COMPARATIVE GENOMICS OF GENE EXPRESSION IN EUKARYOTES

by

YUPENG WANG

(Under the Direction of Romdhane Rekaya and Andrew Paterson)

ABSTRACT

Advances made in genome sequencing in the past decade have produced a massive amount of genetic information. There is a need for developing quantitative methods aimed at exploiting this information, with the ultimate objective being to attain a better understanding of the biological processes taking place. Microarrays, which can characterize the transcriptional profiles of tens of thousands of genes simultaneously, have been widely used in comparative genomic studies. However, the analysis of microarray data is still challenging. To date, assessing cross-species conservation of gene expression using microarray data has been mainly based on comparison of expression patterns across corresponding tissues, or comparison of coexpression of a gene with a reference set of genes. We compared one corresponding tissue-based method and three coexpression-based methods for assessing conservation of gene expression, in terms of their pair-wise agreements, using a frequently used human-mouse tissue expression dataset. Gene expression patterns were then compared between human and mouse genomes using both corresponding tissue-based and coexpression-based methods. To detect and analyze synteny and collinearity, we have developed the *MCScanX* toolkit, which implements an adjusted MCScan algorithm and incorporates 14 utility programs. In *Arabidopsis thaliana* and *Oryza sativa* (rice), species that deeply sample botanical diversity and for which expression data are available from a

wide range of tissues and physiological conditions, we have compared expression divergence between genes duplicated by six different mechanisms (whole-genome, tandem, proximal, DNA-based transposed, retrotransposed and dispersed duplications), and between positional orthologs. The findings imply that gene duplication modes differ in contribution to genetic novelty and redundancy, but show some parallels in taxa separated by hundreds of millions of years of evolution.

INDEX WORDS: Microarray, gene expression, conservation, divergence, coexpression, synteny, collinearity, algorithm, software, gene family, gene duplication, genetic novelty, genetic redundancy, ortholog, phylogenetic analysis

COMPARATIVE GENOMICS OF GENE EXPRESSION IN EUKARYOTES

by

YUPENG WANG

BS, Xiamen University, China, 2004

MS, Xiamen University, China, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHYLOSOPHY

ATHENS, GEORGIA

2011

COMPARATIVE GENOMICS OF GENE EXPRESSION IN EUKARYOTES

by

YUPENG WANG

Major Professors:	Romdhane Rekaya
	Andrew Paterson
Committee:	Paul Schliekelman
	Jessica Kissinger

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

DEDICATION

I would like to dedicate this work to my loving wife, who always encouraged all of my pursuits.

ACKNOWLEDGEMENTS

I would like to thank Dr. Romdhane Rekaya and Dr. Andrew Paterson, my major professors, for guidance and assistance in completion of my doctoral dissertation. Also, I would like to thank the rest of my committee members, Dr. Paul Schliekelman and Dr. Jessica Kissinger, for valuable time and guidance.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Changes in gene expression directed by transcriptional regulation often give rise to new phenotypes in eukaryotes. Advances in microarray technology have made the systematic study of gene expression evolution possible. However, microarray data contain a high level of noise, rendering the data analysis at a genomic scale very challenging. Gene duplication is a key biological mechanism for providing gene expression diversification both within species and across species. Gene duplication is a complicated biological process, which may occur by different modes including whole-genome, tandem, proximal, transposed and dispersed duplications. All these factors highlight the importance of developing advanced computational techniques and mining new biological significance in the field of comparative gene expression in eukaryotes. The aims of this dissertation include

1) Evaluation of existing computational techniques for assessing gene expression conservation across species.

2) Unraveling gene expression evolution and its functional significance between humans and rice.

3) Development of publicly available software for synteny and collinearity detection.

4) Investigation of heterogeneous patterns of expression divergence between duplicate genes in *Arabidopsis* and rice.

CHAPTER 2

LITERATURE REVIEW

Advances made in genome sequencing in the past decade have produced a massive amount of genetic information. There is a need for developing quantitative methods aimed at exploiting this information, with the ultimate objective being to attain a better understanding of the biological processes taking place. Sequence comparison is the most popular tool for comparative genomics. However, sequence similarity is not necessarily proportional to functional similarity (Gerlt and Babbitt 2000). The biological functions of a gene not only rely on its molecular functions but also its spatiotemporal expression pattern. Changes in gene expression often trigger changes in gene networks, which may further give rise to new phenotypes. Thus, it is of great importance to study both gene expression and sequence evolution to fully understand gene evolution.

Microarrays, which can characterize the transcriptional profiles of tens of thousands of genes simultaneously, have been widely used in comparative genomic studies. Studies of gene expression levels in different species often rely on cross-species hybridization (Fortna et al. 2004; Khaitovich et al. 2004; Nuzhdin et al. 2004; Khaitovich et al. 2005). This method is limited to closely related species as it is based on the hybridization of target RNA and gene probes designed for other species (Oshlack et al. 2007), and when the probe and target RNA sequences are inconsistent to some extent, this method fails. Even in related species, several studies found that this approach may be problematic (Gilad et al. 2005; Bar-Or et al. 2006).

In animals including humans, mice, *Drosophila, Caenorhabditis elegans* and *Xenopus*, the conservation/divergence of gene expression across species has been extensively and systematically assessed. However, results of such studies are often conflicting. Yanai et al. (2004) concluded that no expression conservation exists in human and mouse orthologous gene pairs because the evolution in the expression profiles of orthologous gene pairs was shown to be comparable to that of randomly paired genes. In contrast, Liao and Zhang (2006a) found that the expression profile divergence for the majority of orthologous genes between humans and mice is significantly lower than expected under neutrality. Khaitovich et al. (2004) suggested that the majority of expression divergences between species are selectively neutral and are non-functional adaptations, while Jordan et al. (2005) suggested that gene expression divergence among mammalian species is subject to the effects of purifying selection and could also be substantially influenced by positive Darwinian selection. Yang et al. (2005) found that broadly expressed genes have lower rates of gene expression profile evolution than narrowly expressed genes in mammals, while Liao and Zhang (2006b) demonstrated the opposite. Furthermore, several studies found a strong correlation between gene expression divergence and coding sequence divergence in humans, mice, *Drosophila* and *Xenopus* (Nuzhdin et al. 2004; Khaitovich et al. 2005; Lemos et al. 2005; Liao and Zhang 2006a; Sartor et al. 2006), while other studies suggested little correlation between them in humans, mice, *Drosophila* and *C. elegans* (Jordan et al. 2004; Yanai et al. 2004; Jordan et al. 2005; Dutilh et al. 2006; Tirosh and Barkai 2007; Tirosh and Barkai 2008).

Some of these conflicting conclusions on gene expression evolution may be due, in part, to improper comparisons of gene expression across genomes, such as direct comparisons of expression levels across probes or platforms and/or cross-species microarrays hybridization. To

overcome these limitations, indirect comparisons of gene expression across species have become a popular method for assessing conservation of gene expression. Liao and Zhang (2006a) introduced the method of using relative mRNA abundance over 26 common tissues between humans and mice to make cross-species expression comparisons possible. However, their method can be only implemented in closely related species, as it requires that the two microarray experiments sample orthologous tissues and use the same experimental procedures. Based on the conceptual framework of comparing co-expression patterns across species proposed by Ihmels et al. (2005), Dutilh et al. (2006), Tirosh and Barkai (2007), and Essien et al. (2008) used either all or part of the 1-1 orthologs (i.e. in both species, there is only one corresponding ortholog) as a reference set between species and computed the correlations of a gene's expression profile with those of the reference set to facilitate the assessment of the degree of gene expression conversation across genomes. Theoretically, this framework can be applied to any species and any microarray data types. However, the use of the whole 1-1 ortholog set, as references in the study by Dutilh et al. (2006), may be problematic because the subset of 1-1 orthologs with fast expression evolution may distort the true relationship of query genes. Tirosh and Barkai (2007) identified this limitation and tried to minimize the influence of 1-1 orthologs with fast expression evolution by giving larger weights to orthologous pairs with conserved expression. Essien et al. (2008) used the 1-1 orthologs in conserved co-expression networks, instead of whole ortholog set, as a reference set between *Plasmodium* species.

