430.1	GGCT	183	LOC_Os06g35480.1	0	GGGT	178
LOC_Os09g39720.1	ATAC	194	LOC_Os08g44590.1	0	ATAT	215
LOC_Os02g54640.1	GTGT	243	LOC_Os09g26160.1	0	GCAT	223
LOC_Os08g39370.1	GTGC	246	LOC_Os09g31130.1	0	ATCA	230
LOC_Os08g41880.1	ATGA	214	LOC_Os09g32840.1	0	ATGA	214
LOC_Os03g01820.1	GCGG	173	LOC_Os10g39810.1	0	ATGG	232
LOC_Os05g38420.1	GCGA	205	LOC_Os01g62490.1	1	GCGA	205
LOC_Os06g12960.1	AGGT	156	LOC_Os02g50810.1	1	AGGT	156
LOC_Os09g26160.1	GGCA	229	LOC_Os02g54640.1	1	AGGA	226
LOC_Os06g51050.1	GCGG	156	LOC_Os03g04060.1	1	GTGG	255
LOC_Os02g46780.1	GATT	244	LOC_Os04g50770.1	1	GTTT	251
LOC_Os01g50760.1	GAAA	246	LOC_Os05g46580.1	1	GGAA	247
LOC_Os11g09020.1	CCAA	156	LOC_Os12g08090.1	1	CCAA	156
LOC_Os05g04690.1	GAAC	235	LOC_Os01g18400.1	2	GAAC	235
LOC_Os05g48700.1	GGCG	189	LOC_Os01g55240.1	2	GGCC	158
LOC_Os05g39720.1	GAGG	218	LOC_Os01g61080.1	2	GAGG	218
LOC_Os07g49280.1	CTTC	163	LOC_Os03g18140.1	2	GTTC	251
LOC_Os07g49150.1	GTAC	255	LOC_Os03g18690.1	2	GTAT	256
LOC_Os07g49000.1	GTAC	255	LOC_Os03g19200.1	2	GTAC	255
LOC_Os09g26360.1	GTAC	255	LOC_Os08g34910.1	2	GTAC	255
LOC_Os08g41730.1	CACG	97	LOC_Os09g32800.1	2	GACG	158
LOC_Os12g08090.1	CGCC	18	LOC_Os11g09020.1	2	GGCG	189
LOC_Os01g09540.1	GTAC	255	LOC_Os05g10210.1	2	AACT	134
LOC_Os05g10210.1	GCCT	182	LOC_Os01g09540.1	2	GTCG	194
LOC_Os03g21820.1	CGTG	153	LOC_Os05g39990.1	2	GGTG	233

a: Locus name of the rice gene model with intron loss.
b: The exonic 4-mer at the acceptor splice site of the lost intron was inferred from the pair-wise alignment of the coding sequences as illustrated in Fig 5.
c: Each 4-mer is associated with an intron phase-dependent rank ranging from 1 to 256 as its based on the frequency of occurrence calculated from exonic 4-mers at the exon-intron boundary of all 33,011 FLS introns.
d: The corresponding rice duplicated gene with retained intron.

Table 5. Sum of the ranks of the exonic 4-mers at the donor and acceptor splice site of lost introns and simulated introns

	Sum of the ranks	
	Donor site	Acceptor site
Lost introns[a]	6,737	6,410
Simulation average[b] (std)	7,647 (253)	6,679 (337)
P-value of lost introns[c]	0.0007	> 0.05

a: Sum of the ranks of the exonic 4-mers at the donor and acceptor splice site of the 34 lost introns.
b: A total of 10,000 iterations were generated. In each iteration, a total of 34 ranks were randomly generated according to the frequencies obtained from all the exonic 4-mers at the exon-intron boundaries of 33,011 FLS introns. Standard deviation is listed in the parenthesis.
c: The P-value for the sums of the ranks of the donor and acceptor splice site.

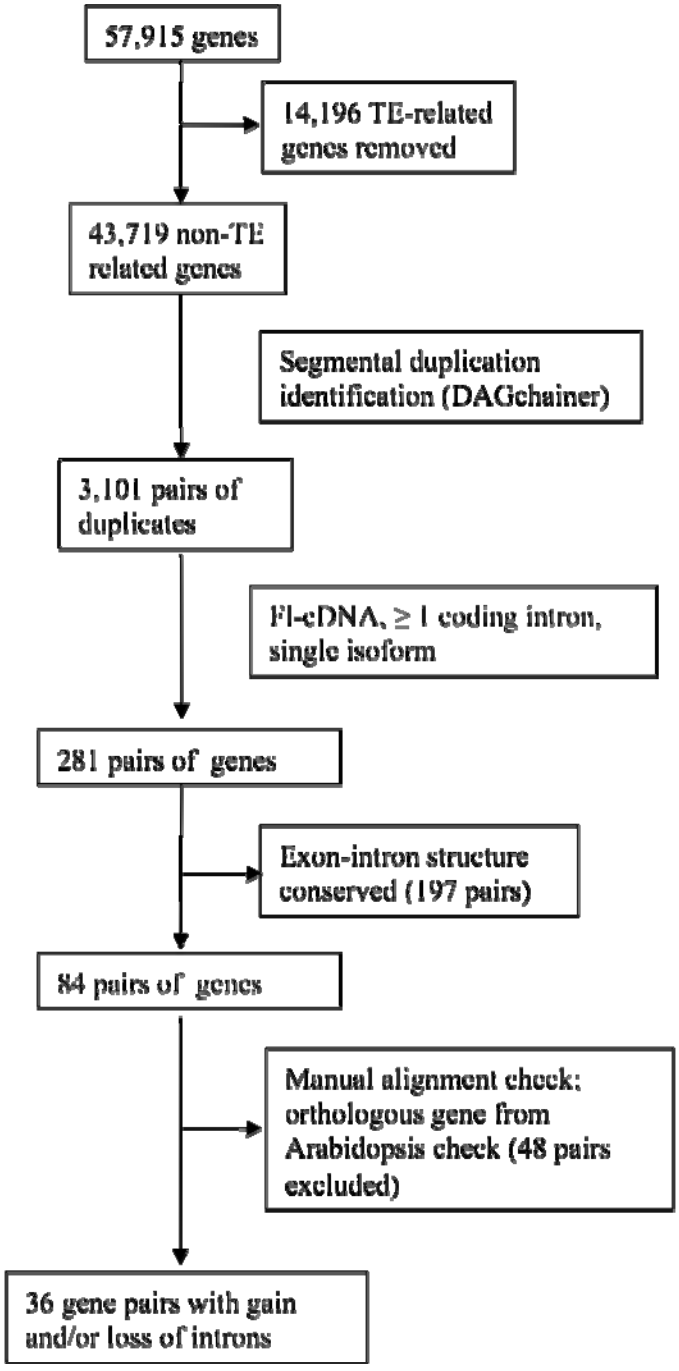


Figure 1. Flow chart for the identification of intron gain and intron loss within segmentally duplicated rice genes. TE, Transposable Element.

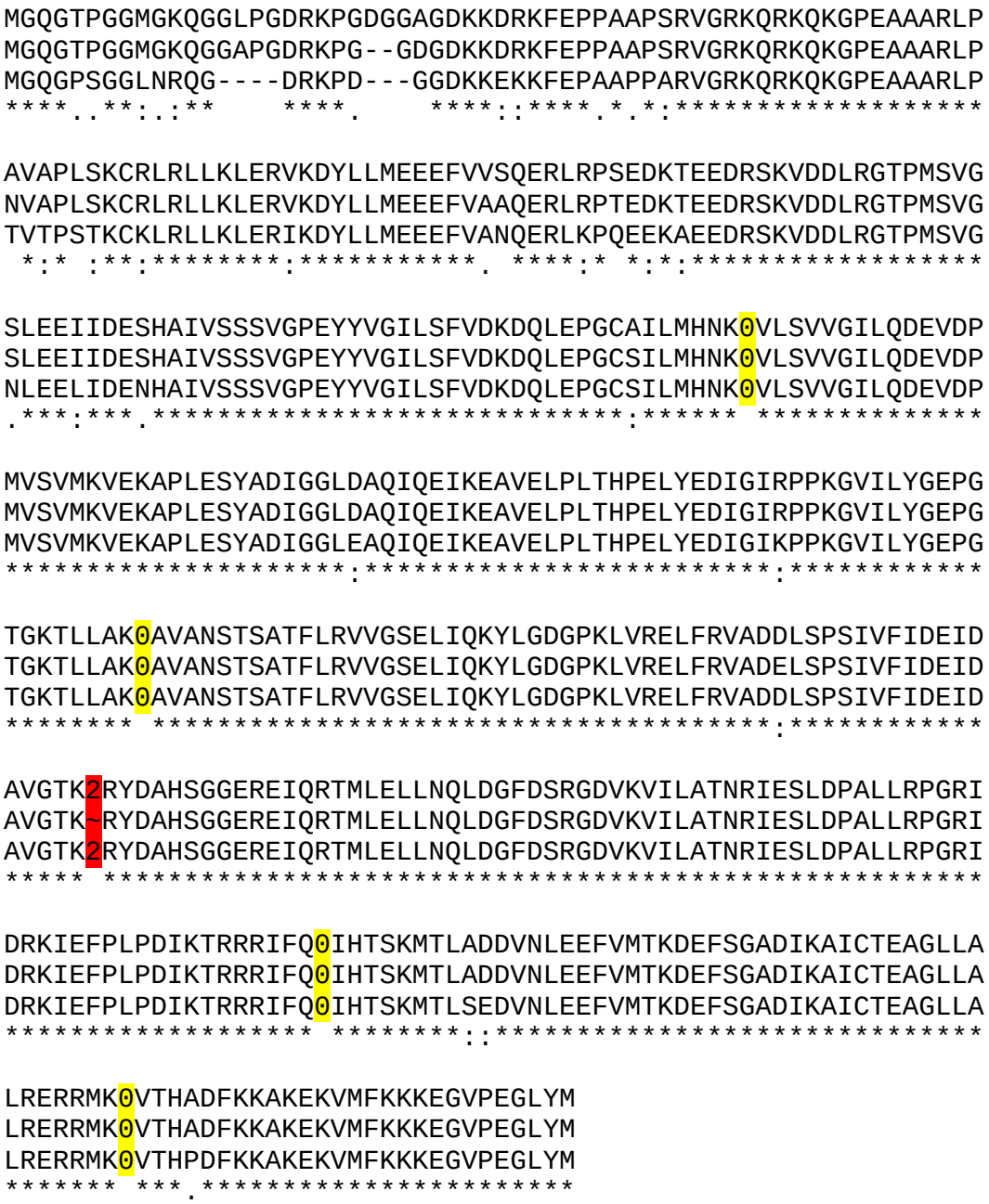


Figure 2. Example of intron loss. Multiple alignment of the two duplicated rice genes (top: LOC_Os03g18690.1, LOC_Os07g49150.1) and their putative orthologous Arabidopsis gene (bottom: At4g29040.1) suggests that the third intron of LOC_ Os07g49150.1 was lost. Yellow inserts indicate conserved introns across the three genes while red indicates lost intron. The phase of the intron is inserted into the alignment. All conserved introns are phase 0 whereas the lost intron is phase 2. The two rice genes and putative Arabidopsis ortholog encode a 26S proteasome regulatory subunit 4.

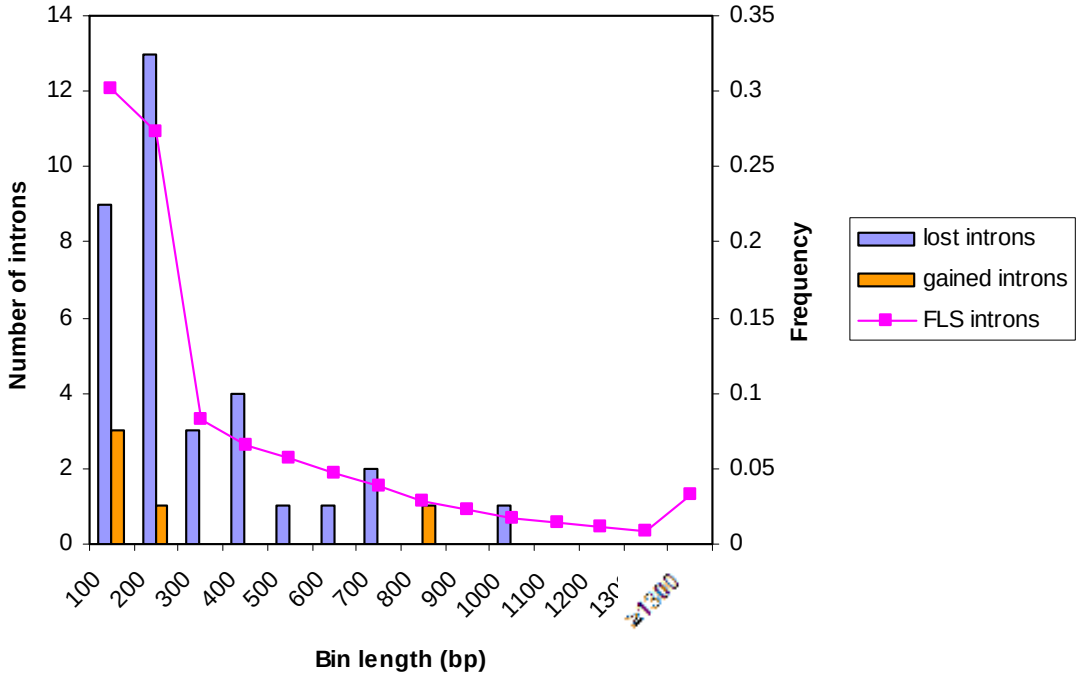


Figure 3. Distribution of the sizes of the lost and gained introns. Intron lengths were binned into 100 bp bins and the number of lost and gained introns in each bin was determined and plotted against the frequency of 33,011 FLS introns within the rice genome.

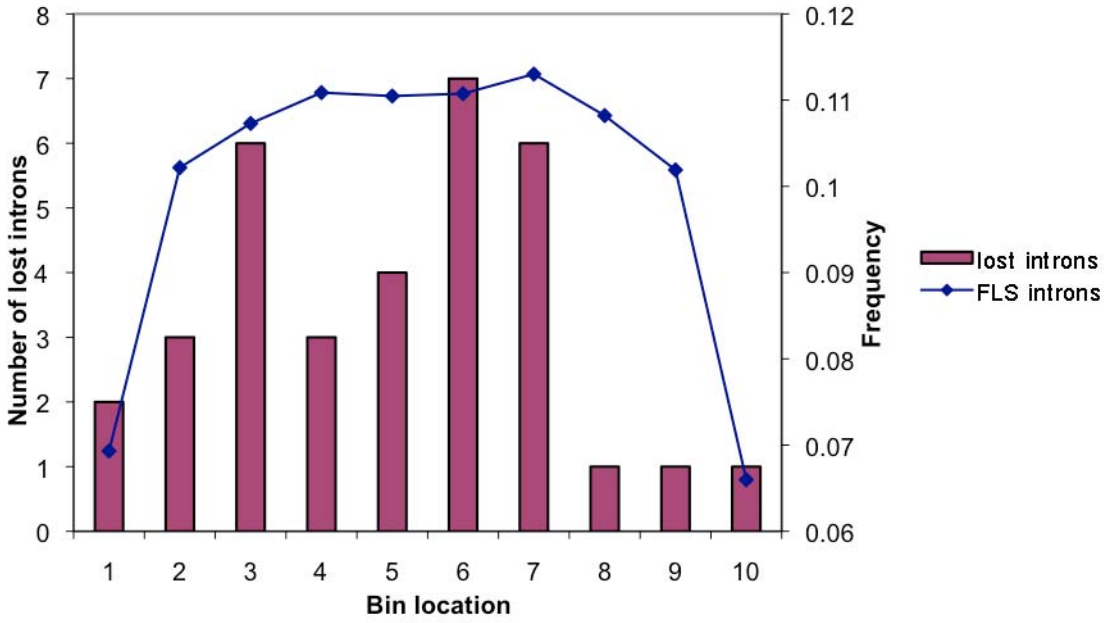


Figure 4. Intron loss along the coding sequence. The positions of the lost introns were inferred from the retained intron of its corresponding duplicated gene. The whole coding sequence was divided into 10 bins. The position of independently lost introns were placed into the corresponding bin and plotted against the frequency of all 33,011 FLS introns within the rice genome which had been binned into the same 10 bins.

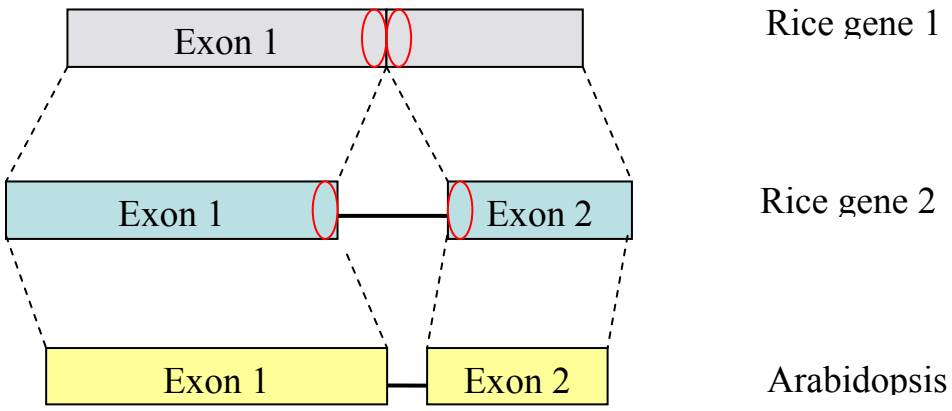


Figure 5. Extraction of the exonic 4-mers at the donor and acceptor splice sites of lost and retained introns. Duplicated rice gene 1 with a single exon and rice gene 2 and Arabidopsis orthologous gene with two exons and a single intron are shown in colored rectangles. Dashed lines indicate similar regions. Phylogeny analysis with Arabidopsis suggests an intron was lost in rice gene 1. The red ovals show the 4-mers extracted for rank sum analysis.

CHAPTER 3. CHARACTERIZATION OF PARALOGOUS PROTEIN FAMILIES IN RICE

A paper published in *BMC Plant Biology* 8: 18

Haining Lin, Shu Ouyang, Amy Egan, Kan Nobuta, Brian J. Haas, Wei Zhu, Xun Gu, Joana C Silva, Blake C. Meyers, and C. Robin Buell

Abstract

Background

High gene numbers in plant genomes reflect polyploidy and major gene duplication events. *Oryza sativa*, cultivated rice, is a diploid monocotyledonous species with a ~390 Mb genome that has undergone segmental duplication of a substantial portion of its genome. This, coupled with other genetic events such as tandem duplications, has resulted in a substantial number of its genes, and resulting proteins, occurring in paralogous families.