The aforementioned methods represent two computational models for assessing conservation/divergence of gene expression across species: 1) comparison of gene expression patterns across corresponding tissues, and 2) comparison of co-expression of a gene with a reference set of genes. Although the separate application of either model has yielded significant

biological insights (Dutilh et al. 2006; Liao and Zhang 2006a; Liao and Zhang 2006b; Tirosh and Barkai 2007; Essien et al. 2008; Liao and Zhang 2008; Liao et al. 2010), a systematic assessment of these models, especially their agreement with each other has yet to be reported.

In plant genomes, the study of gene expression evolution can be confounded by diverse gene duplication modes. Whole-genome duplications (WGDs) have frequently occurred in the lineages of plants (Paterson et al. 2010), with possible consequences including evolution of novel or modified gene functions (Ohno 1970; Lynch and Conery 2000; Zhang and Cohn 2008; Kassahn et al. 2009), provision of "buffer capacity" (Chapman et al. 2006; VanderSluis et al. 2010) or genetic redundancy that increases genetic robustness (Gu et al. 2003; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; DeLuna et al. 2010). Genome duplication may also increase opportunities for nonreciprocal recombination that may result in gene conversion or crossing-over (Wang et al. 2007; Wang et al. 2009a; Wang et al. 2011). Rapid DNA loss and restructuring of low-copy DNA (Song et al. 1995; Ozkan et al. 2001; Shaked et al. 2001; Kashkush et al. 2002), retrotransposon activation (O'Neill et al. 1998; Kashkush et al. 2003; Paterson et al. 2009) and epigenetic changes (Chen and Pikaard 1997; Comai et al. 2000; Lee and Chen 2001; Rodin and Riggs 2003; Adams and Wendel 2005; Rapp and Wendel 2005) following WGD may further provide materials for evolutionary change.

However, computational identification of WGD events as well as the duplicate genes that were created and retained from WGDs has been a challenging task (Van de Peer 2004; Paterson et al. 2010). In general, this task is often solved through analyzing synteny (i.e. genes remaining on corresponding chromosomal regions) and collinearity (i.e. genes remaining in corresponding orders along the chromosomes) among several related species (Tang et al. 2008a; Tang et al. 2008b). One classical theme for synteny detection is to use all versus all BLASTP searches as

inputs, and model the matches in a homology matrix for synteny detection through clustering neighboring matches inside the matrix. This approach was implemented in ADHoRe (Vandepoele et al. 2002a), DiagHunter (Cannon et al. 2003) and other derived algorithms (Calabrese et al. 2003). Another classical theme for synteny detection is to use dynamic programming to detect synteny and statistical strategies to evaluate synteny, e.g. DAGchainer (Haas et al. 2004) and ColinearScan (Wang et al. 2006). However, the aforementioned tools detect only pairwise collinear segments. Thus, they are insufficient for distinguishing the different WGD events that a genome has experienced.

Early approaches for computational detection of paleopolyploidy were "bottom-up", starting with the most recent duplication event, and then resolving more ancient ones sequentially through recursively merging duplicated segments to generate hypothetical intermediate chromosomal segments. Alternatively, top-down algorithms can be used instead (Tang et al. 2008a; Tang et al. 2008b). Pair-wise collinear segments are picked from whole-genome BLASTP results to produce multi-alignments of collinear segments against reference chromosomes, revealing cryptic synteny based on transitive homology, which has been referred to as ghost duplications (Vandepoele et al. 2002a; Vandepoele et al. 2002b; Vandepoele et al. 2003). Based on this idea, a tool named MCScan was developed (Tang et al. 2008b). MCScan was first used to analyze the duplication relationships among *A. thaliana*, *Populus trichocarpa* and *Carica papaya*, using *Vitis vinifera* as the reference genome (Tang et al. 2008b). A shared ancient hexaploidy (γ) event was revealed among these taxa (Figure 2.1). The implementation of MCScan also revealed that proportions of genes created and retained from WGDs or segmental duplications fluctuated among taxa. For example, 54% of *Arabidopsis* genes and 80% of *Populus* genes were created by WGDs or segmental duplications, versus only 11% of *Carica* genes and

18% of *Vitis* genes. Segmental duplication can be regarded as a type of small-scale duplications, but is often difficult to distinguish from WGD.

Genes may be duplicated by several mechanisms other than WGDs. These are referred to as small-scale duplications (Maere et al. 2005) or single-gene duplications (Cusack and Wolfe 2007; Freeling 2009). Tandem duplicates are adjacent to one another in the genome and presumed to arise through unequal crossing over (Freeling 2009), while proximal duplicates, which are near to one another but separated by a few genes, are inferred to occur by localized transposon activities (Zhao et al. 1998). Dispersed duplicates are neither adjacent to each other in the genome nor within homeologous chromosome segments (Ganko et al. 2007). Single-gene transposition may explain the widespread existence of dispersed duplicates within and among genomes (Freeling 2009). It has been suggested that single-gene transposition duplication (referred to as transposed duplication) may occur by DNA-based or RNA-based mechanisms (Cusack and Wolfe 2007). DNA transposons such as packmules (rice) (Jiang et al. 2004), helitrons (maize) (Brunner et al. 2005), and CACTA elements (sorghum) (Paterson et al. 2009) may relocate duplicated genes to new chromosomal positions. RNA-based transposed duplication, often referred to as retrotransposition, typically creates a single-exon retrocopy from a multi-exon parental gene, by reverse transcription of a spliced messenger RNA. It is presumed that the retrocopy duplicates only the transcribed sequence of the parental gene, detached from the parental promoter. The new retrogene is often deposited in a novel chromosomal environment with a different set of gene neighbors and is likely to survive as a functional gene only if a new promoter is acquired (Brosius 1991; Kaessmann et al. 2009).

Population genetic theory suggests that a likely consequence of gene duplication is reversion to single copy (singleton), unless at least one gene copy evolves new functions (Ohno

1970). Recently, the subfunctionalization model, which proposes that duplicated gene copies might both be retained if they partition the functions of the ancestral gene between them, has attracted researchers' attention (Force et al. 1999; Lynch and Conery 2000). Some studies have also shown evidence to support the value of genetic redundancy (Hughes and Hughes 1993; Hughes 1994; Gu et al. 2003; Chapman et al. 2006; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; DeLuna et al. 2010).

The angiosperms (flowering plants) are an outstanding model in which to elucidate the consequences of gene duplication. All angiosperms are now thought to be paleopolyploids (Bowers et al. 2003), many of which underwent multiple WGDs (Paterson et al. 2004; Tang et al. 2008a). It has been shown that *Arabidopsis* experienced two 'recent' WGDs, i.e. since its divergence from other members of the Brassicales clade (α and β), and a more ancient triplication (γ) shared with most if not all eudicots (Bowers et al. 2003; Tang et al. 2008a; Tang et al. 2008b). Likewise, rice appears to have experienced at least two WGDs, one shared with most if not all cereals (ρ), and another more ancient event (σ) (Tang et al. 2010). Proven WGD in angiosperms events are shown in Figure 2.1. Single-gene duplications in angiosperms are also widespread (Freeling et al. 2008; Freeling 2009; Woodhouse et al. 2010).

One avenue for systematic investigation of functional divergence between duplicate genes is comparison of their spatiotemporal profiles of gene expression, comparing degrees of divergence with proxies of duplication age such as synonymous (Ks) substitution rates between duplicate genes. In *Arabidopsis*, the rate of protein sequence evolution is asymmetric in >20% of duplicate pairs and functional diversification of the surviving duplicate genes has been proposed to be a major feature of the long-term evolution of polyploids (Blanc and Wolfe 2004). *Arabidopsis* genes created by large-scale duplication events are more evolutionarily conserved in

gene expression than those created by small-scale duplication or those that do not lie in duplicate segments, and the time since duplication plays important roles in the functional divergence of genes (Casneuf et al. 2006). Further, there may be also a strong positive correlation between expression divergence and non-synonymous (Ka) in *Arabidopsis*, and that the different modes (segmental, tandem and dispersed) of duplication may affect patterns of expression divergence (Ganko et al. 2007). In addition, compared with singletons, *Arabidopsis* duplicate genes increase expression diversity in closely related species and allopolyploids (Ha et al. 2009). In rice, expression correlation is significantly higher for gene pairs from WGDs or tandem duplications than those from dispersed duplications, and expression divergence is closely related to divergence time (Li et al. 2009).

Though many studies have investigated the functional divergence and retention of duplicate genes, conclusions are often contradictory, e.g. gene retention has been attributed to both neofunctionalization (Zhang and Cohn 2008; Kassahn et al. 2009) and genetic redundancy (Gu et al. 2003; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; DeLuna et al. 2010), and expression divergence between duplicate genes has been suggested to be both time dependent (Casneuf et al. 2006; Li et al. 2009) and selection dependent (Ganko et al. 2007). The fates of duplicate genes may be influenced by different modes of gene duplication, which have been suggested to retain genes in a biased manner (Freeling 2009). With the increase of available expression and annotation data, and improved ability to discern various mechanisms of gene duplication, there is merit in re-examining some existing hypotheses on gene duplication as well as exploring some new hypothesis.

Along with the development of high-throughput sequencing technologies, RNA-seq, the first sequencing-based method that allows the entire transcriptome to be surveyed in a very high-

throughput and quantitative manner, has been used to facilitate the study of comparative gene expression (McManus et al. 2010; Brawand et al. 2011). RNA-Seq offers both single-base resolution for annotation and 'digital' gene expression levels at the genome scale, often at a much lower cost than either tiling arrays or large-scale Sanger EST sequencing (Wang et al. 2009b). RNA-seq faces several bioinformatics challenges such as the development of efficient methods to store, retrieve and process large amounts of data, reduce errors in image analysis and base-calling and remove low-quality reads (Wang et al. 2009b). However, as the cost of sequencing continues to fall and software tools for analyzing RNA-seq data are increasingly developed, RNA-Seq is expected to replace microarrays for many applications that should involve comparative genomics of gene expression.

References

Adams KL, Wendel JF. 2005. Novel patterns of gene expression in polyploid plants. *Trends Genet* **21**(10): 539-543.

Bar-Or C, Bar-Eyal M, Gal TZ, Kapulnik Y, Czosnek H, Koltai H. 2006. Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. *BMC Genomics* **7**: 110.

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

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930): 433-438.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M *et al.* 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**(7369): 343-348.

Brosius J. 1991. Retroposons - Seeds of Evolution. *Science* **251**(4995): 753-753.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**(2): 343-360.

Calabrese PP, Chakravarty S, Vision TJ. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19 Suppl 1**: i74-80.

Cannon SB, Kozik A, Chan B, Michelmore R, Young ND. 2003. DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol* **4**(10): R68.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol* **7**(2): R13.

Chapman BA, Bowers JE, Feltus FA, Paterson AH. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc Natl Acad Sci U S A* **103**(8): 2730-2735.

Chen ZJ, Pikaard CS. 1997. Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proc Natl Acad Sci U S A* **94**(7): 3442-3447.

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B. 2000. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell* **12**(9): 1551-1568.

Cusack BP, Wolfe KH. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* **24**(3): 679-686.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet* **4**(7): e1000113.

DeLuna A, Springer M, Kirschner MW, Kishony R. 2010. Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS Biol* **8**(3): e1000347.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. *Nat Genet* **40**(5): 676-681.

Dutilh BE, Huynen MA, Snel B. 2006. A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics* **7**: 10.

Essien K, Hannenhalli S, Stoeckert CJ, Jr. 2008. Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to *Plasmodium* phenotypic diversity. *PLoS One* **3**(9): e3122.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-1545.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T *et al.* 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* **2**(7): E207.

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

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order *Brassicales*. *Genome Res* **18**(12): 1924-1937.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol* **24**(10): 2298-2309.

Gerlt JA, Babbitt PC. 2000. Can sequence determine function? *Genome Biol* **1**(5): REVIEWS0005.

Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res* **15**(5): 674-680.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**(6918): 63-66.

Ha M, Kim ED, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A* **106**(7): 2295-2300.

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

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *P Roy Soc Lond B Bio* **256**(1346): 119-124.

Hughes MK, Hughes AL. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus-laevis*. *Mol Biol Evol* **10**(6): 1360-1369.

Ihmels J, Bergmann S, Berman J, Barkai N. 2005. Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet* **1**(3): e39.

Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**(7008): 569-573.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* **345**(1): 119-126.

Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* **21**(11): 2058-2070.

Kaessmann H, Vinckenbosch N, Long MY. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**(1): 19-31.

Kafri R, Dahan O, Levy J, Pilpel Y. 2008. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A* **105**(4): 1243-1248.

Kashkush K, Feldman M, Levy AA. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**(4): 1651-1659.

Kashkush K, Feldman M, Levy AA. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* **33**(1): 102-106.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res* **19**(8): 1404-1418.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**(5742): 1850-1854.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol* **2**(5): E132.

Lee HS, Chen ZJ. 2001. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc Natl Acad Sci U S A* **98**(12): 6753-6758.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* **22**(5): 1345-1354.

Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J. 2009. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics* **10 Suppl 6**: S8.

Liao BY, Weng MP, Zhang JZ. 2010. Contrasting genetic paths to morphological and physiological evolution. *Proc Natl Acad Sci U S A* **107**(16): 7353-7358.

Liao BY, Zhang JZ. 2006a. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**(3): 530-540.

Liao BY, Zhang JZ. 2006b. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol* **23**(6): 1119-1128.

Liao BY, Zhang JZ. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**(19): 6987-6992.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494): 1151-1155.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102**(15): 5454-5459.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res* **20**(6): 816-825.

Musso G, Costanzo M, Huangfu MQ, Smith AM, Paw J, Luis BJS, Boone C, Giaever G, Nislow C, Emili A *et al.* 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res* **18**(7): 1092-1099.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol* **21**(7): 1308-1317.

O'Neill RJ, O'Neill MJ, Graves JA. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**(6680): 68-72.

Ohno S. 1970. *Evolution by gene duplication*. Springer Verlag, New York.

Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2007. Using DNA microarrays to study gene expression in closely related species. *Bioinformatics* **23**(10): 1235-1242.

Ozkan H, Levy AA, Feldman M. 2001. Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* **13**(8): 1735-1747.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al.* 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**(7229): 551-556.

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

Paterson AH, Freeling M, Tang H, Wang X. 2010. Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* **61**: 349-372.

Rapp RA, Wendel JF. 2005. Epigenetics and plant evolution. *New Phytologist* **168**(1): 81-91.