Results

Using a computational pipeline that utilizes Pfam and novel protein domains, we characterized paralogous families in rice and compared these with paralogous families in the model dicotyledonous diploid species, *Arabidopsis thaliana*. Arabidopsis, which has undergone genome duplication as well, has a substantially smaller genome (~120 Mb) and gene complement compared to rice. Overall, 53% and 68% of the non-transposable element-related rice and Arabidopsis proteins could be classified into paralogous protein families, respectively. Singleton and paralogous family genes differed substantially in their likelihood of encoding a protein of known or putative function; 26% and 66% of singleton genes compared to 73% and 96% of the paralogous family genes encode a known or putative

protein in rice and Arabidopsis, respectively. Furthermore, a major skew in the distribution of specific gene function was observed; a total of 17 Gene Ontology categories in both rice and Arabidopsis were statistically significant in their differential distribution between paralogous family and singleton proteins. In contrast to mammalian organisms, we found that duplicated genes in rice and Arabidopsis tend to have more alternative splice forms. Using data from Massively Parallel Signature Sequencing, we show that a significant portion of the duplicated genes in rice show divergent expression although a correlation between sequence divergence and correlation of expression could be seen in very young genes.

Conclusions

Collectively, these data suggest that while co-regulation and conserved function are present in some paralogous protein family members, evolutionary pressures have resulted in functional divergence with differential expression patterns.

Background

Gene duplication is a major contributor to genetic novelty and proteomic complexity. Evolutionary pressures on duplicated genes differ from single copy (singleton) genes and several models have been proposed for the evolutionary fate of duplicated genes. In the non/neofunctionalization model, one of the duplicated genes becomes a pseudogene through the accumulation of deleterious mutations although on a rare occasion, it may acquire a new function [1]. In the subfunctionalization model [2-4], duplicated genes adopt a subset of functions of the ancestral gene. Functional redundancy of duplicated genes has been shown to increase the robustness of biological systems [5].

Gene duplication occurs frequently in plants, either in the form of segmental duplication, tandem duplication, and at the level of whole genome duplication [6-14]. Genome duplication has been reported in rice (*Oryza sativa*), an important agricultural species and model species for the grass family (Poaceae) [15-19]. Depending on the methods, parameters, and genome assemblies used, 15% to 62% [15-19] of the rice genome underwent one round of large-scale segmental duplication that occurred approximately 70 Million Years Ago (MYA) [15, 16, 18]. A more recent duplication, on the short arms of chromosomes 11 and 12, occurred approximately 5 ~ 8 MYA [15, 20]. With respect to tandem duplications, depending on the parameters utilized, 14-29% of rice genes occur in tandem [21]. Paralogous families, composed of tandemly and segmentally duplicated genes, have been studied to a limited extent in rice, typically in a comparative context with the finished genome of the dicotyledonous plant species, *Arabidopsis thaliana* [22-27]. To date, only limited genome-wide analyses of paralogous protein families have been reported in rice [28, 29]. In Horan *et al*. [28], Arabidopsis and rice proteins were co-clustered using Pfam domain-based or BLASTP-based similarity clustering which allowed for the clustering of proteins into families common between these two model species and for the identification of proteins that were species-specific.

In this study, we classified proteins from the predicted rice proteome into paralogous protein families using a computational pipeline that utilizes both Pfam and BLASTP-based novel domains [30]. While the focus in our study was analysis of the rice paralogous families, for comparative purposes, we performed a similar classification with the predicted Arabidopsis proteome to compare and contrast paralogous family composition and features

in two model species which represent two major divisions of the angiosperms, monocots and dicots. In rice, we characterized alternative splicing, functional classification of paralogous family proteins, expression patterns, and duplication age and compared these data to those observed in single copy proteins. A parallel analysis of alternative splicing and functional domain composition of paralogous family proteins was performed with Arabidopsis to compare and contrast with the findings in rice. To highlight our observations, we examined in depth two rice protein families, prolamin and Bowman-Birk inhibitor. This study provides a comprehensive analysis of rice paralogous families in parallel with a comparative analysis in Arabidopsis thereby providing novel insight into paralogous gene family evolution in these two model plant species.

Results and Discussion

Classification of paralogous protein families in rice and Arabidopsis

A total of 3,865 paralogous protein families containing 21,998 proteins were identified [see Additional file 1] from the 42,653 total non-transposable element (TE)-related proteins predicted in the rice genome, leaving 20,655 putative singleton proteins encoded by single copy genes. On average, a rice family contained six family members, ranging in size from two to 214 family members (Fig. 1). A total of 11 paralogous protein families with more than one hundred member proteins were identified in rice which encoded proteins such as zinc finger proteins, protein kinases, Myb-like proteins, and transducins [see Additional file 2], similar to the largest protein families reported in Arabidopsis [30]. Paralogous protein family genes of rice were distributed throughout the genome and within chromosomes in a pattern similar to the singleton genes [see Additional file 3A]. Although paralogous protein

family genes were more frequently located in the euchromatic regions, this was consistent with previous reports that non-TE-related genes are found more prevalently in euchromatic regions. A comparison of segmentally duplicated genes with the paralogous protein family genes suggested that our classification pipeline was robust. Of the 2,403 segmentally duplicated gene pairs within 163 segmentally duplicated blocks, 1,570 duplicated gene pairs (65%) were classified in the same paralogous protein family. For the remainder of the segmentally duplicated genes, 175 pairs (7%) were classified in different paralogous protein families and 268 (11%) had one gene classified in a paralogous protein family and the other gene classified as a singleton. We observed that 390 segmentally duplicated gene pairs (16%) were not included in any paralogous protein family. Note that in our computational pipeline, four or more members were required to define a BLASTP-based domain. Consequently, a single pair of segmentally duplicated genes alone is insufficient to define a BLASTP-based domain. The lack of 100% correspondence between segmental duplication and paralogous family classification may be due to the acquisition of new domain(s) or loss of existing domain(s) within one of the duplicated genes as in our computational pipeline, only proteins with the identical domain composition were classified into the same paralogous protein family. Alternatively, the difference could be due to the different classification methods employed in each method. For example, LOC_Os08g37350 and LOC_Os09g28940 are segmentally duplicated genes from chromosomes 8 and 9, respectively. These two protein sequences had a 56% identity over 70% of the length of the longer sequence and were within a segmentally duplicated block of 43 collinear gene pairs. LOC_Os08g37350 has two Pfam domains (PF00443: Ubiquitin carboxyl-terminal hydrolase; PF01753: MYND finger) while

LOC_Os09g28940 has only one Pfam domain (PF00443: Ubiquitin carboxyl-terminal hydrolase). As a consequence, these loci were classified in two different paralogous families (LOC_Os08g37350 is classified in Family 1545; LOC_Os09g28940 is in Family 3650). In a second example, LOC_Os11g03210 and LOC_Os12g02960 are from a segmental duplication event involving chromosomes 11 and 12 which includes 160 collinear gene pairs. LOC_Os11g03210 has a single Pfam domain (PF02798: Glutathione S-transferase, N-terminal domain) and thus is classified in Family 3362 while LOC_Os12g02960 is classified as a singleton as although it has two Pfam domains (PF02798: Glutathione S-transferase, N-terminal domain; PF00043: Glutathione S-transferase, C-terminal domain) no other protein has exactly the same domain profile. Note that in our computational pipeline, a paralogous family must have at least two members with identical domain profiles. In a third example, segmentally duplicated genes LOC_Os01g41900 and LOC_Os05g51160 are from chromosomes 1 and 5. These two genes were derived from full length cDNAs (FLcDNAs) and had a 59% identity over approximately three-quarters of the longer protein sequence. LOC_Os01g41900 has two Pfam domains (PF00249: Myb-like DNA-binding domain and PF00098: Zinc knuckle) while LOC_Os05g51160 has only one single Pfam domain (PF00249: Myb-like DNA-binding domain). As a consequence, they were classified in different families, Family 1452 and Family 3863, respectively. Manual inspection of these three sets of loci revealed that they were correctly annotated and that the lack of clustering into a single paralogous family could not be attributed to incorrect structural annotation which is another potential cause for lack of 100% correspondence between segmentally duplicated genes and paralogous families.

A parallel construction of paralogous protein families in Arabidopsis identified 3,092 paralogous protein families (18,183 proteins) and 8,636 single copy genes from a total of 26,819 protein coding genes from TAIR7 release [31]. A similar size distribution of Arabidopsis protein families was observed, ranging from two to 182 (Fig. 1). In Arabidopsis, the largest families encode Myb-like proteins, zinc finger proteins, and protein kinases, consistent with what has been reported previously [30]. Arabidopsis paralogous protein family genes distributed similarly to singleton genes and were more frequently located in the euchromatic regions [see Additional file 3B].

Function of paralogous protein families in rice and Arabidopsis

We examined the functional annotation of paralogous family and singleton proteins. A total of 21,403 and 23,081 genes were annotated as encoding known or putative proteins in rice and Arabidopsis, respectively, due to strong similarity with proteins with a known function or the presence of Pfam domains above the trusted cutoff. Genes with no known or putative function can be supported by experimental transcript evidence (i.e., encode an "expressed protein") or are predicted solely by an *ab initio* gene finder and lack expression support as well as sequence similarity to known proteins with the exception of other hypothetical proteins (i.e., encode a "hypothetical protein"). In rice, a total of 6,913 genes encode expressed proteins as shown by experimental transcript evidence from Expressed Sequence Tags (ESTs), FLcDNAs, Massively Parallel Signature Sequencing [32], Serial Analysis of Gene Expression, and/or proteomic data [33]. In Arabidopsis, 2,270 genes encode expressed proteins as shown by experimental transcript in the form of ESTs and/or cDNA evidence (see Methods). The remaining 14,337 rice genes [33] and 1,468 Arabidopsis

genes (see Methods) encode hypothetical proteins. A majority of rice paralogous family genes (73%) encode either a known or putative protein (Fig. 2). The remaining rice paralogous family genes encode expressed proteins (9%) and hypothetical proteins (18%). In contrast, rice singletons had a larger portion of hypothetical genes (50%) and a smaller portion of genes with a known or putative function (26%). Even though Arabidopsis overall has a smaller number of genes with unknown function than rice, a similar bias of genes with a known or putative function in paralogous family genes was observed in a parallel analysis in Arabidopsis (Fig. 2).

Using Plant GOSlim annotations [34], we compared the function of the proteins within rice paralogous families to that in the singletons. Within the 26 molecular function GOSlim categories identified in our analyses, rice paralogous protein families showed different patterns from singletons in a number of GOSlim categories (Fig. 3A). Although, the relative abundance of each GOSlim category varied with the size of the rice paralogous family, no obvious correlation was observed (Fig. 3A). For each category, a two-tailed two-sample binomial test was performed by comparing the abundance of that category in rice paralogous families with that in the singletons. Multiple testing was corrected using the Benjamini and Hochberg false discovery rate control at a level of 0.05 [35]. The statistical test revealed a substantial enrichment of 12 categories in rice paralogous family proteins including transcription factor activity, hydrolase activity, DNA binding, and transporter activity while a substantial reduction was seen in five categories including receptor activity, nucleotide binding and carbohydrate binding (Table 1). A similar skew in GOSlim categories was observed in a parallel analysis in Arabidopsis (Table 2 & Fig. 3B), consistent with a

previous report in Arabidopsis [36] that non-random loss and retention of paralogous genes with different functions occurred after gene duplication.

Paralogous protein family genes tend to have more alternative isoforms than singletons

Alternative splicing has been regarded as a mechanism to increase genetic novelty. In the rice genome, 6,253 non-TE-related genes have evidence of alternative splicing (see Methods) and we used this set of genes to examine alternative splicing in singleton versus paralogous protein family genes. The percentage of alternative splicing in single copy genes is 2,094/20,655 = 10.1%, while that in paralogous family genes is 4,159/21,998 = 18.9%; a statistically significant difference (χ^2 test, P < 1e-5). To remove any bias due to genes that lack transcript evidence, we restricted our analysis to genes with EST and/or FLcDNA evidence. The percentage of alternative splicing in singletons is 2,094/ 8,619 = 24.3%, while that in paralogous protein family genes is 4,159/14,072 = 29.6%; a statistically significant difference (χ^2 test, P < 1e-5). We further restricted our analysis to high confidence genes whose structures were completely supported by ESTs and/or FLcDNAs. The percentage of alternative splicing in singletons increases to 1,826/5,964 = 30.6%, while that in paralogous protein family genes increases to 3,765/11,235 = 33.5%; a statistically significant difference (χ^2 test, P < 1e-3).

To confirm that our observation was not restricted to rice, we performed a parallel analysis with Arabidopsis. Using data on alternative splicing as provided with the TAIR7 release (see Methods), the percentage of alternative splicing in Arabidopsis single copy genes is 943/8,636 = 9.8%, while that in paralogous protein family genes is 2,856/18,183 = 15.7%. This difference is also statistically significant (χ^2 test, P < 1e-5), similar to that observed in

rice. Restricting the analysis to only those Arabidopsis genes with EST and/or cDNA support as provided in the TAIR7 release revealed that the percentage of alternative splicing in singletons is 942/6,663 = 14.1%, while that in paralogous family genes is 2,852/15,369 = 18.6%; a statistically significant difference (χ^2 test, P < 1e-5). Our findings are contradictory to previous reports in model animal species in which duplicated genes tend to have fewer alternative spliced isoforms thereby supporting the 'function-sharing model' that alternative splicing and gene duplication are two mechanisms that are complementary with respect to proteomic function diversity [37, 38]. Our results suggested that plants may employ multiple mechanisms for proteomic complexity, gene duplication and alternative splicing.

Age of paralogous protein families in rice

While there are previous reports on gene duplication in rice [15-19], they utilized alternative assemblies and annotation datasets of the rice genome. To provide information on the age of paralogous families identified in this study, we estimated the age of a paralogous family from the maximum value of the distribution of pairwise d_S calculated among all members of that protein family (see Methods). We found that the origin of most paralogous families dates back to over 115 Million Years (MY), the point at which synonymous sites are saturated and dating becomes unreliable (d_S ~1.5) [see Additional file 4A]. Among protein families for which the maximum pairwise d_S value is less than 1.5, the distribution of maximum d_S is fairly flat, with the exception of a recent peak at d_S between 0 and 0.1 [see Additional file 4B]. This suggests that paralogous families have been arising at a relatively constant pace within the past 115 MY, but that a burst of duplication took place within the last 7.5 MY. Alternatively, paralogous families arise at a rate similar to that observed for the

first few million years, but about 2/3 of them revert to single-gene status soon thereafter, accounting for the quick decline after the first 7.5 MY. The fairly constant number of older paralogous families can be due to selective constraints maintaining the elevated copy number or if the loss of paralogs is dependent on sequence similarity, such that after ~10% sequence divergence, paralog loss is negligible. Finally, for each family we identified the largest peak below 1.5 (if there was one) in the distribution of all pairwise d_S values. The distribution of this peak value across all families is bimodal [see Additional file 5], and it confirms the presence of a large number of recently duplicated genes ($0 \leq d_S < 0.1$). In addition, the peak at $0.7 \leq d_S \leq 1$ most likely results from the large-scale segmental duplication event that occurred ~70 MYA.

Expression of paralogous protein families in rice

We further examined the expression patterns of the paralogous families using MPSS data from 18 libraries [32]. MPSS tags were searched against our release 4 pseudomolecules and cDNA sequences of all annotated gene models to ensure that all MPSS tags would be identified even if they spanned the intron(s). We found 11,619 genes within the paralogous protein families that were associated with unique, reliable, and significant MPSS tags, which were referred as MPSS-qualifying genes.

Suitable summary statistics of correlation for expression divergence of a gene family can be found in Gu [39] and Gu *et al*. [40], though microarray data were the primary focus in these studies. To be concise, we restricted our analysis of expression correlation in the libraries and tissues to paralogous families with exactly two MPSS-qualifying genes (674 protein families). To measure the expression correlation, the Pearson's Correlation

Coefficient (*r*) of their expression was computed for each pair of MPSS-qualifying genes from each of the 674 protein families across all 18 MPSS libraries. It is important to note that we excluded MPSS tags which mapped to multiple locations, as most of these are likely to match to closely-related paralogs and could have confounded our analyses. We employed the method used by Blanc and Wolfe [36] to determine a minimum cutoff value for Pearson's Correlation Coefficient (*r*) to classify two duplicated genes as having divergent expression. Basically, a total of 10,000 gene pairs were generated by random shuffling of the singleton genes and the Pearson's Correlation Coefficient (*r*) was calculated similarly for each pair. Ninety five percent of the random shuffled gene pairs had a correlation value $r < 0.59$. As random shuffled gene pairs should have divergent function and expression patterns, we utilized $r < 0.59$ as an indicator of divergent expression. Our results show that the expression correlation value (*r*) of the paralogous protein family genes ranged from -0.6 to 1.0 although the majority of the gene pairs had little correlation with *r* peaking at -0.2 ~ 0, similar to that observed with the singletons (Fig. 4). Using the correlation cutoff ($r = 0.59$), a total of 598 (89%) paralogous protein families with two-qualifying MPSS genes exhibited divergent expression patterns, consistent with what has been reported in Arabidopsis [36] and in yeast in which more than 80% of the older duplicated gene pairs ($ds > 1.5$) showed divergence in expression [41].

To gain a better understanding of the expression patterns of paralogous protein family members in different organs/tissues, we classified the 18 MPSS libraries [32] into four groups by organs/tissues: roots, leaves, reproductive organs/tissues, and "other tissues". Within the 674 paralogous families with exactly two MPSS-qualifying genes, 239, 168, 223,

and 200 paralogous families had only a single member of the pair expressed in roots, leaves, reproductive organs/tissues, and "other tissues", respectively, which demonstrated their diverged expression patterns, and possible tissue-specific expression. To further examine the tissue-specific or stress-induced expression patterns of paralogous protein family members, we calculated the Preferential Expression Measure (PEM) for each of the 1,348 genes from the 674 paralogous families (see Methods) in the 18 MPSS libraries. The PEM shows the base-10 log of ratio of the observed expression level in a given tissue/treatment to the expected expression level assuming uniform expression across all tissues/treatments. A PEM value of 1 means the observed expression level in a given tissue/treatment is 10 times that of expected and indicates strong tissue specific expression. For each gene, tissue(s) with a stringent cutoff of PEM ≥ 1 were compared with the other member of the duplicated gene pair. A total of 375 (375/674 = 55.6%) of the paralogous families showed little tissue-specific expression as none of the associated PEMs had a value equal to or greater than 1. Two hundred ninety-nine families showed strong tissue specific expression patterns; 19 families were preferentially expressed in the same tissue or treatment, 49 families were preferentially expressed in different tissues or treatments, and 231 families had only one of the duplicated genes with preferential tissue-specific expression.

We further examined the correlation between expression divergence and sequence divergence. For each family, we calculated the Pearson's Correlation Coefficient (r) for all possible pairs of the MPSS-qualifying genes to measure expression divergence. We then used *ds* as a proxy of divergence time for each gene pair. We restricted our analysis to $d_S \leq 1.5$ so that the synonymous sites are not saturated. The Pearson's Correlation Coefficient (r)

values were plotted against the *d*s values for each interval of 0.1 to gain better resolution. That is, we plotted for gene pairs with $0 < d_S \leq 0.1$, $0.1 < d_S \leq 0.2$, $0.2 < d_S \leq 0.3$, and so on. We found no correlation between d_S and correlation of expression except for gene pairs with $0 < d_S \leq 0.1$ (R = 0.33, P < 1e-4) where duplicated genes were relatively young [see Additional file 6]. The number of non-synonymous substitutions per site (*dN*) was also calculated for each gene pair and plotted against correlation of expression. No correlation was observed between *dN* and correlation of expression (data not shown). This is consistent with reports in Arabidopsis in which expression divergence is not strictly coupled with sequence divergence as shown by no appreciable change for the majority of gene duplicates with highly diverged amino acid sequences in expression pattern in developing roots [42].