Rodin SN, Riggs AD. 2003. Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol* **56**(6): 718-729.

Sartor MA, Zorn AM, Schwanekamp JA, Halbleib D, Karyala S, Howell ML, Dean GE, Medvedovic M, Tomlinson CR. 2006. A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in *Xenopus*. *Nucleic Acids Res* **34**(1): 185-200.

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. 2001. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* **13**(8): 1749-1759.

Song K, Lu P, Tang K, Osborn TC. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proc Natl Acad Sci U S A* **92**(17): 7719-7723.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008a. Synteny and collinearity in plant genomes. *Science* **320**(5875): 486-488.

Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci U S A* **107**(1): 472-477.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**(12): 1944-1954.

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* **8**(4): R50.

Tirosh I, Barkai N. 2008. Evolution of gene sequence and gene expression are not correlated in yeast. *Trends Genet* **24**(3): 109-113.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* **5**(10): 752-763.

Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y. 2002a. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res* **12**(11): 1792-1801.

Vandepoele K, Simillion C, Van de Peer Y. 2002b. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet* **18**(12): 606-608.

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

VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol* **6**: 429.

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. 2006. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics* **7**: 447.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177**(3): 1753-1763.

Wang X, Tang H, Bowers JE, Paterson AH. 2009a. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res* **19**(6): 1026-1032.

Wang X, Tang H, Paterson AH. 2011. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell* **23**(1):27-37.

Wang Z, Gerstein M, Snyder M. 2009b. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**(1): 57-63.

Woodhouse MR, Pedersen B, Freeling M. 2010. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet* **6**(5): e1000949.

Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* **8**(1): 15-24.

Yang J, Su AI, Li WH. 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Mol Biol Evol* **22**(10): 2113-2118.

Zhang G, Cohn MJ. 2008. Genome duplication and the origin of the vertebrate skeleton. *Curr Opin Genet Dev* **18**(4): 387-393.

Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**(5): 479-492.

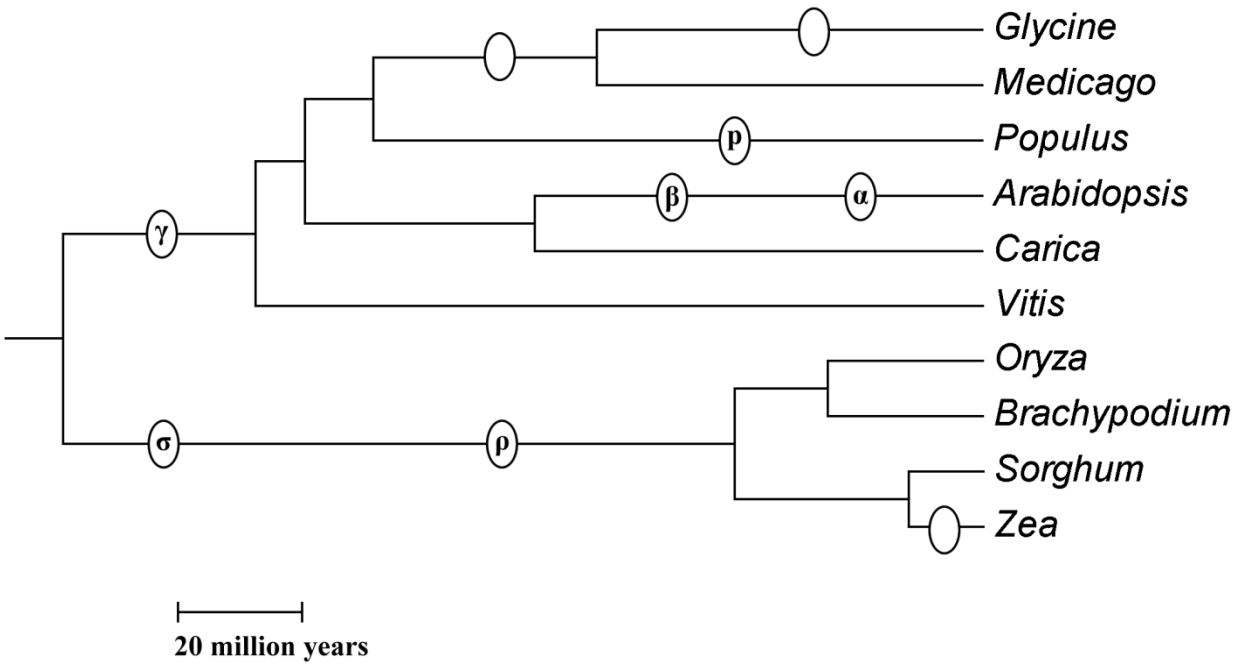

Figure 2.1. Proven WGD events in angiosperms.

CHAPTER 3

COMPARISON OF COMPUTATIONAL MODELS FOR ASSESSING CONSERVATION OF GENE EXPRESSION ACROSS SPECIES[1]

[1]Yupeng Wang, Kelly R Robbins and Romdhane Rekaya. 2010. *PLoS One*. 5(10): e13239.

Reprinted here with permission from the publisher.

Abstract

Assessing conservation/divergence of gene expression across species is important for the understanding of gene regulation evolution. Although advances in microarray technology have provided massive high-dimensional gene expression data, the analysis of such data is still challenging. To date, assessing cross-species conservation of gene expression using microarray data has been mainly based on comparison of expression patterns across corresponding tissues, or comparison of coexpression of a gene with a reference set of genes. Because direct and reliable high-throughput experimental data on conservation of gene expression are often unavailable, the assessment of these two computational models is very challenging and has not been reported yet. In this study, we compared one corresponding tissue-based method and three coexpression-based methods for assessing conservation of gene expression, in terms of their pair-wise agreements, using a frequently used human-mouse tissue expression dataset. We find that 1) the coexpression-based methods are only moderately correlated with the corresponding tissue-based methods, 2) the reliability of coexpression-based methods is affected by the size of the reference ortholog set, and 3) the corresponding tissue-based methods may lose some information for assessing conservation of gene expression. We suggest that the use of either of these two computational models to study the evolution of a gene's expression may be subject to great uncertainty, and the investigation of changes in both gene expression patterns over corresponding tissues and coexpression of the gene with other genes is necessary.

Introduction

The biological functions of a gene, not only rely on its molecular composition and structure, but also on its spatiotemporal expression pattern. For example, duplicate genes, which are usually associated with highly consistent coding sequences but diverse biological functions,

have only a weak correlation between rates of sequence and expression divergences (Wagner 2000). Thus, it is of great importance to study both gene expression and sequence information to fully understand gene evolution.

Thanks to advances in microarray technology, the conservation/divergence of gene expression across species has been extensively and systematically assessed. However, results of such studies are often conflicting. Yanai et al. (2004) concluded that no expression conservation exists in human and mouse orthologous gene pairs because the evolution in the expression profiles of orthologous gene pairs was shown to be comparable to that of randomly paired genes. In contrast, Liao and Zhang (2006a) found that the expression profile divergence for the majority of orthologous genes between humans and mice is significantly lower than expected under neutrality. Khaitovich et al. (2004) suggested that the majority of expression divergences between species are selectively neutral and are non-functional adaptations, while Jordan et al. (2005) suggested that gene expression divergence among mammalian species is subject to the effects of purifying selection and could also be substantially influenced by positive Darwinian selection. Yang et al. (2005) found that broadly expressed genes have lower rates of gene expression profile evolution than narrowly expressed genes, while Liao and Zhang (2006b) proved the opposite. Furthermore, several studies found a strong correlation between gene expression divergence and coding sequence divergence (Nuzhdin et al. 2004; Lemos et al. 2005; Paabo et al. 2005; Liao and Zhang 2006a; Tomlinson et al. 2006), while other studies (Yanai et al. 2004; Jordan et al. 2005; Dutilh et al. 2006; Tirosh and Barkai 2007; Tirosh and Barkai 2008) suggested little correlation between them.