Positive correlation of expression patterns among paralogous protein family members would suggest that similar transcriptional regulation was retained in both members and possibly, similar functions. However, we observed a large number of gene pairs with little expression correlation which could be an indication of subfunctionalization or neofunctionalization after gene duplication. The duplication-degeneration-complementarity (DDC) model proposed by Force et al. [3] and Lynch and Force [4] suggests that subfunctionalization is a major mechanism for retention of duplicated genes as a result of differential expression caused by accumulation of mutations in regulatory regions rather than protein coding regions. The 49 families with preferential expression in two different tissues or treatments, along with the 231 families having only one member of the paralogous pair preferentially expressed, is a strong indicator of subfunctionalization. As our paralogous protein family classification required that each family member have the same domain profile,

the differential expression may be attributable to mutations in regulatory regions rather than gene coding regions, consistent with the DDC model.

Case studies of rice paralogous protein families

Prolamin protein family

Prolamin is one of the major endosperm storage proteins in cereal grains such as wheat, barley, rye, maize, and sorghum [43-46]. It was named prolamin due to its high content of proline and glutamine. In rice, prolamin contributes 35% of the total seed protein [47]. Three classes of prolamins have been identified in *Oryza* by their molecular weights: 10, 13, and 16 kDa [48]. The major prolamin families in rice are Family 3722 (20 members) and Family 3193 (seven members). Members of both families have a BLASTP-based domain. Members of Family 3193 have a Pfam domain (PF00234; Protease inhibitor/seed storage/LTP family) in addition to the common BLASTP-based domain and thus were not clustered within Family 3722 as the exact same domain profile is required for each family member in our computational pipeline [see Additional file 7]. All of the prolamin genes were single-exon genes as reported previously [49] with the exception of four genes that contained a single intron which were further examined and found that based on the EST alignments they were single-exon genes that had not been properly annotated (data not shown). The length of the deduced amino acids of the prolamin proteins (excluding the four inaccurate genes) varied from 101 to 156 bp with two peaks at 101~110 and 145~160 bp, consistent with what had been reported in rice prolamin proteins [49, 50].

Only five prolamin family members (LOC_Os05g26720.1, LOC_Os05g26770.1, LOC_Os06g31070.1, LOC_Os12g16880.1, LOC_Os12g16890.1) were associated with

unique, reliable, and significant MPSS tags, which, as expected, were exclusively expressed in 3-day germinating seeds with relatively high abundances (198, 562, 1042, 148, and 670 Transcripts Per Million (TPM), respectively) [see Additional file 8]. We also examined the expression of the two prolamin families with that of Family 3856 (123 members) which contained the same Pfam domain (PF00234) that was in prolamin family 3193 [see Additional file 7]. A total of 54 genes from Family 3856 were associated with unique, reliable, and significant MPSS tags. However, the expression pattern observed in Family 3856 substantially differed from that of the prolamin families (Family 3722 and Family 3193) in that most of the genes were expressed in multiple organs/tissues [see Additional file 9].

Interestingly, we observed that genes encoding the prolamin protein family seemed to localize closely on the chromosomes. A total of 16 prolamin protein family genes were located together on chromosome 5 with a large number of TE-related genes inserted between the family members [see Additional file 10]. Other prolamin protein family genes were located on chromosome 6 (two genes in tandem), chromosome 7 (in two gene clusters), and chromosome 12 (three genes with TE-related genes inserted between them), suggestive of tandem duplication(s) of the prolamin protein family genes followed by insertion of transposable elements throughout the course of evolution. This is consistent with previous report on the compact expansion of α-zein gene family of maize [13].

Bowman-Birk Inhibitor (BBI) type protein family

BBI is a cysteine-rich protein which has trypsin and chymotrypsin inhibitory activities [51]. It was first characterized in soybean [52, 53] and later found widely

distributed in monocot and dicot species [54-58]. It has been extensively studied due to its possible role in plant defense [51, 54, 58] and its potential application in cancer chemoprevention [59-61]. The major BBI type protein families in rice are Family 3328 (eight members) and Family 1493 (three members). While both families have the Pfam domain PF00228 (Bowman-Birk serine protease inhibitor family), Family 3328 also has a second domain identified via BLASTP [see Additional file 11]. Amino acid composition analysis showed that 31% and 47% of the conserved residues of Family 3288 and Family 1493, respectively, was cysteine suggesting that this amino acid has an important role in the protease inhibitory activity of BBI. These composition data also revealed subtle differences between the two BBI type protein families. The phylogenetic tree generated by MEGA version 3.1 [62] for family 3328 [see Additional file 12] suggests that after the original duplication event, only one of the paralogs underwent further rounds of duplication, consistent with the physical clustering of this set of BBI genes on chromosome 1 [see Additional file 13].

MPSS analysis showed that the BBI genes were differentially expressed in a wide range of tissues and organs, consistent with previously reported expression patterns [58]. Seven genes of Family 3328 were associated with unique, reliable, and significant MPSS tags with the pairwise Pearson's Correlation Coefficient values ranging from -0.35 to 0.71. Two genes within Family 1493 were associated with unique, reliable, and significant MPSS tags, which showed little correlation in expression (r = -0.12). It would be interesting to determine expression levels of the BBI genes following wounding, as seven proteins of the Family 3328 were annotated as Bowman-Birk type bran trypsin inhibitor precursors, a type

which was reported to play an important role in plant defense [54, 58], and two members of the Family 1493 were annotated as wound-induced BBI type WIP1 precursors [33].

Conclusions

We demonstrated that even relatively small plant genomes such as rice and Arabidopsis have a significant portion of their proteomes in paralogous families, resulting in a partially redundant proteome. The origin of most paralogous gene families in the rice genome seems to be very old, but duplicates have continued to arise at a fairly steady pace, with a peak in duplication being coincident with a major segmental duplication that took place at ~ 70 MYA. While conservation of protein domains was clearly observed within rice and Arabidopsis paralogous families, we did observe a major skew in types of proteins and protein domains within paralogous families versus singleton proteins, suggesting an impact of selection occurred during genome evolution and gene duplication. Another level of potential functionality in paralogous family proteins could also occur through alternative splicing which was statistically more frequent in paralogous family proteins compared to singletons in both rice as well as Arabidopsis. In rice, while some paralogous family members were transcriptionally co-regulated, divergence in expression patterns was clearly evident, thereby allowing an expanded range of functionality for the protein. These data suggested that multiple mechanisms are present in plant genomes to generate protein diversity and that these two model plant species share at least a subset of these mechanisms.

Methods

Construction of paralogous protein families

In release 4 of the TIGR Rice Genome Annotation [33], a total of 55,890 genes were annotated, of which 13,237 were related to TE. The TE-related genes were excluded from all further analyses. As alternative splicing occurs in the rice genome and some genes have multiple splice forms, the largest peptide sequence was used whenever alternative isoforms existed. Short protein sequences (<50 amino acids) were excluded from this analysis. A total of 42,653 rice protein sequences were used to classify paralogous protein families using protein domain compositions as described in Haas *et al*. [30]. The basic approach for generating the protein families involved identification of the domains followed by organization of the families based on domains. Two different types of domains were used for the generation of paralogous families: Pfam/HMM domains and BLASTP-based domains. For the Pfam/HMM domains, the predicted rice proteome was searched against the Pfam HMM domain database [63] using HMMER2 [64] and proteins with scores above the trusted cutoff value were retained. For the BLASTP-based domain, peptide regions that were not covered by the Pfam HMM profiles were then clustered based on homology derived from an all versus all BLASTP search [65]. Links were made if two peptides had an >45% identity over >75 amino acids with an E-value <0.001. To prevent multi-domain proteins that are not related from artificially clustering due to single linkages, the Jaccard coefficient of community [66], also known as link score, was used in the clustering process. As described in Haas *et al*. [30], a link score was calculated for the pairs of linked peptide sequences *a* and *b* as follows:

$$J_{a,b} = \frac{\#\,distinct\ sequences\ matching\ a\ and\ b\ including(a,b)}{\#\,distinct\ sequences\ matching\ either\ a\ or\ b}$$

Peptides with a link score above the cut-off value (0.66) were selected to generate single linkage clusters. Clustered peptides were then aligned using CLUSTALW [67, 68] and used to develop BLASTP-based domains, which were used to build the families if the domain alignments contained four or more members. Protein families were then organized based on the domain composition that refers to the type and number of the domains, which included both Pfam HMM domains and BLASTP-based domains. Proteins with identical domain composition were then classified into putative protein families. Paralogous protein families in Arabidopsis were constructed similarly with a total of 26,819 protein coding genes from the TAIR7 release of the predicted proteome [31].

Identification of segmentally duplicated genes

Segmentally duplicated genes in the rice genome were defined in Release 4 as described previously [69]. In brief, similar gene pairs were identified by all versus all BLASTP search (WU-BLASTP, parameters "V = 5 B = 5 E = 1e-10 -filter seg") [65], which were then used to define segmentally duplicated blocks by running DAGchainer [70] with parameters "-s -I –D 100000".

Functional classification of Arabidopsis proteome

A total of 26,819 Arabidopsis protein coding genes were downloaded from the TAIR7 release of the predicted proteome [31] and searched against an in-house non-redundant amino acid database that contains all publicly available protein sequences (e.g. GenBank, Swissprot, etc.) using BLASTP [65] and the Pfam HMM domain database [63] using HMMER2 [64]. BLASTP matches to Arabidopsis sequences were excluded unless they were from Swissprot. BLASTP matches to conserved hypothetical or hypothetical

proteins were excluded as well. Arabidopsis proteins with a BLASTP match (< 1e-10 and > 30% identity over 50% coverage) or Pfam domains with scores above the trusted cutoff value were classified as known or putative proteins. The remaining Arabidopsis genes were classified as expressed genes or hypothetical genes according to the gene set downloaded from TAIR7 release [31] which had at least one supporting cDNA and/or EST.

GOSlim assignment

To assign Gene Ontologies (GO) [71], the predicted rice proteome was searched against the predicted Arabidopsis proteome (TAIR6 Genome Release) [31] using BLASTP. Using an E-value cutoff of 1e-10, plant GOSlim annotations [34] were transitively annotated using the GO terms from Arabidopsis. Hypothetical/expressed proteins, TE-related proteins, and proteins assigned with GO terms with "unknown" definitions were excluded from this analysis. The GOSlim assignment of Arabidopsis proteins was obtained form TAIR7 release [31].

Identification of alternatively spliced genes

Approximately 780,000 rice EST sequences were released subsequent to the generation of the Release 4 gene models [33]. Thus, we utilized the PASA program [72] to re-annotate the gene models and comprehensively identify alternatively spliced genes with the latest set of rice transcript data. Alternative splicing information on Arabidopsis was obtained from TAIR7 release [31].

Estimation of the age of the paralogous protein families

A multiple protein sequence alignment was obtained for each family using CLUSTALW with default parameter settings [67, 68]. From each protein family of size *n*, all

(n^2-n)/2 pairwise alignments were extracted from the global family alignment, maintaining the position and length of all gaps. A maximum likelihood estimate of the number of synonymous substitutions per synonymous site (d_S) was obtained for all pairwise alignments. All calculations were performed using the codon-based substitution model of Goodman and Yang [73] implemented in *codeml*, of the PAML package, version 3.15 [74], running in pairwise mode (runmode = –2), with codon equilibrium frequencies estimated from average nucleotide frequencies at each codon position (codonFreq = 2).

The age of a paralogous protein family is defined by the duplication that gave rise to its second member, and can be approximated by the divergence between the most distantly-related pair of genes in the family. Given the rate of synonymous substitutions in grasses, estimated to be ~6.5×10^{-9} per site per year [75], the number of synonymous substitutions per site (d_S) between the most divergent gene pair in a family can be converted into a divergence time, provided synonymous sites are not saturated ($d_S < \sim 1$). In addition, peaks in the distribution of intra-family pairwise d_S values suggest periods of family diversification. For each family, the distribution of pariwise d_S values was determined, plotted within the range of 0 to 1.5, with bin size of 0.1. Both the modal bin of each distribution (usually resulting from the most ancient split in the family tree) and the largest modal value of $d_S <$ 1.5 (reflecting a burst in diversification within the last 100 MY) were recorded.

Massively parallel signature sequencing data and mapping

A total of 106,521 significant (>3 TPM) and reliable (observed in more than one sequencing run) MPSS [32] tags were obtained from the Rice MPSS Project [32, 76]. These MPSS tags are derived from nine treated or untreated organs/tissues including callus, leaf,

seed, crown vegetative meristematic tissue, ovary, stigma, pollen, panicle and stem. To reduce background noise, the method of Haberer *et al*. [77] was used to remove tags if the total minimal abundance across all libraries was ≤ 10 TPM or if the tag was not detected at ≥ 5 TPM in at least a single library, resulting in a total of 74,748 tags for subsequent analyses. The final set of MPSS tags were searched against TIGR rice pseudomolecules [33] using the Vmatch program [78]. As tags can span an intron(s), MPSS tags were also searched against all the cDNA sequences of the annotated genes. MPSS tags that mapped to the anti-sense sequence of the annotated genes or that mapped to multiple locations of the genome were excluded, which is important to minimize false correlations among closely related paralogs. If a gene was associated with multiple MPSS tags, only the most 3' tag was used for the expression analysis. Paralogous genes that were associated with unique, reliable, and significant MPSS tags were analyzed. Pearson's Correlation Coefficient (r) was calculated for each gene pair to determine the expression correlation using the following formula [79]:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}},$$

Where n is the number of DNA libraries. x_i and y_i represent the expression level of the gene pair in the i-th library.

Tissue specific expression analysis

To determine if a gene was preferentially expressed in a specific tissue, we employed the PEM devised by Huminiecki *et al* [80]. PEM is defined as $\log_{10}(O/E)$. Basically, it compares the observed (O) expression level in a given tissue with that of expected (E) level, assuming uniform expression across all tissues. The PEM value of the *i*-th gene in the *j*-th tissue was calculated as followed:

$$PEM_{i,j} = \log_{10}(x_{i,j} / (\sum_{k=1}^{m} x_{k,j} \sum_{l=1}^{n} x_{i,l} / \sum_{k=1}^{m} \sum_{l=1}^{n} x_{k,l}))$$

Where *m* and *n* represent the total number of MPSS-qualifying genes and tissues, respectively. $x_{i,j}$ is the expression level of the i-th gene in the j-th tissue.

Abbreviations

BBI: Bowman-Birk Inhibitor; EST: Expressed Sequence Tag;FLcDNA: Full Length cDNA; MPSS: Massively Parallel Signature Sequencing; MY: Million Years; MYA: Million Years Ago; PEM: Preferential Expression Measure; TE: Transposable Element; TPM: Transcripts Per Million;

Authors' contributions

HL designed the study, performed the analyses, and drafted the manuscript. SO participated in the analysis of GOSlim and made Additional file 3. KN and BM provided rice MPSS data. AE and JS carried out the age analysis of paralogous families. BH identified alternative splicing isoforms in rice. WZ identified the high confidence gene set in rice. XG participated in the analysis of alternative splicing. RB designed the study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Zhe Zhang for comments on statistical analyses. We thank Francoise Thibaud-Nissen for critical review of the article. This work was supported by a National Science Foundation Plant Genome Research Program grant to C. R. B. (DBI-0321538). The MPSS data were supported by NSF grant to B.C.M. (DBI-0321437).

References

1. Ohno S: **Evolution by Gene Duplication**: Springer-Verlag, New York; 1970.
2. Hughes AL: **The evolution of functionally novel proteins after gene duplication**. *Proc Biol Sci* 1994, **256**(1346):119-124.
3. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**(4):1531-1545.
4. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**(1):459-473.
5. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate genes in genetic robustness against null mutations**. *Nature* 2003, **421**(6918):63-66.
6. Grant D, Cregan P, Shoemaker RC: **Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis**. *Proc Natl Acad Sci U S A* 2000, **97**(8):4168-4173.
7. Settles AM, Baron A, Barkan A, Martienssen RA: **Duplication and suppression of chloroplast protein translocation genes in maize**. *Genetics* 2001, **157**(1):349-360.

8. Blanc G, Hokamp K, Wolfe KH: **A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome**. *Genome Res* 2003, **13**(2):137-144.

9. Dias AP, Braun EL, McMullen MD, Grotewold E: **Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication**. *Plant Physiol* 2003, **131**(2):610-620.

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

11. Leister D: **Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene**. *Trends Genet* 2004, **20**(3):116-122.

12. Leseberg CH, Li A, Kang H, Duvall M, Mao L: **Genome-wide analysis of the MADS-box gene family in Populus trichocarpa**. *Gene* 2006, **378**:84-94.

13. Song R, Llaca V, Linton E, Messing J: **Sequence, regulation, and evolution of the maize 22-kD alpha zein gene family**. *Genome Res* 2001, **11**(11):1817-1825.

14. Messing J, Dooner HK: **Organization and variability of the maize genome**. *Curr Opin Plant Biol* 2006, **9**(2):157-163.

15. Wang X, Shi X, Hao B, Ge S, Luo J: **Duplication and DNA segmental loss in the rice genome: implications for diploidization**. *New Phytol* 2005, **165**(3):937-946.

16. Vandepoele K, Simillion C, Van de Peer Y: **Evidence that rice and other cereals are ancient aneuploids**. *Plant Cell* 2003, **15**(9):2192-2202.

17. Simillion C, Vandepoele K, Saeys Y, Van de Peer Y: **Building genomic profiles for uncovering segmental homology in the twilight zone**. *Genome Res* 2004, **14**(6):1095-1106.

18. Paterson AH, Bowers JE, Chapman BA: **Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics**. *Proc Natl Acad Sci U S A* 2004, **101**(26):9903-9908.

19. Guyot R, Keller B: **Ancestral genome duplication in rice**. *Genome* 2004, **47**(3):610-614.

20. The Rice Chromosomes 11 and 12 Sequencing Consortia: **The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications**. *BMC Biol* 2005, **3**:20.

21. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.

22. Vij S, Tyagi AK: **Genome-wide analysis of the stress associated protein (SAP) gene family containing A20/AN1 zinc-finger(s) in rice and their phylogenetic relationship with Arabidopsis**. *Mol Genet Genomics* 2006.

23. Tripathi LP, Sowdhamini R: **Cross genome comparisons of serine proteases in Arabidopsis and rice**. *BMC Genomics* 2006, **7**:200.

24. Martinez M, Abraham Z, Carbonero P, Diaz I: **Comparative phylogenetic analysis of cystatin gene families from arabidopsis, rice and barley**. *Mol Genet Genomics* 2005, **273**(5):423-432.

25. Ito Y, Takaya K, Kurata N: **Expression of SERK family receptor-like protein kinase genes in rice**. *Biochim Biophys Acta* 2005, **1730**(3):253-258.

26. Abel S, Savchenko T, Levy M: **Genome-wide comparative analysis of the IQD gene families in Arabidopsis thaliana and Oryza sativa**. *BMC Evol Biol* 2005, **5**:72.

27. Yuan JS, Yang X, Lai J, Lin H, Cheng ZM, Nonogaki H, Chen F: **The Endo-beta-Mannanase gene families in Arabidopsis, rice, and poplar**. *Funct Integr Genomics* 2006.

28. Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T: **Genome cluster database. A sequence family analysis platform for Arabidopsis and rice**. *Plant Physiol* 2005, **138**(1):47-54.

29. Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskiewich R *et al*: **Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana**. *Genome Res* 2007, **17**(2):175-183.

30. Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Jr., Maiti R, Chan AP, Yu C, Farzad M, Wu D *et al*: **Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release**. *BMC Biol* 2005, **3**:7.

31. TAIR: **http://www.arabidopsis.org**.

32. The Rice MPSS Database: **http://mpss.udel.edu/rice/**.

33. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L *et al*: **The TIGR Rice Genome Annotation Resource:**

improvements and new features. *Nucleic Acids Res* 2007, **35**(Database issue):D883-887.

34. The Gene Ontology: **http://www.geneontology.org/GO.slims.shtml**.

35. Benjamini Y, Hochberg Y: **Controlling the false positive discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society* 1995, **Series B, 57**:289-300.

36. Blanc G, Wolfe KH: **Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution**. *Plant Cell* 2004, **16**(7):1679-1691.

37. Kopelman NM, Lancet D, Yanai I: **Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms**. *Nat Genet* 2005, **37**(6):588-589.

38. Su Z, Wang J, Yu J, Huang X, Gu X: **Evolution of alternative splicing after gene duplication**. *Genome Res* 2006, **16**(2):182-189.

39. Gu X: **Statistical framework for phylogenomic analysis of gene family expression profiles**. *Genetics* 2004, **167**(1):531-542.

40. Gu X, Zhang Z, Huang W: **Rapid evolution of expression and regulatory divergences after yeast gene duplication**. *Proc Natl Acad Sci U S A* 2005, **102**(3):707-712.

41. Gu Z, Nicolae D, Lu HH, Li WH: **Rapid divergence in expression between duplicate genes inferred from microarray data**. *Trends Genet* 2002, **18**(12):609-613.

42. Hughes AL, Friedman R: **Expression patterns of duplicate genes in the developing root in Arabidopsis thaliana**. *J Mol Evol* 2005, **60**(2):247-256.

43. Kreis M, Forde BG, Rahman S, Miflin BJ, Shewry PR: **Molecular evolution of the seed storage proteins of barley, rye and wheat**. *J Mol Biol* 1985, **183**(3):499-502.

44. Shewry PR, Tatham AS: **The prolamin storage proteins of cereal seeds: structure and evolution**. *Biochem J* 1990, **267**(1):1-12.

45. Shewry PR, Tatham AS, Halford NG: **The prolamins of the Triticeae**. *Shewry PR, Casey R, eds Seed proteins* 1999:35–78.

46. Leite A, Neto GC, Vettore AL, Yunes JA, Arruda P: **The prolamins of sorghum, Coix and millets**. *Shewry PR, Casey R, eds Seed proteins* 1999:141–157.

47. Krishnan HB, White JA: **Morphometric Analysis of Rice Seed Protein Bodies (Implication for a Significant Contribution of Prolamine to the Total Protein Content of Rice Endosperm)**. *Plant Physiol* 1995, **109**(4):1491-1495.

48. Barbier P, Ishihama A: **Variation in the nucleotide sequence of a prolamin gene family in wild rice**. *Plant Mol Biol* 1990, **15**(1):191-195.

49. Wen TN, Shyur LF, Su JC, Chen CS: **Nucleotide sequence of a rice (Oryza sativa) prolamin storage protein gene, RP6**. *Plant Physiol* 1993, **101**(3):1115-1116.

50. Mullins IM, Hilu KW: **Amino acid variation in the 10 kDa Oryza prolamin seed storage protein**. *J Agric Food Chem* 2004, **52**(8):2242-2246.

51. Ryan CA: **Proteinase inhibitors in plants: genes for improving defenses against insects and pathogens**. *Annu Rev Phytophathol* 1990, **28**:425-449.

52. Birk Y, Gertler A, Khalef S: **A pure trypsin inhibitor from soya beans**. *Biochem J* 1963, **87**:281-284.

53. Bowman DE: **Differentiation of soy bean anti-tryptic factors**. *Proc Soc Exp Biol Med* 1946, **63**:547-550.

54. Masumura T, Fujioka M, Matsui Y, Kumazawa Y, Tashiro M, Morita S, Tanaka K: **Cloning, expression and localization pattern of a trypsin inhibitor gene from rice**. *Plant & Animal Genomes XI Conference* 2003.

55. Odani S, Koide T, Ono T: **Wheat germ trypsin inhibitors. Isolation and structural characterization of single-headed and double-headed inhibitors of the Bowman-Birk type**. *J Biochem (Tokyo)* 1986, **100**(4):975-983.

56. Tanaka AS, Sampaio MU, Marangoni S, de Oliveira B, Novello JC, Oliva ML, Fink E, Sampaio CA: **Purification and primary structure determination of a Bowman-Birk trypsin inhibitor from Torresea cearensis seeds**. *Biol Chem* 1997, **378**(3-4):273-281.

57. Norioka S, Ikenaka T: **Amino acid sequence of trypsin chymotrypsin inhibitors (AI, AII, BI and BII) from peanut (*Arachis hypogaea*): a discussion on the molecular evolution of legume Bowman-Birk type inhibitors.** *J Biochem* 1983, **94**:589-599.

58. Qu LJ, Chen J, Liu M, Pan N, Okamoto H, Lin Z, Li C, Li D, Wang J, Zhu G *et al*: **Molecular cloning and functional analysis of a novel type of Bowman-Birk inhibitor gene family in rice**. *Plant Physiol* 2003, **133**(2):560-570.

59. Kennedy AR, Zhou Z, Donahue JJ, Ware JH: **Protection against adverse biological effects induced by space radiation by the Bowman-Birk inhibitor and antioxidants**. *Radiat Res* 2006, **166**(2):327-332.

60. Chen YW, Huang SC, Lin-Shiau SY, Lin JK: **Bowman-Birk inhibitor abates proteasome function and suppresses the proliferation of MCF7 breast cancer cells through accumulation of MAP kinase phosphatase-1**. *Carcinogenesis* 2005, **26**(7):1296-1306.

61. Dittmann KH, Mayer C, Rodemann HP: **Radioprotection of normal tissue to improve radiotherapy: the effect of the Bowman Birk protease inhibitor**. *Curr Med Chem Anticancer Agents* 2003, **3**(5):360-363.

62. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment**. *Brief Bioinform* 2004, **5**(2):150-163.

63. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**(Database issue):D138-141.

64. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

65. Gish W: **http://blast.wustl.edu**. In.; 1996 - 2006.

66. Jaccard P: **The Distribution of the Flora in the Alpine Zone**. *The New Phytologist* 1912, **11**(2):37-50.

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

68. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**(13):3497-3500.

69. Lin H, Zhu W, Silva JC, Gu X, Buell CR: **Intron gain and loss in segmentally duplicated genes in rice**. *Genome Biol* 2006, **7**(5):R41.

70. Haas BJ, Delcher AL, Wortman JR, Salzberg SL: **DAGchainer: a tool for mining segmental genome duplications and synteny**. *Bioinformatics* 2004, **20**(18):3643-3646.

71. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

72. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucleic Acids Res* 2003, **31**(19):5654-5666.

73. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences**. *Mol Biol Evol* 1994, **11**(5):725-736.

74. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput Appl Biosci* 1997, **13**(5):555-556.

75. Gaut BS, Morton BR, McCaig BC, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene**

Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci U S A* 1996, **93**(19):10274-10279.

76. Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC: **Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA**. *Nucleic Acids Res* 2006, **34**(Database issue):D731-735.

77. Haberer G, Hindemitt T, Meyers BC, Mayer KF: **Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis**. *Plant Physiol* 2004, **136**(2):3009-3022.

78. Kurtz S: **The Vmatch large scale sequence analysis software (http://www.vmatch.de/)**.

79. Rosner B: **Fundamental of Biostatistics**, 4th edn: Duxbury Press; 1995.

80. Huminiecki L, Lloyd AT, Wolfe KH: **Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases**. *BMC Genomics* 2003, **4**(1):31.

Additional data files

Additional data files are provided with the online version of the paper "Characterization of paralogous protein families in rice", *BMC Plant Biology* 2008, 8:18. The additional data files can be downloaded from http://www.biomedcentral.com/1471-2229/8/18/additional/.

Additional file 1

Description: Putative paralogous protein families within the rice genome.

Additional file 2

Description: Rice paralogous protein families with more than one hundred member proteins.

Additional file 3

Description: Distribution of non-transposable element-related genes in rice and Arabidopsis. In panel A, the 12 rice chromosomes are shown with paralogous gene family members plotted in blue while single copy genes are plotted in red. Segmental duplicated blocks are indicated in green and centromeres are denoted by a white box. In panel B, the five Arabidopsis chromosomes are shown with paralogous gene family members plotted in blue while single copy genes are plotted in red.

Additional file 4

Description: The age distribution of rice paralogous protein families. **A)** an expanded view of the age distribution. **B)** the enlarged distribution of rice paralogous protein families with largest *d*s ≤ 1.5.

Additional file 5

Description: Distribution of modal values under $d_S = 1.5$ across rice paralogous protein families. Of all 3,865 paralogous protein families, 2,388 showed a peak under 1.5 in the distribution of all pairwise d_S values and are plotted.

Additional file 6

Description: Pearson's correlation coefficient (*r*) versus *ds* values. **A)** $0 < d_S \leq 0.1$; **B)** $0.4 < d_S \leq 0.5$; **C)** $1.0 < d_S \leq 1.1$; **D)** $1.4 < d_S \leq 1.5$.

Additional file 7

Description: Schematic illustration of the domain composition of three related rice paralogous protein families: Family 3722, Family 3193, and Family 3856.

Additional file 8

Description: Expression abundance of the rice prolamin genes from Family 3722 and Family 3193 in 18 libraries which were associated with unique, reliable, and significant MPSS tags.

Additional file 9

Description: Expression abundance of genes from rice paralogous protein family Family 3856 (contained PF00234) in 18 libraries which were associated with unique, reliable, and significant MPSS tags.

Additional file 10

Description: Genome Browser view of the genes encoding rice prolamin proteins with TE-related genes inserted between putative tandem duplications.

Additional file 11

Description: Schematic illustration of the domain composition of two rice BBI-related paralogous protein families which have Pfam domain PF00228: Family 3328 and Family 1493.

Additional file 12

Description: Neighbor-Joining tree of the rice Bowman-Birk inhibitor protein family Family 3328.

Additional file 13

Description: Browser view of the rice genes encoding BBI proteins on chromosome 1.

Table 1. Two-sample binomial tests for GOSlim assignments of paralogous family and singleton proteins in rice

GOSlim assignment[a]	Singletons (%)	Paralogous genes (%)	P-value[d]
Binding, other[b]	3.3	6.5	<1e-5
Carbohydrate binding[c]	2.7	0.6	<1e-5
DNA binding[b]	4.8	8.0	<1e-5
Hydrolase activity[b]	7.8	12.7	<1e-5
Kinase activity[c]	16.0	6.2	<1e-5
Nucleotide binding[c]	13.4	4.2	<1e-5
Protein binding, other[c]	14.2	9.5	<1e-5
Receptor activity[c]	2.3	0.4	<1e-5
Transcription factor activity[b]	4.3	9.3	<1e-5
Catalytic activity, other[b]	8.7	12.2	<1e-5
Structural molecule activity[b]	0.8	2.2	<1e-5
Oxygen binding[b]	0.7	1.9	<1e-5
Transcription regulator activity[b]	1.1	2.3	<1e-5
Transporter activity[b]	5.0	7.0	<1e-5
Lipid binding[b]	0.4	1.1	<1e-5
Molecular function, other[b]	0.1	0.4	0.001
Enzyme regulator activity[b]	0.5	0.9	0.008
Motor activity	0.5	0.3	0.051
Transferase activity	7.0	7.7	0.095
Receptor binding	0.0	0.1	0.137
RNA binding	1.8	2.1	0.369
Translation factor activity, nucleic acid binding	0.5	0.7	0.353
Signal transducer activity	1.0	0.9	0.43
Chromatin binding	0.3	0.2	0.465
Nucleic acid binding, other	1.8	1.9	0.882
Nuclease activity	0.8	0.8	0.888

[a] GoSlim assignment classifications were performed as described in the Materials and Methods.

[b] Enrichment of GOSlim annotations in paralogous protein families compared to singletons.

[c] Reduction of GOSlim annotations in paralogous protein families compared to singletons.

[d] Benjamini and Hochberg correction for multiple testing.

Table 2. Two-sample binomial tests for GOSlim assignments of paralogous family and singleton proteins in Arabidopsis

GOSlim assignment[a]	Singletons (%)	Paralogous genes (%)	P-value[d]
Hydrolase activity[b]	7.5	12.6	<1e-5
Kinase activity[c]	10.4	5.5	<1e-5
Nucleotide binding[c]	10.2	4.6	<1e-5
Protein binding, other[c]	12.9	8.2	<1e-5
Transcription factor activity[b]	4.2	9.0	<1e-5
Receptor activity[c]	1.9	0.7	<1e-5
DNA binding[b]	4.1	7.2	<1e-5
Oxygen binding[b]	0.1	1.4	<1e-5
Receptor binding[c]	0.5	0.1	<1e-5
Carbohydrate binding[c]	0.7	0.3	<1e-3
Lipid binding[b]	0.3	0.8	0.001
Structural molecule activity[b]	1.6	2.5	0.002
Enzyme regulator activity[b]	0.7	1.4	0.005
Molecular function, other[b]	1.8	2.5	0.011
Transporter activity[b]	5.0	6.0	0.019
Nucleic acid binding, other[c]	2.6	2.0	0.027
Motor activity[b]	0.2	0.5	0.03
Transferase activity	5.3	6.1	0.053
RNA binding	1.5	1.9	0.099
Binding, other	12.3	11.3	0.102
Signal transducer activity	1.0	0.8	0.132
Catalytic activity, other	12.4	11.7	0.244
Transcription regulator activity	1.3	1.5	0.743
Chromatin binding	0.2	0.1	0.803
Translation factor activity, nucleic acid binding	0.6	0.6	1
Nuclease activity	0.7	0.8	1

[a] GoSlim assignment classifications were performed as described in the Materials and Methods.

[b] Enrichment of GOSlim annotations in paralogous protein families compared to singletons.

[c] Reduction of GOSlim annotations in paralogous protein families compared to singletons.

[d] Benjamini and Hochberg correction for multiple testing.

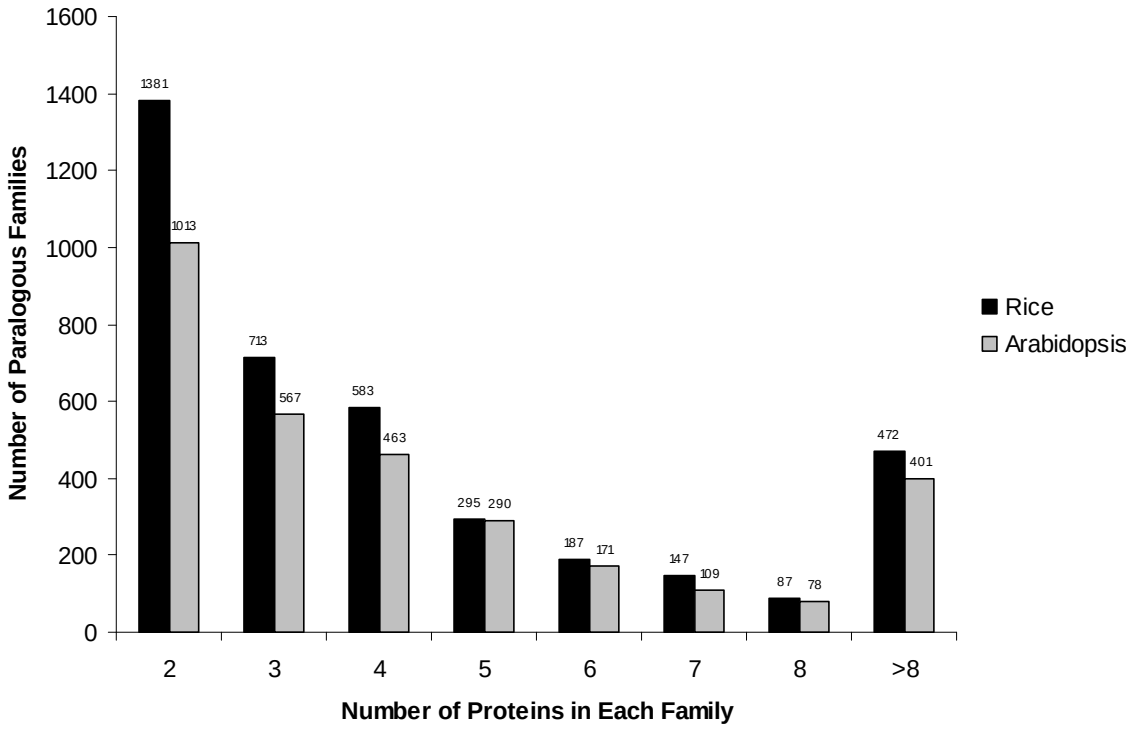


Figure 1. Size distribution of paralogous protein families in rice and Arabidopsis. The exact number of families is listed above the bars.

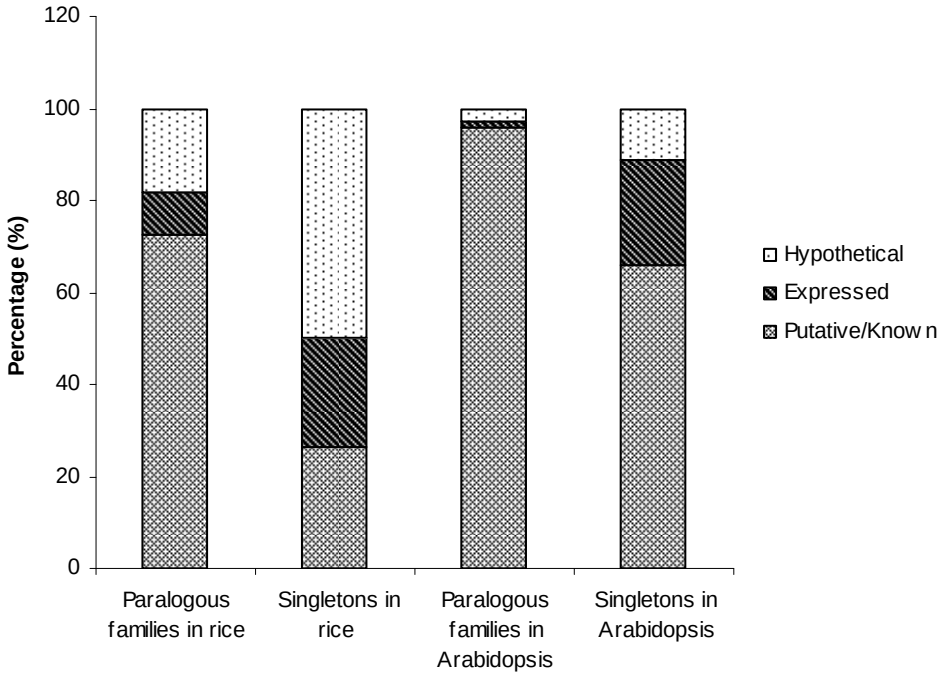


Figure 2. Functional classification of paralogous family and singleton proteins in rice and Arabidopsis.