Some of these conflicting conclusions on gene expression evolution may be due, in part, to improper comparisons of gene expression across genomes, such as direct comparisons of

expression levels across probes or platforms, as suggested by Liao and Zhang (2006a). Furthermore, cross-species microarrays hybridization may be problematic even when applied to closely related species (Gilad et al. 2005; Bar-Or et al. 2006). To overcome these limitations, indirect comparisons of gene expression across species have become a popular method for assessing conservation of gene expression. Liao and Zhang introduced the method of using relative mRNA abundance over 26 common tissues between humans and mice to make cross-species expression comparisons possible (Liao and Zhang 2006a). However, their method can be only implemented in closely related species, as it requires that the two microarray experiments sample orthologous tissues and use the same experimental procedures. Based on the conceptual framework of comparing coexpression patterns across species proposed by Ihmels et al. (2005), Dutilh et al. (2006), Tirosh and Barkai (2007), and Essien et al. (2008) used either all or part of the 1-1 orthologs as a reference set between species and computed the correlations of a gene's expression profile with those of the reference set for facilitating the study of assessing the degree of gene expression conversation across genomes. Theoretically, this framework can be applied to any species and any microarray data types. However, the use of the whole 1-1 ortholog set (WOS), as references in the study by Dutilh et al. (2006), may be problematic because the subset of 1-1 orthologs with fast expression evolution may distort the true relationship of query genes. Tirosh and Barkai (2007) identified this limitation and tried to minimize the influence of 1-1 orthologs with fast expression evolution by giving larger weights to orthologous pairs with conserved expression. Essien et al. (2008) used the 1-1 orthologs in conserved coexpression networks (CCNs), instead of WOS, as a reference set between species.

The aforementioned methods represent two computational models for assessing conservation/divergence of gene expression across species: 1) comparison of gene expression

patterns across corresponding tissues, and 2) comparison of coexpression of a gene with a reference set of genes. Although the separate application of either model has yielded significant biological insights (Liao and Zhang 2006a; Liao and Zhang 2006b; Tirosh and Barkai 2007; Essien et al. 2008; Liao and Zhang 2008; Tirosh and Barkai 2008; Liao et al. 2010), a systematic assessment of these models, especially their agreement with each other has yet to be reported. Until most recently, our group (Wang and Rekaya 2009) implemented both of these models to assess gene expression evolution between humans and mice. Surprisingly, we found little overlap between the conserved Gene Ontology (GO) terms detected by the two models. This observation has raised our concern about the usefulness and accuracy of the biological conclusions obtained using indirect comparison methods.

In this study, we assessed one corresponding tissue-based method: Liao and Zhang's method (Liao and Zhang 2006a) and three coexpression-based methods: Dutilh et al.'s method (Dutilh et al. 2006), Tirosh and Barkai's method (Tirosh and Barkai 2007) and Essien et al.'s method (Essien et al. 2008), in terms of their pair-wise agreements. The comparisons were conducted using the human-mouse tissue gene expression data from Su et al. (2004), one of the most frequently used dataset for the study of gene expression evolution.

Methods

Microarray data and annotations

A public human and mouse expression dataset was downloaded from GNF SymAtlas V1.2.4. at http:// symatlas.gnf.org/SymAtlas/ (GEO accession number: GSE1133) (Su et al. 2004). The dataset consisted of 79 human and 61 mouse tissues using specially designed Affymetrix microarray chips (human: HG-U133A&GNF1H; mouse: GNF1M). The gene expression levels were obtained using MAS 5.0 algorithms (Hubbell et al. 2002). To minimize

the random effects of low expression values on estimating correlations (Pereira et al. 2009), probes with an expression level < 200 were removed from analyses. The annotation files for GNF1H and GNF1M were downloaded from GNF SymAtlas along with the data files. The annotation file for HG-U133A was downloaded from the Affymetrix website (http://www.affymetrix.com). To assign the Ensembl ID for each gene, the annotation files (humans: uniprot_sprot_human.dat; mice: uniprot_sprot_rodents.dat) were downloaded from the Uniprot FTP site at ftp://us.expasy.org/databases/uniprot/current_release/knowledgebase/taxonomic_divisions. The orthologous gene pairs between humans and mice were downloaded from the Ensembl FTP site (ftp://ftp.ensembl.org). Only 1-1 orthologs were considered in this study. The number of available 1-1 orthologous gene pairs was 7182, out of which 3142 had multiple probe sets. For a gene with multiple probe sets, the selection of a probe set that best represents the gene's expression profile according to a general rule has not been resolved yet (Elbez et al. 2006). Thus, in this study and in order to remove a potential additional source of variation in the data, the 1-1 orthologs with multiple probe sets were removed from analyses. The final number of human and mouse 1-1 orthologous gene pairs used for this study was 4040. These 4040 human-mouse1-1 orthologs constituted the WOS.

Liao and Zhang's method for assessing conservation of gene expression between humans and mice

The expression data of 26 common tissues from two species were extracted and normalized by their relative abundance (RA) values calculated as:

$$RA_H(i,j) = S_H(i,j) / \sum_{j=1}^{n} S_H(i,j)$$

$$RA_M(i,j) = S_M(i,j) / \sum_{j=1}^{n} S_M(i,j)$$

where n is the number of common tissues, H represents humans, M represents mice, and $S_H(i,j)$ and $S_M(i,j)$ are the expression levels of gene i in human and mouse tissue j, respectively. The expression conservation (EC) for human-mouse orthologous pair i is calculated as:

$$EC(i) = \frac{\sum_{j=1}^{n}[RA_H(i,j)RA_M(i,j)] - \frac{\sum_{j=1}^{n} RA_H(i,j)\sum_{j=1}^{n} RA_M(i,j)}{n}}{\sqrt{(\sum_{j=1}^{n}[RA_H(i,j)]^2 - \frac{[\sum_{j=1}^{n} RA_H(i,j)]^2}{n})(\sum_{j=1}^{n}[RA_M(i,j)]^2 - \frac{[\sum_{j=1}^{n} RA_M(i,j)]^2}{n})}}.$$

Its corresponding expression divergence measured by Euclidian distance is computed as:

$$d(i) = \sqrt{\sum_{j=1}^{n}(RA_H(i,j) - RA_M(i,j))^2}$$

Existing coexpression-based methods for assessing conservation of gene expression

Expression datasets with different dimensions under different conditions between any two species, A and B, can be compared. The expression matrices, **A** and **B**, in species A and B respectively, are restricted to genes for which 1-1 orthology relationships have been identified and ordered accordingly (i.e., equivalent rows of the two matrices correspond to the expression profiles of a pair of orthologs):

$$\mathbf{A} = [\mathbf{a}_i]_{i=1,\ldots,k}$$
$$\mathbf{B} = [\mathbf{b}_i]_{i=1,\ldots,k}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the vectors of expression profiles for any pair i of 1-1 orthologs for species A and B, respectively, and k is the number of 1-1 orthologous gene pairs.