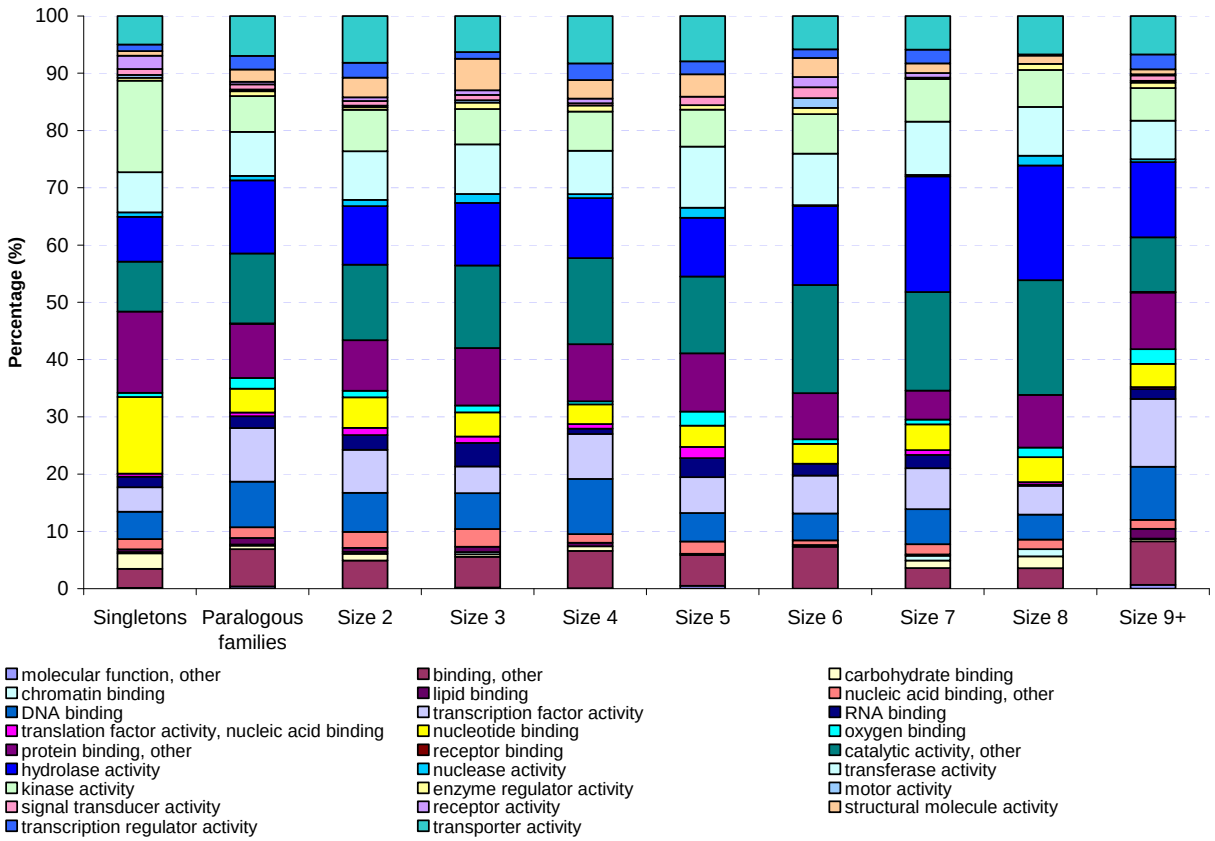


Figure 3A. GOSlim assignment of rice paralogous families and singletons. The paralogous protein families are further classified by family size.

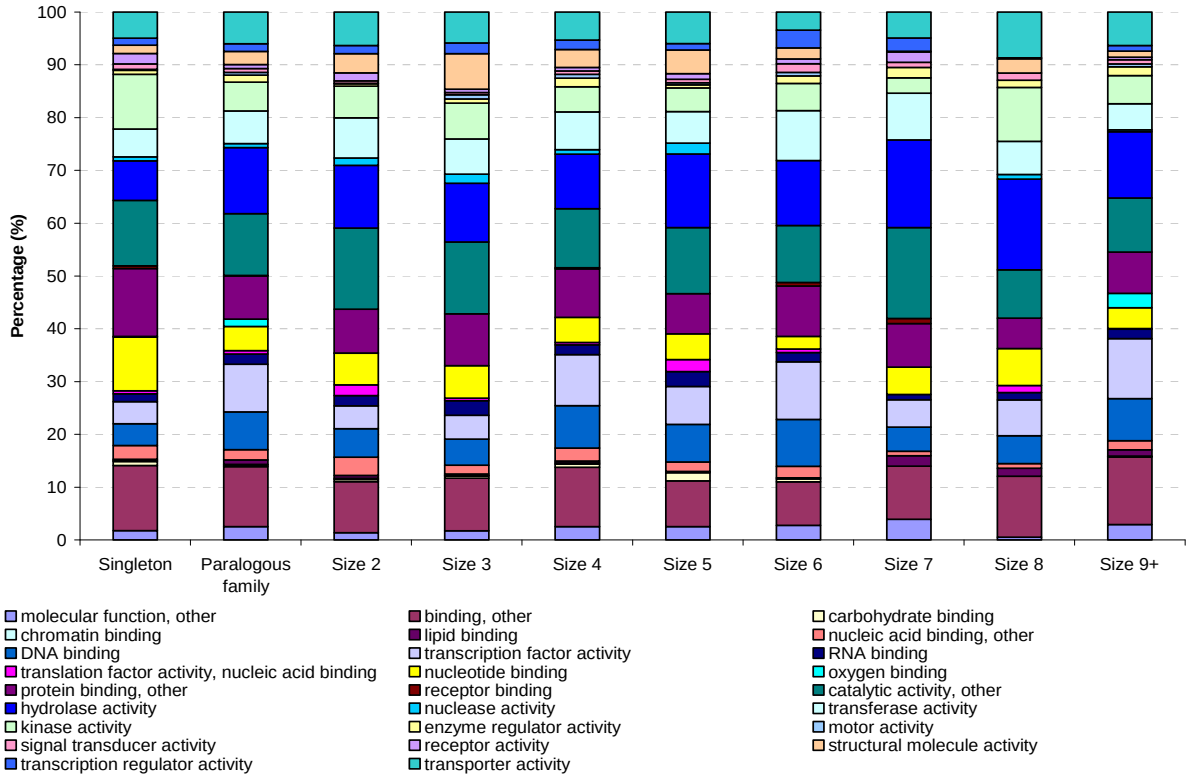


Figure 3B. GOSlim assignment of Arabidopsis paralogous families and singletons. The paralogous protein families are further classified by family size.

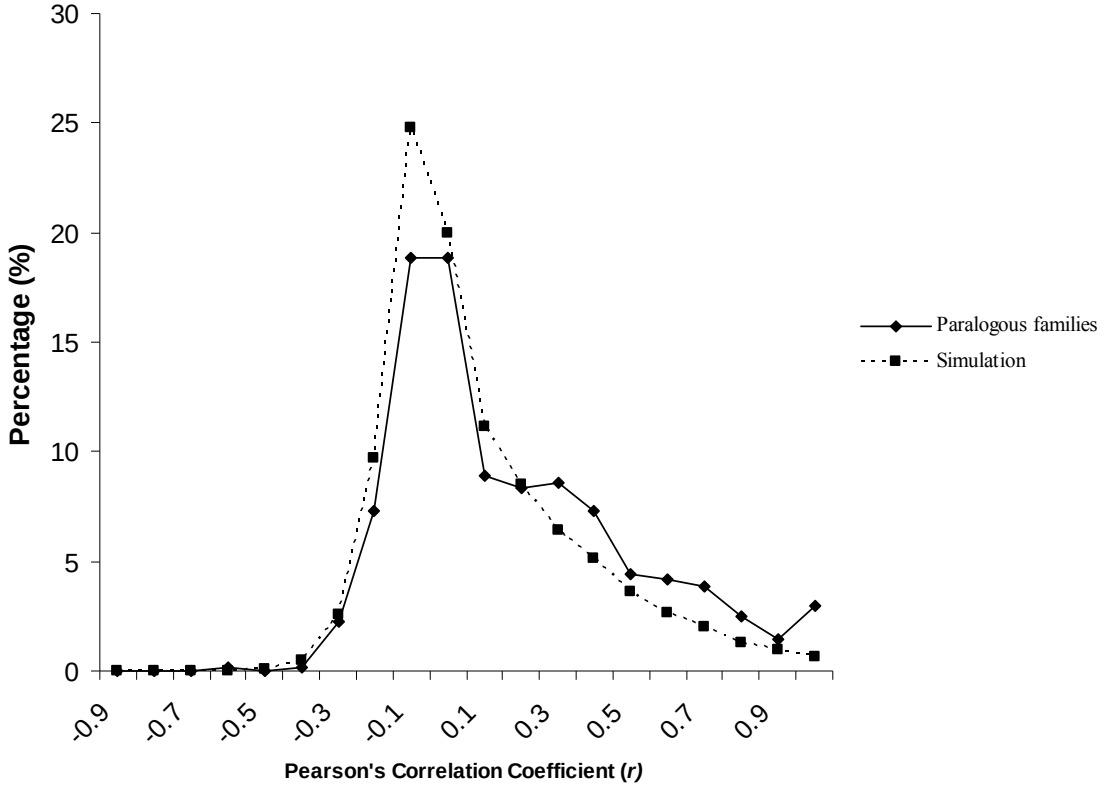


Figure 4. Histogram of Pearson's Correlation Coefficients of expression (*r*) of rice paralogous protein families with exactly two MPSS-qualifying genes.

CHAPTER 4. COMPARATIVE ANALYSES REVEAL DISTINCT SETS OF LINEAGE SPECIFIC GENES WITHIN *ARABIDOPSIS THALIANA*

A paper to be submitted to *BMC Evolutionary Biology*

Haining Lin, Gaurav Moghe, Shu Ouyang, Shin-han Shiu, Xun Gu, C. Robin Buell

Abstract

Background

The availability of genome and transcriptome sequences for a number of species permits the identification and characterization of conserved as well as divergent genes such as lineage-specific genes which have no detectable sequence similarity to genes from other lineages. While genes conserved among taxa provide insight into the core processes among species, lineage specific genes provide insights into evolutionary processes and biological functions that are clade or species specific.

Results

Using comparative sequence analyses between *Arabidopsis thaliana* and 156 plant species, we identified a set of 24,554 (91.4%) Arabidopsis genes that are evolutionarily conserved (EC) within the Plant Kingdom as defined by sequence similarity with at least one of 147 plant species and two sets of genes that are restricted in their distribution within the Plant Kingdom. The Conserved Brassicaceae-Specific Genes (957, 3.6%, CBSG) share sequence similarity only to sequences within the Brassicaceae family while the Arabidopsis Lineage-Specific Genes (1,351, 5.0%), ALSG) lack sequence similarity to any sequence outside *A. thaliana*. While the CBSGs (76.4%) and ALSGs (53.6%) are transcribed, the

majority of the CBSGs (75.7%) and ALSGs (93.9%) have no known function making them an enigma within the Arabidopsis genome. Co-expression analysis of CBSGs and ALSGs indicated enrichment in 19 and 13 of 27 FunCat categories, repectively, suggesting a wide range of biological functions within the cell. Subcellular localization prediction revealed that CBSGs were significantly enriched in secretory proteins (441, 46.1%) and among the 119 putatively secreted CBSGs with a known function, 72 encode putative pollen coat proteins or S locus cysteine-rich proteins which are involved in the self-incompatibility response. Single Nucleotide Polymorphism (SNP) analysis showed an elevated ratio of non-synonymous to synonymous SNPs within the ALSGs (2.01) and CBSGs (1.65) relative to the EC set (0.92), mainly caused by an elevated number of non-synonymous SNPs, indicating that they are fast-evolving at protein sequence level.

Conclusions

Our analyses suggest that while a significant fraction of the Arabidopsis proteome is conserved within the Plant Kingdom, evolutionarily distinct sets of genes have arisen in the Brassicaceae and Arabidopsis that may function in defining biological processes unique to these lineages.

Background

Lineage-specific genes are genes that have no detectable sequence similarity to genes from other lineages. It has been reported that lineage-specific genes evolve rapidly and as a consequence, no sequence similarity to genes from other species can be detected [1]. With the availability of complete or near-complete genome and transcriptome sequences from a

wide range of species, lineage-specific genes have been extensively studied, especially for microbial species [2-5]. Several hypotheses regarding the origin of lineage-specific genes have been proposed. One model suggests that lateral gene transfer has an important role in generating lineage-specific genes that are not shared by closely related species [6, 7]. The second model proposes that lineage-specific genes may be generated by gene duplication followed by rapid sequence divergence [5, 8]. Other models include *de novo* emergence from non-coding sequences which are more diverged between species [9], differential gene loss of duplicated genes [10], and possible artifacts from genome annotation [11]. Although the origin and evolution of lineage-specific genes remains unresolved, the identification and characterization of putative lineage-specific genes provide insight into evolutionary processes such as speciation (divergence) and adaptive variance [5]. In addition, they provide insight into biological functions that are species specific.

Within the Plant Kingdom, identification and characterization of lineage-specific genes has been performed through comparative analysis of the Expressed Sequence Tags (ESTs) and/or the finished genomic sequences of Arabidopsis and rice [12-14], the model species for dicotyledonous and monocotyledonous plants. For example, a comparative analysis of unigene sets from legumes to unigene sets from non-legumes, GenBank's nonredundant and EST databases, and the genome sequences of Arabidopsis and rice revealed that approximately 6% of the legume unigene sets were legume-specific [14]. In a more recent analysis, a set of 861 rice genes termed "Conserved Poaceae Specific Genes" that are evolutionarily conserved within the Poaceae family yet lack significant sequence similarity to non-Poaceae species was identified by searching the finished rice genome

sequence against the genomic sequences from Arabidopsis, Medicago, Poplar, and EST clusters from 184 plant species [15]. This set of conserved Poaceae-specific genes provides a starting point for further research experiments to better understand the unique morphology, physiological and developmental characteristics of the Poaceae. Despite the ongoing comparative studies, to the best of our knowledge, only one large-scale comparative genomic analysis of Arabidopsis and rice has been performed to date, in which 116 clusters that have at least two Arabidopsis sequences but no rice sequence were identified as Arabidopsis-specific genes [16, 17].

In this study, we identified and characterized Conserved Brassicaceae-Specific Genes (CBSGs) and Arabidopsis Lineage-Specific Genes (ALSGs) using the completed and well-annotated Arabidopsis genome, the genomes of *Medicago truncatula* (medicago), *Populus trichocarpa* (poplar), *Vitis vinifera* (grapevine), and *Oryza sativa* (rice) [18-21], and EST clusters from 156 plant species. By our definition, CBSGs are Arabidopsis genes that have significant sequence similarity only to sequences within the Brassicaceae family while ALSGs are Arabidopsis genes that are specific to *A. thaliana*. As a large portion of CBSGs and ALSGs have no known function, co-expression and subcellular localization analyses were performed to infer possible biological function. To assess evolutionary pressures within these two sets of genes, Single Nucleotide Polymorphisms (SNPs) within the coding region of CBSGs and ALSGs were analyzed.

Results

Identification of CBSGs and ALSGs

Using TBLASTN, a total of 26,862 Arabidopsis protein-coding genes were searched against the genomic sequences of poplar, medicago, grapevine, rice, and the PlantGDB-assembled Unique Transcripts (PUT) [22] from 147 species outside the Brassicaceae family. A total of 24,483 Arabidopsis genes with significant sequence similarity (E-value < 1e-5) to either genomic or PUT sequences from species outside the Brassicaceae were defined as the Evolutionarily Conserved (EC) set (Fig. 1). The remaining 2,379 Arabidopsis genes with no significant sequence similarity to any sequence (genomic or PUT) outside the Brassicaceae were further searched against PUT sequences from nine species within the Brassicaceae family: *Brassica napus*, *Brassica oleracea*, *Brassica oleracea* var. *alboglabra*, *Brassica rapa, Raphanus raphanistrum* subsp *landra, Raphanus raphanistrum* ssp*. maritimus, Raphanus raphanistrum* ssp*. raphanistrum, Raphanus sativus*, and *Raphanus sativus* var*. oleiformis*. This resulted in two datasets: 970 CBSGs with no significant sequence similarity to sequences from the Plant Kingdom except species from the Brassicaceae, and 1,409 ALSGs that had no significant sequence similarity to any sequences within the Plant Kingdom (Fig. 1). To further eliminate false positives due to incompleteness of the genomic and transcriptome sequence sets, the CBSGs and ALSGs were searched against the UniProt Knowledgebase (UniProtKB) using BLASTP. Manual inspection of the alignments (E-value < 1e-5) identified 71 Arabidopsis genes (48 CBSGs and 23 ALSGs) with similarity to non-Brassicaceae UniProt entries and were removed from the CBSG and ALSG sets to the SH set. A total of 35 ALSGs with similarity to Brassicaceae UniProt entries and were also removed from the ALSG set to CBSG set. Thus, the final sets of CBSGs, ALSGs, and ECs contain 957, 1,351, and 24,554 Arabidopsis genes, respectively (Fig. 1). The identification of

CBSGs and ALSGs enables us to investigate genes that may have been involved in the adaptation of the Brassicaceae family and Arabidopsis, respectively. The EC gene set serves as reference to compare and contrast characteristics of CBSGs and ALSGs.

Characterization of the CBSGs and ALSGs

To discern whether there are significant differences in the genic features between the three gene sets (CBSGs, ALSGs, ECs) and Transposable Element (TE)-related genes, genic features of CBSGs and ALSGs were characterized and compared to those of the EC and TE gene sets (Table 1). The average exon numbers per gene for CBSGs (2.2) and ALSGs (1.7) were similar to that of the TE set (1.7), but smaller than that of the EC gene set (5.5), consistent with previous finding of shorter gene size of lineage-specific genes in rice [15, 23]. A total of 345 (36.1%) CBSGs, 875 (64.8%) ALSGs, and 4,682 (19.1%) ECs were single-exon genes. However, the average exon length of CBSGs and ALSGs was one fifth of that of the TE set (Table 1). The average intron length of CBSGs and ALSGs was slightly longer than that of the EC and TE sets. CBSGs had a lower average GC content (37.8%) for the whole gene while ALSGs had higher average GC content (41.0%) similar to that of ECs (39.6%) and TE genes (41.5%). Both CBSGs and ALSGs had lower average GC coding sequence content compared to the EC set, with CBSGs having the lowest GC content. The lower GC content observed for CBSGs was consistent with pervious report on lower GC content of lineage-specific genes than non-lineage-specific genes of Drosophila [5]. Overall, both CBSGs and ALSGs seemed distinct gene sets from EC and TE-related gene sets.