A and **B** are then converted into two pair-wise correlation matrices (PCMs), $\mathbf{R}^A$ and $\mathbf{R}^B$, by computing the Pearson's correlation coefficient (denoted by PCC or r) between the expression profiles of each pair of genes over all conditions in each species separately:

$$\mathbf{R}^A = [PCC(\mathbf{a}_i, \mathbf{a}_g)]_{i=1,\ldots,k;g=1,\ldots,k}$$
$$\mathbf{R}^B = [PCC(\mathbf{b}_i, \mathbf{b}_g)]_{i=1,\ldots,k;g=1,\ldots,k}$$

$\mathbf{R}^A$ and $\mathbf{R}^B$, contain all the correlations between genes that have 1-1 orthology relationships. As they have the same dimension *k*, any row $R^A_{i,g}, 1 \le g \le k$ from $\mathbf{R}^A$ and any row $R^B_{j,g}, 1 \le g \le k$ from $\mathbf{R}^B$ can be correlated.

Dutilh et al. (2006) defined the expression conservation (EC) for an orthologous gene pair *i* as:

$$EC(i) = PCC(R^A_{i,g}, R^B_{i,g}), 1 \le g \le k$$

Tirosh and Barkai (2007) suggested that a difference between $R^A_{i,g}$ and $R^B_{i,g}$ does not necessarily correspond to a difference in expression patterns of $\mathbf{a}_i$ and $\mathbf{b}_i$, and thus when calculating the similarity between $\mathbf{a}_i$ and $\mathbf{b}_i$,, larger weight should be given to orthologous pairs whose expression has been conserved. To that aim, they developed the Iterative Comparison of Coexpression (ICC) algorithm. The ICC algorithm extends the above described procedure by iteratively refining the ECs using a weighted correlation, where the weight for each gene is given by the EC of that gene from the previous iteration:

$$EC_l(i) = PCCw(R^A_{i,g'}, R^B_{i,g'})$$

where

$$PCCw(X,Y) = \frac{\sum w_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum w_i (X_i - \bar{X})^2 \sum w_i (Y_i - \bar{Y})^2}}$$
$$w_i = EC_{l-1}(i)$$
$$g' = \{m \in g \mid EC_{l-1}(m) > 0\}$$

This iterative process is repeated until convergence:

$$\sum_{i \in g} [EC_l(i) - EC_{l-1}(i)]^2 < 0.1$$

Essien et al. (2008) computed the inter-species correlation, another expression of EC, in a similar way to how Dutilh et al. (2006) computed the EC, except the reference ortholog set consisted of only the nodes in conserved coexpression networks (CCNs) between species. Thus, the EC by Essien et al.'s method can be computed as:

$$\mathbf{R}^A' = [PCC(\mathbf{a}_i, \mathbf{a}_{g'})]_{i=1,\ldots,k;\, g' \in CCN}$$
$$\mathbf{R}^B' = [PCC(\mathbf{b}_i, \mathbf{b}_{g'})]_{i=1,\ldots,k;\, g' \in CCN}$$

$$EC(i) = PCC(R^A_{i,g'}', R^B_{i,g'}'),\ g' \in CCN$$

For coexpression-based methods, the Euclidian distance between orthologs of gene *i* is computed as:

$$d(i) = \sqrt{\sum_{g \in \text{Reference ortholog set}} (R^A_{i,g} - R^B_{i,g})^2}$$

Identification of reference ortholog set required for application of Essien et al.'s method

To apply Essien et al.'s method, the nodes of CCNs between humans and mice should be identified first. In this study, the identification of the nodes in CCNs, was performed via determination of conserved pair-wise coexpression between species, i.e. the expression profiles of a pair of genes are significantly correlated in both species. Intra-species background distributions of correlations were first constructed based on 20,000 random gene pairs. All two gene combinations were assessed for potential conserved coexpression. Gene pairs whose

expression profiles were significantly correlated (r greater than a certain quantile x of the background correlation distribution, in both humans and mice) were selected as nodes of CCNs. Because the correlation cutoff value may affect the number of CNN nodes and in order to fully assess Essien et al.'s method, we varied the correlation coefficient threshold. Out of 4040 pairs of human-mouse 1-1 orthologs, 3390, 2424 and 1246 pairs were found as nodes of CCNs when the correlation threshold was set to 0.95, 0.975 and 0.99 quantile of the background distribution, respectively.

Results

Because prior knowledge on the expression conservation for human-mouse orthologs is limited (expression conservation may not be associated with sequence conservation (Jordan et al. 2004; Jordan et al. 2005; Dutilh et al. 2006; Tirosh and Barkai 2007; Tirosh and Barkai 2008), it is difficult to establish a benchmark for accurately evaluating the computational methods used for assessing expression conservation in terms of sensitivity and specificity. Given this difficulty and the purpose of this study, to examine whether different computational methods would generate consistent results on expression conservation, the performances of Liao and Zhang's method, Dutilh et al.'s method, ICC and Essien et al.'s method were evaluated based on their pair-wise agreements.

Plots of the distributions of ECs for all human-mouse orthologous gene pairs and 4040 human-mouse random gene pairs, generated by different methods can be found in Figure 3.1. The means and standard deviations of these distributions are shown in Table 3.1. Generally, the comparisons of EC distributions between human-mouse orthologs and random gene pairs by different methods all prove the theory of non-random expression conservation of orthologs. This confirms that all the methods examined in this study are able to detect expression conservation.

Note that there may be two steps in obtaining results of expression conservation of orthologs bioinformatically: the identification of orthologs and the measurement of expression conservation between orthologs. Liao and Zhang's method addresses issues related to the second step, while coexpression-based methods can be applied to both steps. To demonstrate the usefulness of coexpression-based methods in the first step, we re-generated the above results by disturbing the orthology relationships in the reference ortholog set (via permuting the order of columns of $\mathbf{R}^B$). In this case, non-random expression conservation of orthologs is not observed (negative data are not shown), confirming that the 1-1 orthologs are a good reference gene set for coexpression-based methods.

Evaluation of the agreement between corresponding tissue-based methods and coexpression-based methods

Using Liao and Zhang's method as a reference, the three coexpression-based methods generated variable EC distributions (Figure 3.1). For human-mouse random gene pairs, Essien et al.'s method at *x* (see the Methods section)=0.975 generated an EC distribution that best approximated the EC distribution by Liao and Zhang's method; For the human-mouse orthologous gene pairs, when x=0.975, Essien et al.'s method resulted in an EC distribution with a similar mean and a smaller standard deviation by comparison with Liao and Zhang's method. Within relation to Liao and Zhang's method, when x=0.95 and x=0.99, Essien et al.'s method tended to underestimate and overestimate the ECs respectively; Dutilh et al.'s method tended to underestimate the ECs and ICC tended to overestimate the ECs, though ICC had a comparable standard deviation to that obtained by Liao and Zhang's method. Additionally, the ECs of all human-mouse orthologous gene pairs generated by different coexpression-based methods were correlated with those by Liao and Zhang's method. The correlation values are shown in Table

3.2. These results suggest that the coexpression-based methods are only moderately correlated with the corresponding tissue-based methods, and although Essien et al.'s method appears to best agree with Liao and Zhang's method, its performance is affected by the size of the reference ortholog set (i.e., number of the nodes in CCNs). Note that although coexpression-based methods may generated different EC distributions, the ECs of human-mouse 1-1 orthologs computed by different coexpression-based methods are highly correlated ($0.962 \le r \le 0.997$).

The reliability of coexpression-based methods for assessing cross-species conservation of gene expression may be greatly affected by the inclusion of fast evolving genes as references, as suggested by Tirosh and Barkai (2007). As such, a potential underlying problem with ICC is that, because $EC_0(i)$ may be incorrectly computed using equal weights for all orthologous pairs which consist of both conserved and fast evolving genes (in expression), the weights given to the subsequent iterations may also be incorrect. Thus, an alternative approach to minimize the effects of fast evolving genes may rely on using a refined reference set which excludes fast evolving genes, such as Essien et al.'s method. The orthologs that are involved in CCNs have been shown to be more conserved in gene expression between species (Semon and Duret 2006), which should be a better reference set for cross-species comparison of gene expression than WOS. Although it is reasonable to let the reference ortholog set consist of nodes in CCNs, the size of the reference set should be chosen appropriately because large reduction of dimensions may cause the correlation values to be unreliable while a too large size makes the performance of Essien et al.'s method approach that of Dutilh et al.'s method. Based on the analysis in this study, we would suggest that the size of the reference ortholog set range from $0.5|WOS|$ to 0.7 $|WOS|$.

Problems in Liao and Zhang's method

Liao and Zhang's method was based on a subset of the microarray data, represented by the expression profiles over 26 human-mouse common tissues. However, the original human and mouse expression data cover 79 human tissues and 61 mouse tissues respectively. The potential problems for Liao and Zhang's method include 1) the similarity of gene expression profiles over only 26 common tissues may not reflect the expression conservation over all available tissues, and 2) common tissues are not the same tissues, i.e. tissues evolve between humans and mice.

Because there are no means of applying Liao and Zhang's method to the whole human and mouse tissue data, to quantify the effects of using the microarray data over only common tissues, we adopted an indirect approach: comparing coexpression-based methods using the whole microarray data with the expression data over only common tissues (the same data used by Liao and Zhang's method), with the hypothesis that if the results on expression conservation do not differ significantly between the two types of expression data, the use of the expression data over common tissues should not be a factor affecting the assessment of expression conservation, which should be also true to Liao and Zhang's method. However, we found that the properties of EC distributions generated by coexpression-based methods differ greatly between these two types of expression data (Table 3.3), and that the ECs of all human-mouse orthologous gene pairs inferred based on the whole microarray data and the expression data over 26 common tissues are only moderately correlated ($0.60 \le r \le 0.69$), suggesting that the reduction from the whole microarray data to the expression data over 26 common tissues results in loss of information for assessing conservation of gene expression.

Discussion

By applying coexpression-based methods to the expression data of 26 common tissues between humans and mice, i.e. the same data used by Liao and Zhang's method, a maximum agreement between corresponding tissue-based methods and coexpression-based methods can be estimated. Using this dataset, the ECs of all human-mouse 1-1 orthologs generated by different coexpression-based methods were correlated with those generated by Liao and Zhang's method. Though these correlations were increased from (0.48-0.50) to (0.69-0.74), a maximum correlation of 0.74 is still far from a high agreement (say, $r > 0.9$), suggesting that even if the same data are used, corresponding tissue-based methods and coexpression-based methods may still give different estimations of ECs.

In addition to expression conservation, expression divergence between species is also a measure for studying evolution of gene expression. Some studies used 1-EC as a measure of expression divergence (Liao and Zhang 2008; Liao et al. 2010), and in this case the agreement between the assessed computational methods should be the same as the above analysis. Some studies used the Euclidean distance of expression profiles as a measure of expression divergence (Jordan et al. 2005; Kim et al. 2006; Yanai et al. 2006; Urrutia et al. 2008). We further reproduced the results by using Euclidean distances instead of ECs. However, negative correlations ($-0.29 \le r \le -0.24$) were observed between the Euclidean distances of human-mouse 1-1 orthologs computed by Liao and Zhang's method and those by coexpression-based methods. This contradiction is not surprising as some previous studies have showed that Pearson's correlations and Euclidean distances may be completely uncorrelated (Jordan et al. 2005; Liao and Zhang 2006a; Pereira et al. 2009). To assess expression conservation, we would suggest the use of correlations instead of Euclidean distance because 1) they show agreements

between different computational models; 2) unlike Euclidian distance, the scale of correlation ([-1, 1]) is not affected by different degrees of freedom. In addition to the potential contradiction between them, correlation and Euclidian distance have other limitations. They both measure the global similarity/divergence between gene expression profiles over multiple conditions/tissues, which may leave condition-specific / tissue-specific changes of gene expression undetected. However, some of these undetected changes may be caused by striking genetic evolution. Some studies (Gu and Su 2007; Singh and Hannenhalli 2010) have suggested that condition-specific / tissue-specific changes of gene expression should be also surveyed for fully understanding the mechanisms of gene regulation evolution.

In this study, we compared two popular computational models for assessing conservation of gene expression. The corresponding tissue-based methods are only moderately correlated with coexpression-based methods. All the assessed methods have limitations and thus, the use of a combination of Liao and Zhang's method and Essien et al.'s method (Essien et al.'s method appears better than Dutilh et al.'s method and ICC) is recommended. However, the two assessed computational models, which mainly capture the information on the global changes in gene expression patterns over orthologous tissues and in gene coexpression networks, reveal only part of the whole picture of gene expression evolution. Additionally, besides expression abundance as an indicator of gene expression behavior, expression breadth and specificity are also worth investigating (Yang et al. 2005; Liao and Zhang 2006b; Park and Choi 2010). Development of computational methods that properly model the divergence of expression breadth or specificity across species may be an important part of comprehensively assessing conservation of gene expression.

References

Bar-Or C, Bar-Eyal M, Gal TZ, Kapulnik Y, Czosnek H, Koltai H. 2006. Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. *BMC Genomics* **7**: 110.

Dutilh BE, Huynen MA, Snel B. 2006. A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics* **7**: 10.

Elbez Y, Farkash-Amar S, Simon I. 2006. An analysis of intra array repeats: the good, the bad and the non informative. *BMC Genomics* **7**: 136.

Essien K, Hannenhalli S, Stoeckert CJ, Jr. 2008. Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to *Plasmodium* phenotypic diversity. *PLoS One* **3**(9): e3122.

Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res* **15**(5): 674-680.

Gu X, Su ZX. 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc Natl Acad Sci U S A* **104**(8): 2779-2784.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18**(12): 1585-1592.

Ihmels J, Bergmann S, Berman J, Barkai N. 2005. Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet* **1**(3): e39.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* **345**(1): 119-126.

Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* **21**(11): 2058-2070.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol* **2**(5): E132.

Kim RS, Ji H, Wong WH. 2006. An improved distance measure between the expression profiles linking co-expression and co-regulation in mouse. *BMC Bioinformatics* **7**: 44.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* **22**(5): 1345-1354.

Liao BY, Weng MP, Zhang J. 2010. Contrasting genetic paths to morphological and physiological evolution. *Proc Natl Acad Sci U S A* **107**(16): 7353-7358.

Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**(19): 6987-6992.

Liao BY, Zhang JZ. 2006a. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**(3): 530-540.

Liao BY, Zhang JZ. 2006b. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol* **23**(6): 1119-1128.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol* **21**(7): 1308-1317.

Paabo S, Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**(5742): 1850-1854.

Park SG, Choi SS. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol* **10**: 241.

Pereira V, Waxman D, Eyre-Walker A. 2009. A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics* **183**(4): 1597-1600.

Semon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol* **23**(9): 1715-1723.

Singh LN, Hannenhalli S. 2010. Correlated changes between regulatory cis elements and condition-specific expression in paralogous gene families. *Nucleic Acids Res* **38**(3): 738-749.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al.* 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**(16): 6062-6067.

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* **8**(4): R50.

Tirosh I, Barkai N. 2008. Evolution of gene sequence and gene expression are not correlated in yeast. *Trends Genet* **24**(3): 109-113.

Tomlinson CR, Sartor MA, Zorn AM, Schwanekamp JA, Halbleib D, Karyala S, Howell ML, Dean GE, Medvedovic M. 2006. A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association

between mRNA sequence divergence and differential gene expression in *Xenopus*. *Nucleic Acids Res* **34**(1): 185-200.