With respect to function, both the CBSG and ALSG gene sets are enriched in genes of unknown function with 724 CBSGs (75.7%) and 1,269 ALSGs (93.9%) encoding proteins

with no known or putative function (Table 2). However, a large portion of both the CBSG and ALSG gene sets had transcript support from ESTs, cDNAs, or microarray data, increasing the confidence these are bona fide genes and not annotation artifacts (Table 2). A total of 73 CBSGs (7.6%) encode LCR (low-molecular-weight, cysteine-rich) genes and SCRL (S locus cysteine-rich) genes, members of a family of small, secreted, cysteine rich protein with sequence similarity to members in the PCP (pollen coat protein) gene family or SCR (S locus cysteine-rich protein), respectively [24]. Both PCP and SCR genes are involved in the self-incompatibility (SI) response which prevents selfing through interruption of pollen recognition, pollen tube elongation, ovule fertilization, or embryo development [25-28]. Upstream open reading frames (uORF) are small open reading frames present in the 5' UTR of the mature mRNA which occur in 15~40% of eukaryotic transcriptomes [29, 30]. uORF appeared to be involved in post-transcriptional regulation of a main coding sequence [31-34] and 10 CBSGs (1.0%) and 20 ALSGs (1.5%) are uORFs compared to 36 (0.1%) uORFs in the EC set. Intriguingly, a recent study identified over 200 well conserved uORFs between the human and mouse genomes which showed evidence of purifying selection, suggesting that uORFs have other biological functions than simple cis-acting post-transcriptional regulation [35].

Neither the CBSGs nor ALSGs were distributed randomly within the Arabidopsis genome (Additional data file 1). Although large numbers of CBSGs, ALSGs, and EC were located within segmentally duplicated blocks consistent with the substantial segmental duplication that occurred in Arabidopsis [36], the CBSGs and ALSGs were located more frequently in non-segmentally duplicated regions than ECs. A total of 23.4% CBSG and

27.0% ALSG genes, respectively, were located within non-segmentally duplicated regions, compared to 13.8% EC genes (χ^2 test, P < 1e-5). This could be due to differential gene loss of lineage-specific genes (ALSGs, CBSGs) and ECs in segmentally duplicated versus non-segmentally duplicated regions.

In Arabidopsis, a total of 17,911 genes were classified within 3,051 paralogous families (66.7 %) using a computational pipeline that utilizes Pfam and novel BLASTP-based protein domains (see Methods). The identification of novel BLASTP-based domain during the paralogous family classification pipeline allows proteins without a Pfam domain to be classified into paralogous families thereby removing any bias associated with lack of a characterized protein domain. A total of 424 CBSGs (44.3%) were classified within paralogous families while only 66 ALSGs (4.9%) were classified within paralogous families. The percentage of CBSGs within paralogous families is substantially lower than that of the EC set (70.9%), yet consistent with what has been reported for lineage-specific genes within Poaceae [15]. These data are also consistent with previous analyses that showed paralogous families in Arabidopsis are enriched in genes with known function whereas the single copy gene complement in Arabidopsis is enriched in genes with no known function [37].

Functional inference by co-expression analyses

Given the lack of functional assignment for a large percentage of the ALSG and CBSG set, we performed co-expression analysis to associate them to genes with Functional Catalogue (FunCat) annotation which consists of 27 top-most categories that cover functional annotations such as metabolism, information pathways, and perception and response to stimuli [38]. Using expression data from 3,037 *A. thaliana* ATH1 arrays, we computed

Pearson's Correlation Coefficient for the ALSGs and CBSGs in comparison to all other genes on the microarray. Probes for 358 (26%) ALSGs and 328 (34%) CBSGs are present on the ATH1 array slide. Using a cutoff value of 0.6 for the correlation coefficient, we found that a total of 291 ALSGs (81%) and 273 CBSGs (83%) were co-expressed with at least one other gene. A Fisher Exact Test, performed using the top two levels of the FunCat categories with a False Discovery Rate of 5% revealed that 203 out of the 291 ALSGs (70%) and 205 out of the 273 CBSGs (75%) had at least one enriched FunCat category.

After assigning FunCat annotations to the 203 ALSGs and 205 CBSGs, we performed further enrichment analysis to see which categories contain over-represented numbers of ALSGs or CBSGs. This analysis, performed with only the top-most FunCat categories, indicated that both ALSGs and CBSGs were enriched in 13 and 19 of the 27 possible FunCat main categories, respectively (Fig. 2). Most of the genes, 138 of the 189 CBSGs and 153 of the 192 ALSGs, are enriched in the category "Metabolism" (FunCat 01). Over 50% of the genes associated with this category are also enriched in the sub-categories related to carbon and phosphorus metabolism (FunCat 01.05 and FunCat 01.04). This is interesting in the context of the knowledge that ~82% of the 146 CBSGs and ~70% of the 159 ALSGs in the category "Sub-cellular localization" (FunCat 70) are associated with the plastids (FunCat 70.26), and more specifically, chloroplasts (data not shown), which are the sites of carbon fixation and ATP synthesis.

A high number of the CBSGs (60%) are enriched in the function "Interaction with the environment" (FunCat 34), as compared to only 50% of the ALSGs. This category includes functions that function in responses to the environment immediate to the cellular membranes.

Similarly, 75 of the 105 genes in the category “Proteins with binding functions or co-factor requirements” (FunCat 16) are enriched in the sub-category “Protein binding”(FunCat 16.01). This sub-category includes only the genes involved in “Receptor binding” (FunCat 16.01.01). These observations, despite being obtained from a small subset of the lineage specific genes, are still consistent with the predictions of TargetP, of an enrichment of secretory proteins among the CBSGs. A substantial proportion of genes are also enriched in functions related to “Protein synthesis” (FunCat 12, 38% CBSGs), “Nucleotide binding” (FunCat 16.19, 30% CBSGs) and “Protein binding” (FunCat 16.01, 40% CBSGs), suggesting a role in regulatory activities within the cell.

CBSGs are enriched with secretory proteins

To provide additional levels of functional annotation of CBSGs and ALSGs, TargetP was used to deduce the subcellular localization of the predicted Arabidopsis proteome [39]. TargetP determines the putative subcellular localization based on N-terminal amino acid sequences: chloroplast transit peptide, mitochondrial targeting peptide, or secretory pathway signal peptide. The abundance of genes at the whole genome level predicted to be targeted to chloroplast, mitochondria, and secretory pathway were 14.9%, 11.7%, and 20.2%, respectively, comparable to a previous report [40]. A dramatic enrichment of secretory proteins was observed in the CBSG set (46.1%), among which, only 119 (27.0%) have a putative function (Table 3). Examination of the TAIR8 assigned functions of these 119 CBSGs suggested that 94 of them were likely targeted to secretory pathway: 72 secreted proteins similar to PCP/SCR; ten defensin-like family proteins; four putative ligands; and eight Rapid Alkalinization Factor (RALF)-like proteins. As proteins involved in the secretory

pathway (e.g., receptor-ligand signaling proteins, transporters, and extracellular signaling proteins) play fundamental roles in plant development, the finding that the majority of the secreted CBSGs have no known function suggests that Brassicaceae species possess one or more biological processes that are specific to the Brassicaceae family or have diverged significantly from species outside the Brassicaceae family. No bias was seen in the ALSG set for secreted proteins.

A lower percentage of CBSGs and ALSGs were predicted to be targeted to the chloroplast compared to the SH or whole proteome set (as annotated in TAIR8), consistent with the notion that proteins involved in chloroplast function are highly conserved throughout the Plant Kingdom. This same bias was seen in the CBSG set for proteins targeted to the mitochondrion. Surprisingly, the percentage of genes targeted to the mitochondrion differed between the CBSG, ALSG, and EC sets (6.9%, 17.3%, and 11.5%, respectively). However, while the percentages differ, the sheer number of genes with proteins targeted to the mitochondria within the CBSG and ALSG sets are a mere fraction of those within the EC set (CBSG: 66; ALSG: 234; EC: 2,833). The majority of mitochondrial and chloroplast targeted ALSGs and CBSGs have no known function in sharp contrast to the EC set (Table 3) suggesting these encode novel and potentially evolutionarily distinct functions within these two organelles.

CBSGs and ALSGs have a higher ratio of non-synonymous to sysnonymous SNPs

A total of 249,344 SNPs were used to assess the genetic variation of the CBSGs and ALSGs among 20 Arabidopsis ecotypes [41]. A total of 243,963 (97.8%) SNPs, which showed a single variation were used to calculate the number of genes with SNPs within the

coding regions as well as the frequencies of synonymous and non-synonymous SNPs. Compared to EC genes, both the ALSGs and CBSGs had a smaller percentage of genes with SNPs within the coding region. A total of 683 (50.6%), 615 (64.3%), and 21,136 (86.1%) genes from ALSG, CBSG, and EC set, respectively, showed one or more SNPs within their coding regions. Taking the length of the coding sequence into account, both ALSGs (0.45) and CBSGs (0.43) had more SNPs per 100 bp than EC genes (0.34). Further investigation of synonymous and non-synonymous SNPs showed that the elevated number of SNPs per 100 bp in ALSGs and CBSGs is mainly due to the elevated number of non-synonymous SNPs per 100 bp (Fig. 3A) with the number of non-synonymous SNPs per 100 bp higher in ALSGs (0.30) and CBSGs (0.27) than ECs (0.16) while the number of synonymous SNPs per 100 bp is similar among ALSGs (0.15), CBSGs (0.16), and ECs (0.18). A total of 424 (31.4%), 345 (36.1%), and 8,670 (35.3%) genes from ALSG, CBSG, and EC set, respectively, had more non-synonymous SNPs than synonymous SNPs. Among them, ALSGs and CBSGs had greatly elevated ratios of non-synonymous to synonymous SNPs compared to the EC set (Fig. 3A). Our results indicate that approximately one third of ALSGs and CBSGs may be fast-evolving genes. As CBSGs were enriched in secreted genes and secreted proteins are likely to have fundamental function in plant development, SNP analysis was performed for putative secreted ALSGs, CBSGs, and ECs. The pattern of SNP density of secreted genes is similar to all genes within the ALSG, CBSG, and EC sets (Fig. 3B), suggesting that both secreted and non-secreted genes in ALSGs and CBSGs have an elevated non-synonymous SNP density compared to the ECs.

Discussion

The 957 CBSGs and 1,351 ALSGs identified in this study are attractive targets for experimental discovery as they are lineage-specific in nature and the majority (75.7% CBSGs and 93.9% ALSGs) encodes functions yet to be determined. Both CBSGs and ALSGs had shorter genes compared to the EC set, primarily due to fewer numbers of exons per gene and higher percentage of single-exon genes. A total of 68.6% of the 26,862 Arabidopsis genes that have been used in our analyses are high confidence genes in that gene structure (including splice junctions) of at least one or more isoforms of the gene has been confirmed with a single cDNA or multiple overlapping cDNAs [42]. The percentages of high confidence genes within ALSG, CBSG, and EC sets are 19.1%, 37.9%, and 72.5%, respectively. Furthermore, 84.1% and 54.8% of CBSGs and ALSGs, respectively, have transcript evidence from full length-cDNA, ESTs or microarray data, or have a putative function assigned, which provides strong support that they are likely to be bona fide genes rather than false positive gene predictions from the *ab initio* gene prediction programs utilized in genome annotation processes.

The dramatic enrichment of secretory proteins in CBSGs indicates there may be specific or highly evolved secretion processes within Brassicaceae species as significant sequence similarity could not be detected in other sequenced angiosperms including the dicots poplar and grapevine for which genome sequences are available. Previous studies suggested that a large percentage of Brassicaceae species or accessions are self-incompatible [43, 44] and are able to recognize and reject self-pollen or pollen from closely related plants. Specificity of the SI response is genetically determined by the alleles at the S (self incompatibility) locus and involves arrest of pollen development upon self pollination [45].

A total of 72 of the 119 CBSGs with a putative function are predicted by TargetP to be involved in secretory pathways and are similar to SCR or PCP proteins. SCR is the male component of the SI response which is expressed specifically in the anther tapetum and microspores [46] and is predicted to interact with the female component S locus receptor kinase gene expressed in the papillar cells of the stigma [47]. Members of the PCP gene family are structurally related to SCR genes, and encode small-secreted proteins that are strongly implicated in playing a role in pollen-stigma interaction [48]. Interestingly, among the Brassicaceae species used in our analyses, *B. rapa*, *B. oleracea*, *B. oleracea* var. *alboglabra, R. raphanistrum,* and *R. sativus* are self-incompatible while *B. napus* and Arabidopsis are self-compatible. It is possible that similar to SCR and PCP genes, the CBSGs are differentially regulated in self-compatible and self-incompatible Brassicaceae species thereby resulting in variation in the SI response. Alternatively, mutations in S locus related genes could lead to self-compatibility [49] could have occurred during the evolution of Arabidopsis.

ALSGs and CBSGs had more genetic variation among the 20 re-sequenced Arabidopsis ecotypes than EC genes, with ALSG and EC genes having the most (0.45) and least (0.34) SNPs per 100 bp, respectively. This was inversely correlated with the evolutionary conservation of detectable homology of ALSG, CBSG, and EC sets. In addition, ALSGs and CBSGs showed substantially higher non-synonymous SNPs per 100 bp than the EC set, and as a consequence, they had higher ratios of non-synonymous to synonymous SNPs, indicating they are fast-evolving at the protein level. This is not surprising, as by definition, CBSGs and ALSGs lack detectable homology outside Brassicaceae and suggests

that these are fast-evolving genes. Collectively, this suggests ALSGs and CBSGs might be involved in biological processes that are specific to Arabidopsis and the Brassicaeae family.

Conclusions

In summary, we have identified two sets of Arabidopsis genes, CBSGs and ALSGs, which are specific to Brassicaceae family and Arabidopsis, respectively. CBSGs are especially enriched in proteins with binding function such as receptor binding that may play a role in the SI response. The exact functions of a majority of these lineage specific genes are unclear at this time. Further biological experiments would be necessary to fully understand their functions in Arabidopsis and Brassica species.

Methods

Data sources and preparation

The proteome of *Arabidopsis thaliana* was obtained from the TAIR8 release (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release). Pseudogenes and Transposable Elements (TEs) were excluded from the original gene set based on TAIR8 annotation, which resulted in 27,025 protein coding genes. Further screening against two in-house transposon databases identified an additional 163 putative TE-related genes, resulting in 26,862 Arabidopsis genes for further analysis.

The masked assembled scaffolds (v1.0) of poplar (*Populus trichocarpa*) were downloaded from DOE Joint Genome Institute (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html). The masked assembly of the grapevine (*Vitis vinifera*) genome was downloaded from Genoscope (http://www.genoscope.cns.fr/spip/Vitis-

vinifera-whole-genome.html). The release 2.0 assembly of the Medicago (*Medicago truncatula*) genome was downloaded from the Medicago Genome Sequence Consortium (http://www.medicago.org/genome/downloads.php). Release 4 pseudomolecules of rice (*Oryza sativa* ssp. japonica) were downloaded from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). The PUTs from 156 plant species (excluding Arabidopsis in this analysis) were downloaded from PlantGDB on June, 20, 2008 (http://www.plantgdb.org/download/download.php?dir=/Sequence/ESTcontig). UniProtKB (Release 14.6) was downloaded from UniProt (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/).

Genic features

The TE set was comprised of 3,900 TE genes from TAIR8 release and 163 putative TE genes identified by screening against two in-house transposon databases. For the TE set, only three sequence files were created: gene, exon, and intron as they lack CDS or protein sequences. For each of the CBSG, ALSG, and EC set, the sequences of gene, exon, CDS, intron, and protein were either downloaded directly from the TAIR8 release or extracted from the chromosome sequences according to the coordinates provided in the GFF3 file. Perl scripts were used to calculate the exon number, length of gene, CDS, exon, intron, and protein, GC content of CDS, gene, and three codon positions.

Construction of paralogous protein families

A total of 26,862 non-TE Arabidopsis proteins from TAIR8 release were used to construct paralogous protein families in the Arabidopsis proteome using a computational pipeline that utilizes Pfam [50] and novel BLASTP-based novel domains described

previously [37]. In brief, Pfam domains were identified using HMMER2 [51] with scores above the trusted cutoff value. Peptide regions that were not covered by Pfam domains were clustered based on homology (>45% identity over 75 amino acids, E-value < 1e-3) derived from an all versus all BLASTP search (WU-BLASTP 2.0MP-WashU [22-Mar-2006]) [52]. Clustered peptides were then aligned using CLUSTALW [53, 54] to develop BLASTP-based domains. Paralogous protein families were then classified based on the domain composition of each protein.

Identification of segmental duplication

A total of 26,862 non-TE Arabidopsis proteins from TAIR8 release were used to identify segmental duplication in the Arabidopsis genome using a method described previously [55]. In brief, similar protein pairs were identified by all versus all BLASTP search (WU-BLASTP 2.0MP-WashU [22-Mar-2006], parameters "V=5 B=5 E=1e-10") [52], which were then used to defined segmental using DAGChainer [56] with parameters "-s -I -D 100000".

Co-expression Analyses

The ATH1 microarray compendium of 3,037 experiments (hereafter called "supercluster") was downloaded from the NASCarray website (http://affymetrix.arabidopsis.info/narrays/help/usefulfiles.html). Only the genes having probes on the ATH1 array, 358 of the 1351 ALSGs and 328 of the 957 CBSGs, were used for further analysis. Pairwise Pearson's Correlation Coefficient was computed between all lineage-specific genes (ALSGs and CBSGs) with array data and all genes in the supercluster. The threshold value (r=0.6) was defined as the 99 percentile of all pairwise correlation

coefficients obtained during the above computation. Using this threshold, we obtained a set of co-expressed genes for each ALSG and CBSG gene tested. 291 of the 358 ALSGs and 273 of the 328 CBSGs had one or more unique gene with a significantly correlated expression profile.

To define the functional annotations of the lineage specific genes, we used the Functional Catalogue annotation scheme (v 2.1) for all the *A. thaliana* genes present on the ATH1 array. Only the top two hierarchical levels of the FunCat categories were used. For each of the 291 ALSGs and 273 CBSGs, we identified the enriched FunCat categories for the genes significantly co-expressed. This was accomplished by performing a Fisher Exact Test at a False Discovery Rate (FDR) of 5% as defined by Q-value [57]. 203 of the 291 ALSGs and 205 of the 273 CBSGs tested had at least one enriched FunCat category, out of which 192 ALSGs and 189 CBSGs had enrichment in at least one top-most level FunCat category. We performed further enrichment analysis using Fisher Exact Test to determine which of the top-most level FunCat categories were over-represented among the 192 ALSGs and 189 CBSGs, using the same Q-value threshold as above. Custom-made Python scripts were used for all of the above computations.

Determination of subcellular localization

The subcellular localization of 32,419 protein sequences from 26,862 Arabidopsis protein-coding genes was identified by TargetP program [39] using plant networks and default parameters. Subcellular localization prediction with the best (lowest) Reliability Class was used to represent the subcellular localization of the deduced protein if multiple different locations were predicted for isoforms predicted for the gene. If none of the isoforms had a

prediction of 'Chloroplast', 'Mitochondria', or 'Secreted', then the subcellular localization of the gene was assigned 'Other'. If multiple subcellular localizations with equal Reliability Class were predicted for the isoforms of a gene, then the subcellular localization of that gene was assigned 'Uncertain'.

SNP analyses

The SNP data from re-sequencing of 20 diverse Arabidopsis accessions using high-density oligonucleotide arrays [41] was downloaded from TAIR8 release (ftp://ftp.arabidopsis.org/Polymorphisms/Perlegen_Array_Resequencing_Data_2007/SNP_predictions/). The polymorphism GFF3 file that includes the mapping information of the SNP marker was also downloaded from TAIR8 release (ftp://ftp.arabidopsis.org/Polymorphisms/TAIR8_Variation_GFF/TAIR8_GFF3_polymorphisms.gff). PERL scripts were used to parse the data and calculate synonymous and non-synonymous SNPs within protein coding regions. A total of 249,344 SNPs were downloaded. Only base calls from MBML2 dataset [41] were used in our analyses. Base calls of 'N' were ignored. A total of 5,381 SNPs with more than two variations within all 20 accessions were excluded from our analyses. Representative models were used whenever alternative-splicing isoforms existed. SNPs that produce same amino acid as the reference codon (Columbia-0 ecotype) was counted as synonymous SNPs while SNPs that produce a different amino acid than the reference codon was counted as non-synonymous SNPs.

Authors' contributions

HL designed the study, conducted the majority of the computational analyses, and drafted the paper. GM and SS carried out the expression analysis. SO generated the additional data file 1. XG supervised the analysis of single nucleotide polymorphisms and the study. CRB designed the study, supervised the study, and drafted the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by funds to CRB from Michigan State University.

Reference

1. Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY: **Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota**. *J Mol Evol* 2006, **63**(1):1-11.
2. Amiri H, Davids W, Andersson SG: **Birth and death of orphan genes in Rickettsia**. *Mol Biol Evol* 2003, **20**(10):1575-1587.
3. Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM *et al*: **Mechanisms of evolution in Rickettsia conorii and R. prowazekii**. *Science* 2001, **293**(5537):2093-2098.
4. Siew N, Fischer D: **Analysis of singleton ORFans in fully sequenced microbial genomes**. *Proteins* 2003, **53**(2):241-251.
5. Domazet-Loso T, Tautz D: **An evolutionary analysis of orphan genes in Drosophila**. *Genome Res* 2003, **13**(10):2213-2219.
6. Daubin V, Lerat E, Perriere G: **The source of laterally transferred genes in bacterial genomes**. *Genome Biol* 2003, **4**(9):R57.
7. Striepen B, Pruijssers AJ, Huang J, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC: **Gene transfer in the evolution of parasite nucleotide biosynthesis**. *Proc Natl Acad Sci U S A* 2004, **101**(9):3154-3159.
8. Alba MM, Castresana J: **Inverse relationship between evolutionary rate and age of mammalian genes**. *Mol Biol Evol* 2005, **22**(3):598-606.

9. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ: **Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression**. *Proc Natl Acad Sci U S A* 2006, **103**(26):9935-9939.

10. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution**. *Genome Biol* 2006, **7**(5):R43.

11. Schmid KJ, Aquadro CF: **The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes**. *Genetics* 2001, **159**(2):589-598.

12. Rensink WA, Lee Y, Liu J, Iobst S, Ouyang S, Buell CR: **Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts**. *BMC Genomics* 2005, **6**:124.

13. Allen KD: **Assaying gene content in Arabidopsis**. *Proc Natl Acad Sci U S A* 2002, **99**(14):9568-9572.

14. Graham MA, Silverstein KA, Cannon SB, VandenBosch KA: **Computational identification and characterization of novel genes from legumes**. *Plant Physiol* 2004, **135**(3):1179-1197.

15. Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, Haas BJ, Hamilton JP, Buell CR: **Identification and characterization of lineage-specific genes within the Poaceae**. *Plant Physiol* 2007, **145**(4):1311-1322.

16. Conte MG, Gaillard S, Droc G, Perin C: **Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants**. *BMC Genomics* 2008, **9**:183.

17. Conte MG, Gaillard S, Lanau N, Rouard M, Perin C: **GreenPhylDB: a database for plant comparative genomics**. *Nucleic Acids Res* 2008, **36**(Database issue):D991-998.

18. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.

19. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007.

20. Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S: **Sequencing the genespaces of Medicago truncatula and Lotus japonicus**. *Plant Physiol* 2005, **137**(4):1174-1181.

21. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L *et al*: **The TIGR Rice Genome Annotation Resource: improvements and new features**. *Nucleic Acids Res* 2007, **35**(Database issue):D883-887.

22. Dong Q, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V: **Comparative plant genomics resources at PlantGDB**. *Plant Physiol* 2005, **139**(2):610-618.

23. Guo WJ, Li P, Ling J, Ye SP: **Significant Comparative Characteristics between Orphan and Nonorphan Genes in the Rice (Oryza sativa L.) Genome**. *Comp Funct Genomics* 2007:21676.

24. Vanoosthuyse V, Miege C, Dumas C, Cock JM: **Two large Arabidopsis thaliana gene families are homologous to the Brassica gene superfamily that encodes pollen coat proteins and the male component of the self-incompatibility response**. *Plant Mol Biol* 2001, **46**(1):17-34.

25. Franklin-Tong VE, Franklin FC: **The different mechanisms of gametophytic self-incompatibility**. *Philos Trans R Soc Lond B Biol Sci* 2003, **358**(1434):1025-1032.

26. Sage TL, Sampson FB: **Evidence for ovarian self-incompatibility as a cause of self-sterility in the relictual woody angiosperm, Pseudowintera axillaris (Winteraceae)**. *Ann Bot (Lond)* 2003, **91**(7):807-816.

27. Bittencourt NS, Jr., Gibbs PE, Semir J: **Histological study of post-pollination events in Spathodea campanulata beauv. (Bignoniaceae), a species with late-acting self-incompatibility**. *Ann Bot (Lond)* 2003, **91**(7):827-834.

28. Kemp BP, Doughty J: **Just how complex is the Brassica S-receptor complex?** *J Exp Bot* 2003, **54**(380):157-168.

29. Kochetov AV, Sarai A, Rogozin IB, Shumny VK, Kolchanov NA: **The role of alternative translation start sites in the generation of human protein diversity**. *Mol Genet Genomics* 2005, **273**(6):491-496.

30. Rogozin IB, Kochetov AV, Kondrashov FA, Koonin EV, Milanesi L: **Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon**. *Bioinformatics* 2001, **17**(10):890-900.

31. Kwon HS, Lee DK, Lee JJ, Edenberg HJ, Ahn YH, Hur MW: **Posttranscriptional regulation of human ADH5/FDH and Myf6 gene expression by upstream AUG codons**. *Arch Biochem Biophys* 2001, **386**(2):163-171.

32. Xu G, Rabadan-Diehl C, Nikodemova M, Wynn P, Spiess J, Aguilera G: **Inhibition of corticotropin releasing hormone type-1 receptor translation by an upstream AUG triplet in the 5' untranslated region**. *Mol Pharmacol* 2001, **59**(3):485-492.

33. Jin X, Turcott E, Englehardt S, Mize GJ, Morris DR: **The two upstream open reading frames of oncogene mdm2 have different translational regulatory properties**. *J Biol Chem* 2003, **278**(28):25716-25721.

34. Puyaubert J, Denis L, Alban C: **Dual targeting of Arabidopsis holocarboxylase synthetase1: a small upstream open reading frame regulates translation initiation and protein targeting**. *Plant Physiol* 2008, **146**(2):478-491.

35. Crowe ML, Wang XQ, Rothnagel JA: **Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides**. *BMC Genomics* 2006, **7**:16.

36. Blanc G, Hokamp K, Wolfe KH: **A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome**. *Genome Res* 2003, **13**(2):137-144.

37. Lin H, Ouyang S, Egan A, Nobuta K, Haas BJ, Zhu W, Gu X, Silva JC, Meyers BC, Buell CR: **Characterization of paralogous protein families in rice**. *BMC Plant Biol* 2008, **8**:18.

38. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M *et al*: **The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes**. *Nucleic Acids Res* 2004, **32**(18):5539-5545.

39. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence**. *J Mol Biol* 2000, **300**(4):1005-1016.

40. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.

41. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA *et al*: **Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana**. *Science* 2007, **317**(5836):338-342.

42. TAIR: **http://www.arabidopsis.org**.

43. Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME: **Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing A. lyrata and its autogamous relative A. thaliana**. *Plant Cell* 2001, **13**(3):627-643.

44. Bateman AJ: **Self-incompatibility systems in angiosperms: III. Cruciferae**. *Heredity* 1955, **9**:52-68.

45. Nasrallah JB: **Cell-cell signaling in the self-incompatibility response**. *Curr Opin Plant Biol* 2000, **3**(5):368-373.

46. Schopfer CR, Nasrallah ME, Nasrallah JB: **The male determinant of self-incompatibility in Brassica**. *Science* 1999, **286**(5445):1697-1700.

47. Takasaki T, Hatakeyama K, Suzuki G, Watanabe M, Isogai A, Hinata K: **The S receptor kinase determines self-incompatibility in Brassica stigma**. *Nature* 2000, **403**(6772):913-916.

48. Takayama S, Shimosato H, Shiba H, Funato M, Che FS, Watanabe M, Iwano M, Isogai A: **Direct ligand-receptor complex interaction controls Brassica self-incompatibility**. *Nature* 2001, **413**(6855):534-538.

49. Okamoto S, Odashima M, Fujimoto R, Sato Y, Kitashiba H, Nishio T: **Self-compatibility in Brassica napus is caused by independent mutations in S-locus genes**. *Plant J* 2007, **50**(3):391-400.

50. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**(Database issue):D138-141.

51. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

52. Gish W: **http://blast.wustl.edu**. In.; 1996 - 2008.

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

54. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**(13):3497-3500.

55. Lin H, Zhu W, Silva JC, Gu X, Buell CR: **Intron gain and loss in segmentally duplicated genes in rice**. *Genome Biol* 2006, **7**(5):R41.

56. Haas BJ, Delcher AL, Wortman JR, Salzberg SL: **DAGchainer: a tool for mining segmental genome duplications and synteny**. *Bioinformatics* 2004, **20**(18):3643-3646.

57. Storey JD: **A direct approach to false discovery rates**. *Journal Of The Royal Statistical Society Series B* 2002, **64**(3):479-498.

Additional data files

Additional data file 1

File format: PNG

Title: Distribution of CBSGs, ALSGs, and EC genes within the Arabidopsis genome.

Description: The five Arabidopsis chromosomes are shown with CBSGs, ALSGs, and ECs plotted in purple, red, and blue from top to bottom, respectively. Segmentally duplicated blocks are indicated in green and estimated centromeric regions are denoted by a yellow box.

Table 1. Genic features of CBSGs, ALSGs, ECs, and TE-related genes

	CBSGs		ALSGs		ECs		TE-related genes	
Feature	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median	Mean (SD)	Median
Exons/gene	2.2 (1.6)	2	1.7 (1.5)	1	5.6 (5.2)	4	1.7 (2.4)	1
Exon length	257 (246)	183	218 (231)	149	280 (352)	155	1336 (1675)	522
Intron length	205 (263)	109	227 (318)	113	163 (172)	99	160 (186)	96
Gene length	816 (667)	603	550 (673)	276	2319 (1558)	2003	2420 (1742)	2072
Protein length	150 (112)	106	97 (85)	67	431 (298)	370	na	na
Exon GC (%)	41.0 (6.0)	40.7	42.4 (6.1)	42.2	42.6 (4.6)	42.6	42.7 (5.3)	42.3
Intron GC (%)	31.4 (7.4)	31.3	35.1 (7.5)	34.4	32.4 (4.4)	32.7	32.8 (7.9)	31.9
Gene GC (%)	37.8 (5.0)	37.8	41.0 (5.1)	40.9	39.6 (3.3)	39.4	41.5 (4.6)	41.4
CDS/ORF GC(%)	42.2 (4.3)	42	42.8 (4.8)	42.7	44.6 (3.2)	44.2	na	na
1st position GC (%)	45.6 (6.7)	45.6	45.7 (7.6)	45.7	50.1 (4.6)	50.2	na	na
2nd position GC (%)	40.4 (6.5)	40	40.0 (7.7)	40	40.5 (5.4)	40.1	na	na
3rd position GC (%)	40.7 (8.3)	40.9	42.8 (8.1)	42.9	42.9 (6.3)	42.1	na	na

Table 2. Functional annotation of CBSGs, ALSGs, and ECs

	CBSGs		ALSGs		ECs	
	No. of genes	Percentage (%)	No. of genes	Percentage (%)	No. of genes	Percentage (%)
with no known function	724	75.7	1,269	93.9	5,043	20.5
transcript support	572	59.8	658	48.7	4,864	19.8
no transcript support	152	15.9	611	45.2	179	0.7
with a known function	233	24.3	82	6.1	19,511	79.5
transcript support	159	16.6	66	4.9	18,685	76.1
no transcript support	74	7.7	16	1.2	826	3.4
putative PCP or SCR[a]	73	7.6	4	0.3	39	0.2
uORF	10	1.0	20	1.5	36	0.1
beta-galactosidase	0	0.0	13	1.0	34	0.1
Other	151	15.7	45	3.3	19,402	79.0
Total	957	100.0	1,351	100.0	24,554	100.0

[a]PCP (pollen coat protein) gene family or SCR (S locus cysteine-rich protein)

Table 3. Subcellular localization of CBSGs, ALSGs, ECs, and TAIR8 non-TE protein-coding genes

	CBSGs				ALSGs			
	No. of genes	Percentage (%)	No. of known genes	No. of expressed genes	No. of genes	Percentage (%)	No. of known genes	No. of expressed genes
Chloroplast	38	4.0	7	31	64	4.7	5	48
Mitochondria	66	6.9	9	58	234	17.3	12	114
Secreted	441	46.1	119	308	275	20.4	13	113
Other	412	43.1	98	334	778	57.6	52	429
Uncertain	0	0.0	0	0	0	0.0	0	0
Total	957	100	233	731	1,351	100	82	704

Table 3. Subcellular localization of CBSGs, ALSGs, ECs, and TAIR8 non-TE protein-coding genes (Continued)

	ECs				TAIR8 non-TE Protein-coding Genes			
	No. of genes	Percentage (%)	No. of known genes	No. of expressed genes	No. of genes	Percentage (%)	No. of known genes	No. of expressed genes
Chloroplast	3,905	15.9	3,022	3,833	4,007	14.9	3,034	3,912
Mitochondria	2,833	11.5	2,149	2,772	3,133	11.7	2,170	2,944
Secreted	4,718	19.2	3,824	4,472	5,434	20.2	3,956	4,913
Other	13,085	53.3	10,507	12,459	14,275	53.1	10,657	13,222
Uncertain	13	0.1	9	13	13	0	9	13
Total	24,554	100	19,511	23,549	26,862	100	19,826	25,004

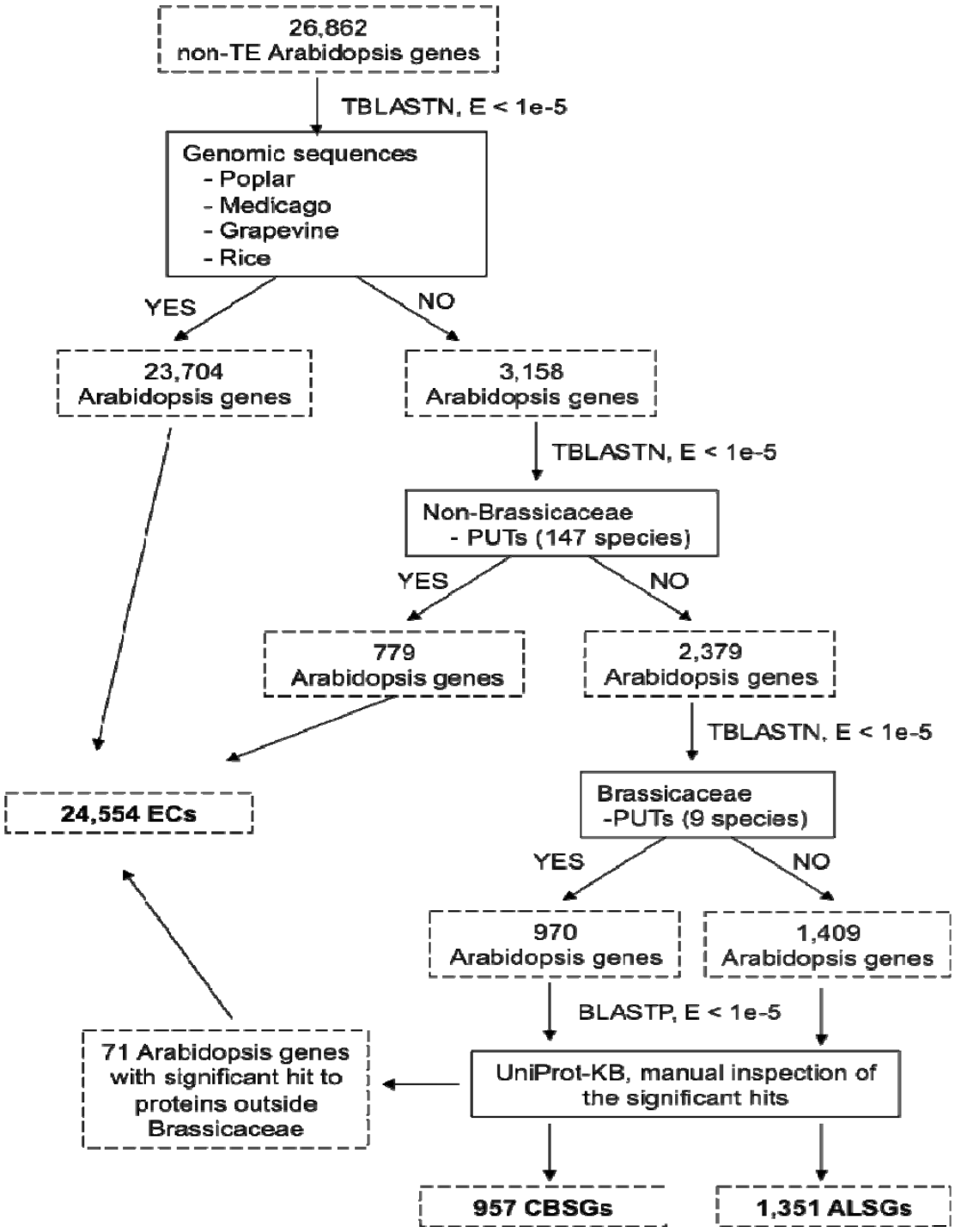


Figure 1. Flowchart of identification of CBSGs and ALSGs. The solid boxes reflect non-Arabidopsis sequences used in the searches while the hashed boxes show the Arabidopsis genes.

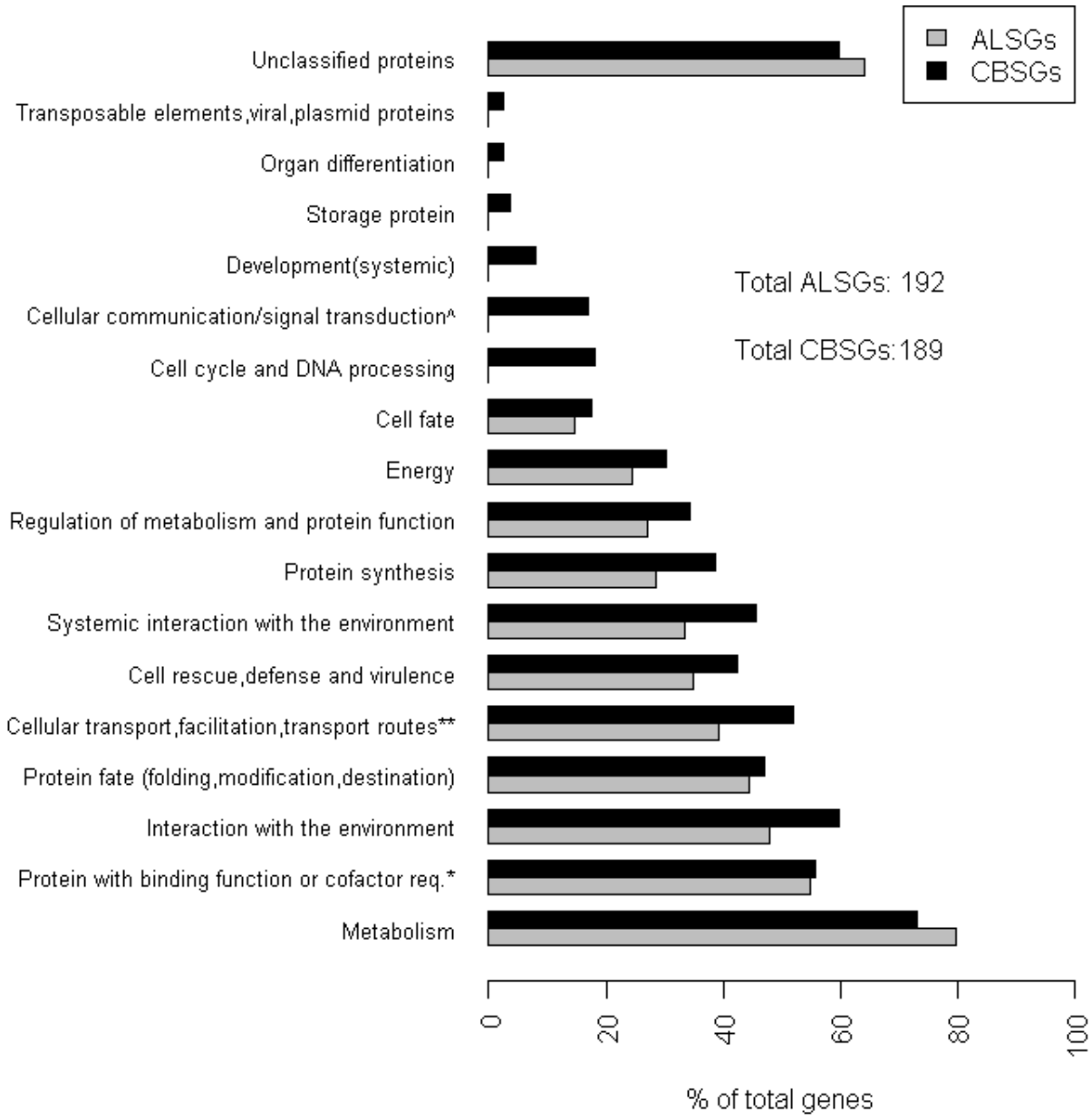


Figure 2. Enriched FunCat categories in ALSG and CBSG gene sets.

Note: The category "Subcellular localization" is not shown here

*: Protein with binding function or cofactor requirement (structural or catalytic)

**: Cellular transport, transport facilitation and transport routes

^: Cellular communication/Signal transduction mechanism

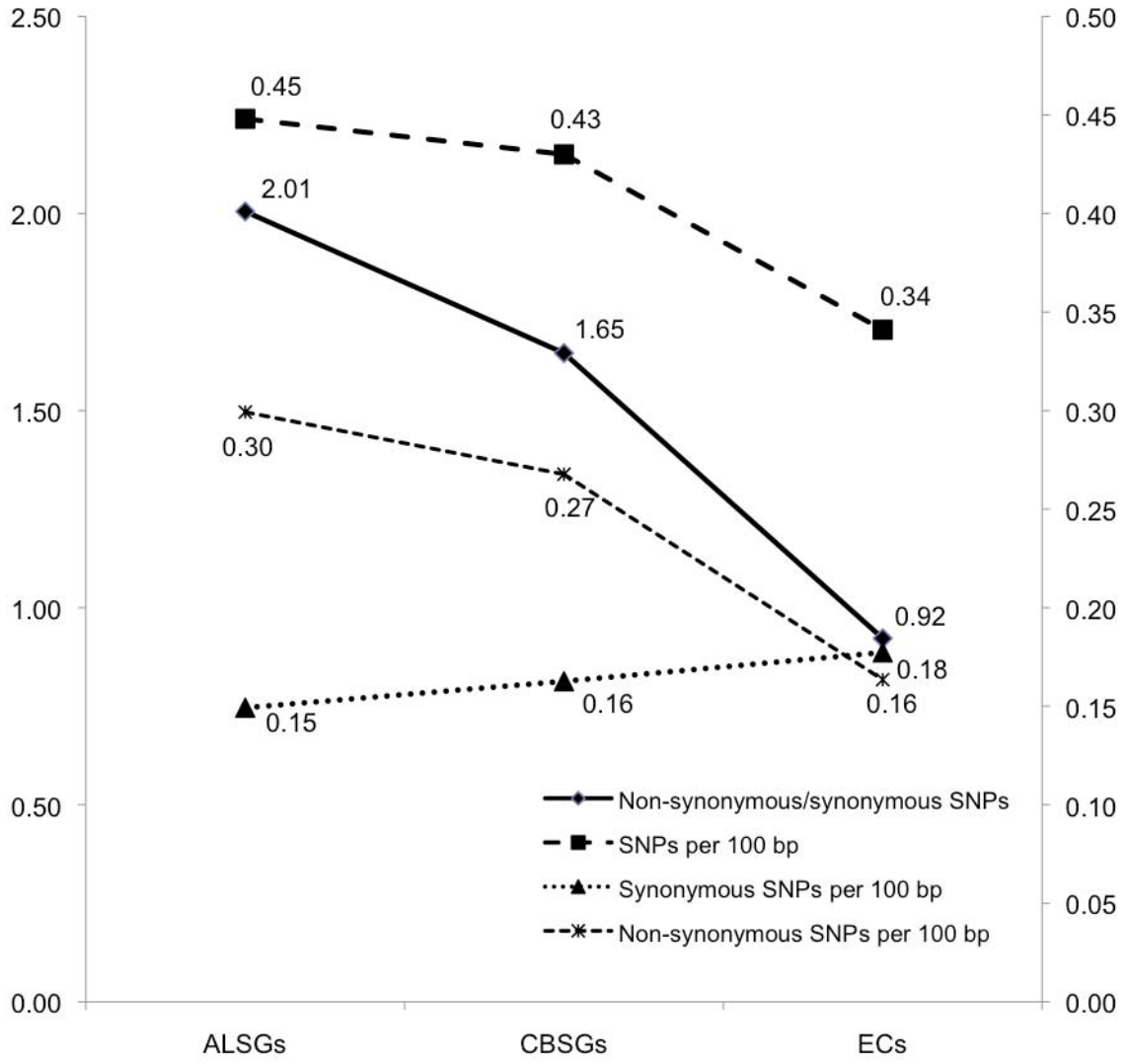


Figure 3A. Ratio of non-synonymous to synonymous SNPs substitutions and number of SNPs, non-synonymous SNPs, and synonymous SNPs per 100 bp within coding region of ALSGs, CBSGs, and ECs. Ratio of non-synonymous to synonymous substitutions SNPs is plotted in solid line. Number of SNPs, non-synonymous SNPs, and synonymous SNPs per 100 bp within coding regions is plotted in dotted line.

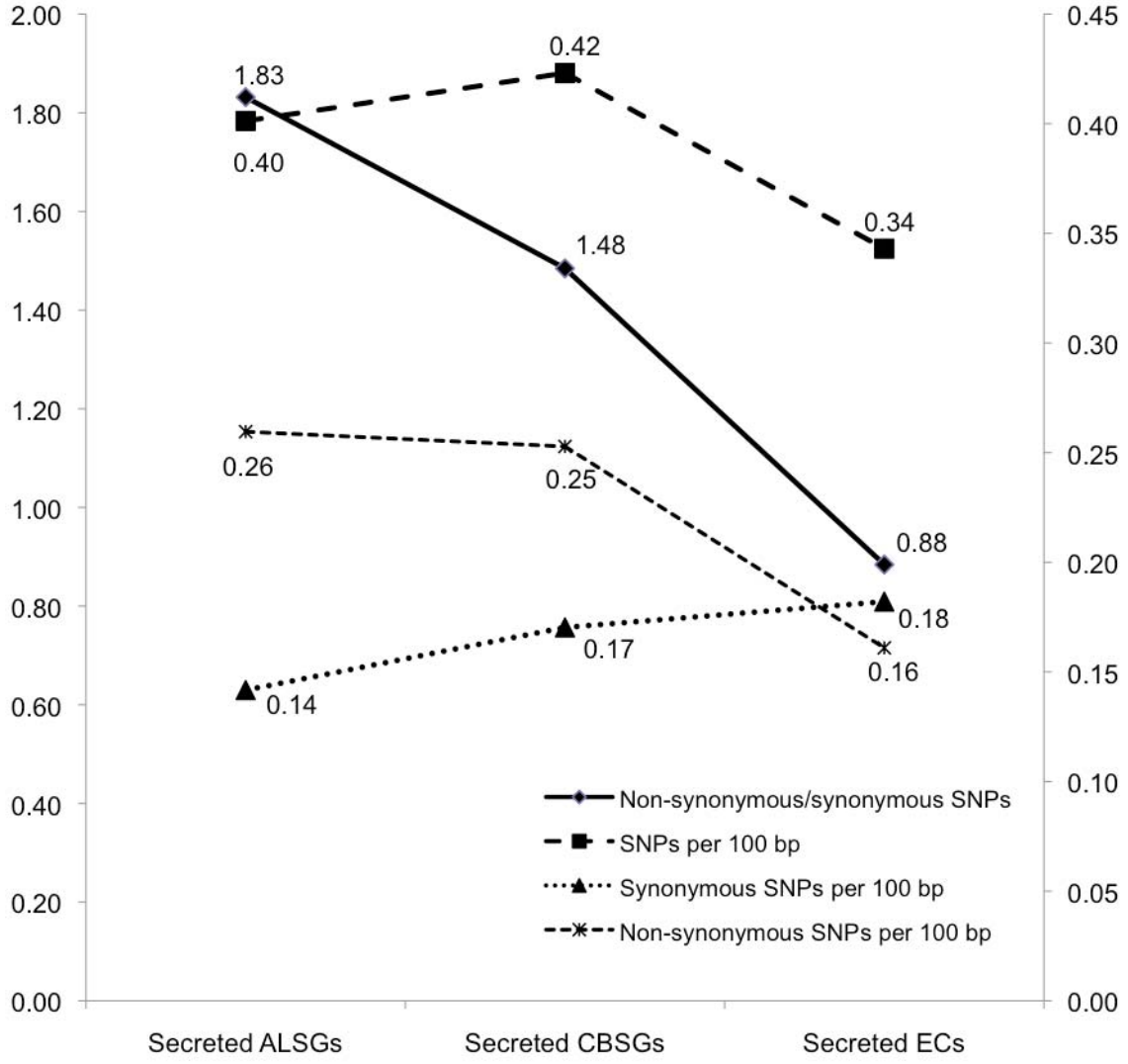
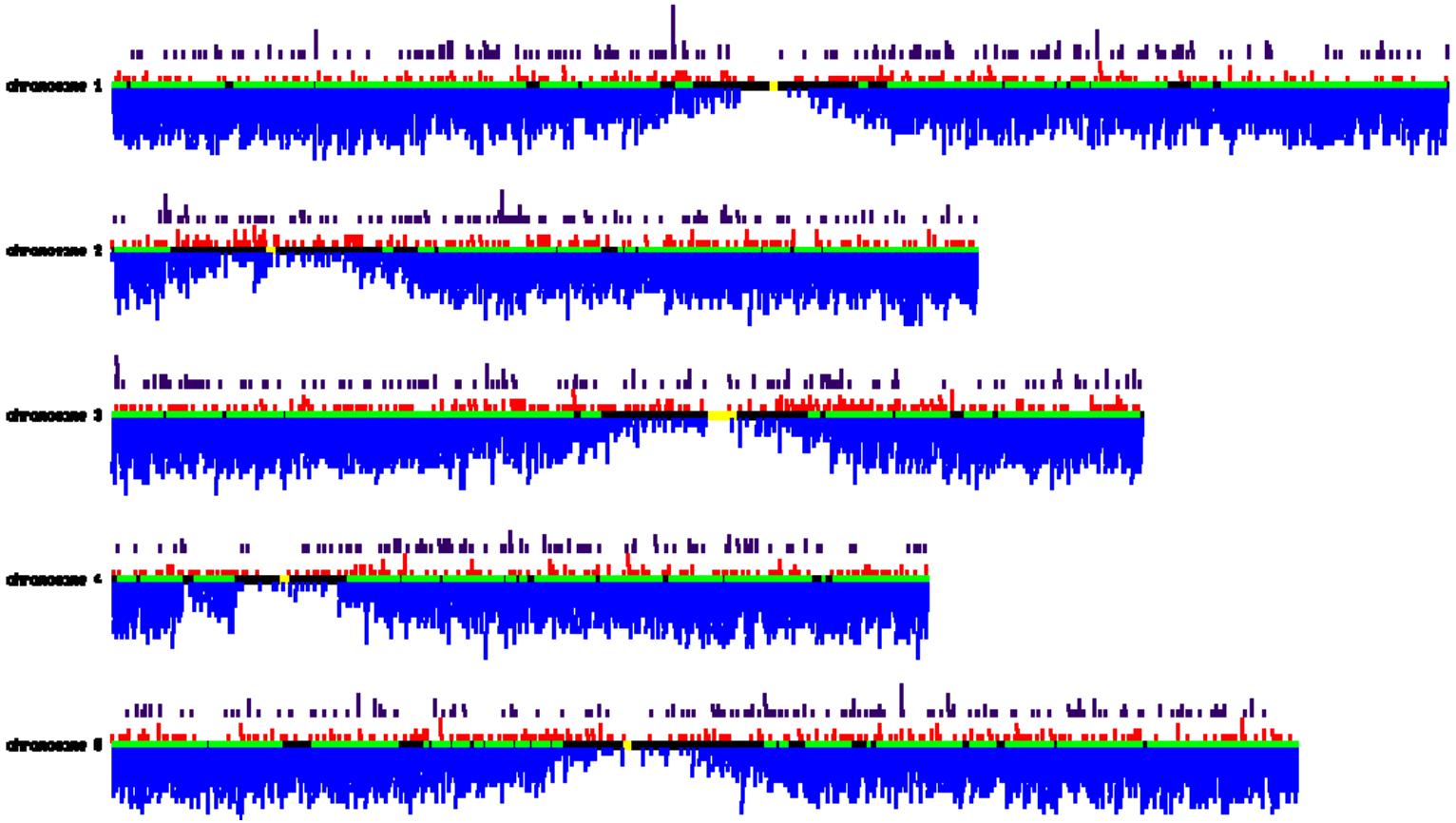


Figure 3B. Ratio of non-synonymous to synonymous SNPs substitutions and number of SNPs, non-synonymous SNPs, and synonymous SNPs per 100 bp within coding region of secreted ALSGs, CBSGs, and ECs. Ratio of non-synonymous to synonymous substitutions SNPs is plotted in solid line. Number of SNPs, non-synonymous SNPs, and synonymous SNPs per 100 bp within coding regions is plotted in dotted line.



Additional data file 1

CHAPTER 5. GENERAL CONCLUSIONS

In this dissertation, we examined segmental duplication, intron turnover, and paralogous protein family composition in the rice genome. In addition, with the availability of more complete or near-complete plant genome and transcriptome sequence datasets across a wide range of species, we identified and characterized conserved Brassicaceae-specific genes and Arabidopsis lineage-specific genes.

Summary of findings

In Chapter 2, we characterized intron gain and loss events in the rice genome with respect to segmental duplication events. We observed that intron evolution within the rice genome following segmental duplication is dominated by intron loss rather than intron gain. We did not observe preferential intron loss at the 3'end of genes as previously reported in mammalian genomes, nor did we observe any statistically significant difference in intron loss rate at different phases. However, we did observe that the four nucleotides of exons that flank the donor splice site of lost introns had less frequently used 4-mers. We found that two of the five gained introns were similar to transposable elements, suggesting transposable elements may play a role in intron evolution.

In Chapter 3, we characterized paralogous protein families in rice in a comparative context with Arabidopsis using a computational pipeline that utilized both Pfam and novel BLASTP-based domains. We found that both rice (53%) and Arabidopsis (68%) proteomes have a substantial fraction of gene families. Paralogous family genes (rice: 73%, Arabidopsis: 96%) are more likely to be genes with a known or putative function, compared

to singleton genes (rice: 26%, Arabidopsis: 66%). Using Plant GOSlim annotation, we found that 17 out of the 26 Gene Ontology categories analyzed were statistically different in their distribution between paralogous family and singleton genes. In contrast to mammalian organisms, we found that paralogous family genes in rice and Arabidopsis tend to have more alternative splice forms, suggesting that plants may employ multiple mechanisms for proteomic complexity. Using data from Massively Parallel Signature Sequencing (MPSS), we show that a significant portion of the duplicated genes in rice show divergent expression and tissue specific expression with correlation between expression and sequence divergence observed only in very young genes.

In chapter 4, we identified 957 (3.6%) Arabidopsis genes that lack sequence similarity with other sequences within the Plant Kingdom except sequences within the Brassicaceae family (Conserved Brassicaceae-Specific Genes, CBSGs) and 1,351 (5.0%) Arabidopsis genes unique to *A. thaliana* (Arabidopsis Lineage-Specific Genes, ALSGs). We compared the CBSGs and ALSGs to the 24,554 (91.4%) Arabidopsis genes (termed EC for Evolutionarily Conserved) that are evolutionarily conserved within the Plant Kingdom as defined by sequence similarity with at least one of 147 plant species. A large majority of the 957 CBSGs (75.7%) and 1,351 ALSGs (93.9%) had no known function. However, based on EST, cDNA, and microarray data, 76.4% of the CBSGs and 53.6% of the ALSGs are transcribed, suggesting that they are likely to be bona fide genes. Co-expression analysis of CBSGs and ALSGs indicated enrichment in 19 and 13 of 27 FunCat categories, respectively, suggesting a wide range of biological functions for these two sets of genes. Subcellular localization prediction revealed that CBSGs were significantly enriched in secretory proteins

(441, 46.1%) compared to the EC set (4,718, 19.2%). Among the 119 putatively secreted CBSGs with a known function, 72 encode putative pollen coat proteins or S locus cysteine-rich proteins which are involved in the self-incompatibility response in Brassicaceae species. Single Nucleotide Polymorphism analysis showed an elevated ratio of non-synonymous to synonymous SNPs within the ALSGs (2.01) and the CBSGs (1.65) relative to the EC genes (0.92), mainly caused by an elevated number of non-synonymous SNPs, suggesting that they are fast-evolving at protein sequence level.

Ongoing and future work

With the promising results we obtained for rice and Arabidopsis using comparative approaches presented here, we are interested in expanding the depth and breadth of our studies by:

- Classification and investigation of orthologous groups among four sequenced flowering plants (rice, Arabidopsis, poplar, grapevine), and consequently improvement of the functional annotation of the rice genome (results not shown in this dissertation).
- Exploration of alternative splicing patterns in orthologous genes between rice and Arabidopsis.
- Validation of ALSGs and CBSGs with no known function, and further exploration of their putative functions.