Urrutia AO, Ocana LB, Hurst LD. 2008. Do Alu repeats drive the evolution of the primate transcriptome? *Genome Biol* **9**(2): R25.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc Natl Acad Sci U S A* **97**(12): 6579-6584.

Wang YP, Rekaya R. 2009. A comprehensive analysis of gene expression evolution between humans and mice. *Evol Bioinform* **5**: 81-90.

Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* **8**(1): 15-24.

Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ. 2006. Similar gene expression profiles do not imply similar tissue functions. *Trends Genet* **22**(3): 132-138.

Yang J, Su AI, Li WH. 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Mol Biol Evol* **22**(10): 2113-2118.

Table 3.1. Means and standard deviations of the EC distributions generated by different methods

Feature of the EC distributions	Liao and Zhang's method	Coexpression-based method				
		Dutilh et al.	ICC	Essien et al.		
				x=0.95	x=0.975	x=0.99
Human-mouse random gene pairs						
Mean	0.004	-0.003	0.002	-0.001	0.004	0.007
Standard deviation	0.217	0.177	0.313	0.192	0.225	0.300
Human-mouse 1-1 orthologs						
Mean	0.253	0.209	0.305	0.226	0.258	0.312
Standard deviation	0.332	0.199	0.321	0.217	0.254	0.327

Table 3.2. Correlations between Liao and Zhang's method and different coexpression-based methods

Correlation method	Dutilh et al.'s method	ICC	Essien et al.'s method		
			x=0.95	x=0.975	x=0.99
Pearson's correlation	0.498	0.456	0.514	0.523	0.510
Spearman's correlation	0.477	0.440	0.492	0.502	0.498

Table 3.3. Comparison of means of the EC distributions for human-mouse 1-1 orthologs based on the whole microarray data with the expression data over 26 common tissues by using coexpression-based methods

Coexpression-based methods	Mean of the EC distribution		P-value by two-sample *t*-test
	Whole microarray data	Data over 26 common tissues	
Dutilh et al.'s method	0.209	0.168	$< 2.2\times10^{-16}$
ICC	0.305	0.274	4.241×10^{-16}
Essien et al.'s method (x=0.975)	0.258	0.214	3.25×10^{-16}

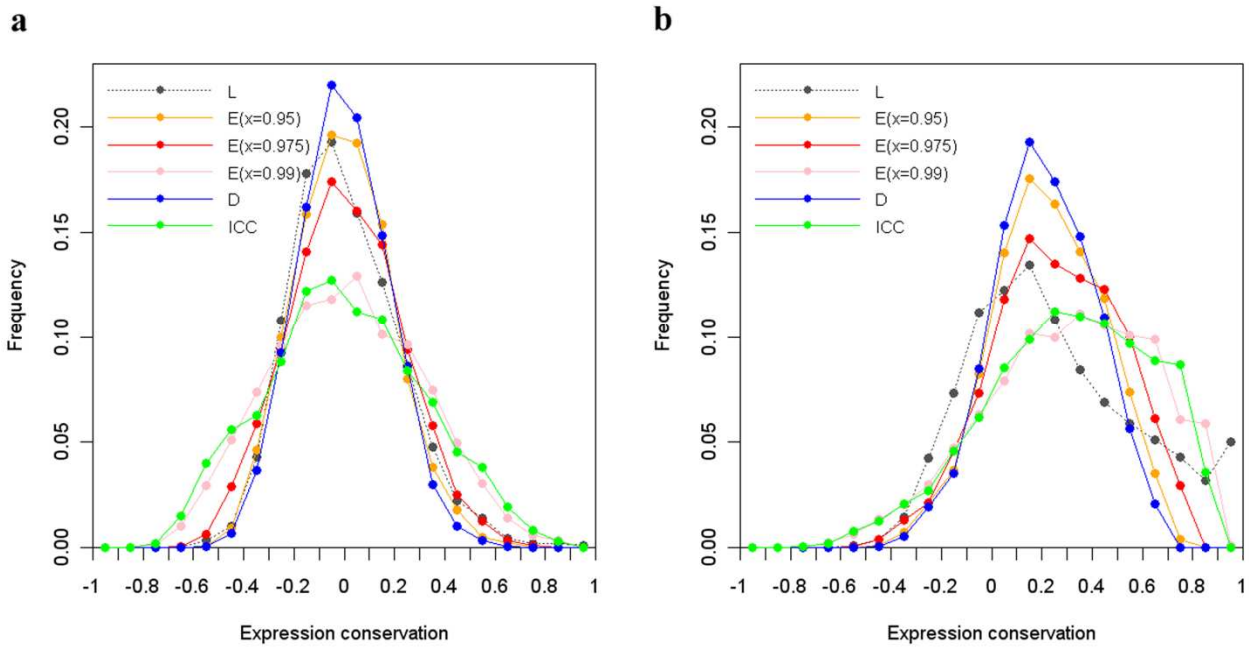

Figure 3.1. Comparison of the EC distributions for (a) human-mouse random gene pairs and (b) human-mouse 1-1 orthologs using Liao and Zhang's method (L), Dutilh et al.'s method (D), ICC and Essien et al.'s method (E).

CHAPTER 4

A COMPREHENSIVE ANALYSIS OF GENE EXPRESSION EVOLUTION BETWEEN HUMANS AND MICE[1]

[1]Yupeng Wang and Romdhane Rekaya. 2009. *Evol Bioinform Online*. 5: 81–90.

Reprinted here with permission from the publisher.

Abstract

Evolutionary changes in gene expression account for most phenotypic differences between species. Advances in microarray technology have made the systematic study of gene expression evolution possible. In this study, gene expression patterns were compared between human and mouse genomes using two published methods. Specifically, we studied how gene expression evolution was related to GO terms and tried to decode the relationship between promoter evolution and gene expression evolution. The results showed that 1) the significant enrichment of biological processes in orthologs of expression conservation reveals functional significance of gene expression conservation. The more conserved gene expression in some biological processes than is expected in a purely neutral model reveals negative selection on gene expression. However, fast evolving genes mainly support the neutrality of gene expression evolution, and 2) gene expression conservation is positively but only slightly correlated with promoter conservation based on a motif-count score of the promoter alignment. Our results suggest a neutral model with negative selection for gene expression evolution between humans and mice, and promoter evolution could have some effects on gene expression evolution.

Introduction

Comparative genomics adopts the assumption that important biological processes are often conserved across related species. Based on that, scientists use animal models to infer human physiological and genetic properties (Bedell et al. 1997; Sell 2003; Meuwissen and Berns 2005). Sequence comparison is the most popular tool for comparative genomics. However, sequence similarity is not necessarily proportional to functional similarity (Gerlt and Babbitt 2000). The biological functions of a gene not only rely on its molecular functions but also its spatiotemporal expression pattern. Changes in gene expression often mean changes in function

(Marques et al. 2008). One example is that, for duplicate genes, which are usually associated with highly consistent coding sequences but diverse biological functions, there is only a weak correlation between rates of sequence divergences and rates of expression divergences (Wagner 2000). It is urgent to make the details of gene expression evolution clear for the aim of making proper functional inferences across species.

Microarrays, which can characterize the transcriptional profiles of tens of thousands of genes simultaneously, have been widely used in biomedical (Alon et al. 1999; Golub et al. 1999; van 't Veer et al. 2002) and comparative genomic (Bergmann et al. 2004; Zhou and Gibson 2004; Lelandais et al. 2006) studies. In the latter applications, studies of gene expression levels in different species often rely on cross-species hybridization (Fortna et al. 2004; Khaitovich et al. 2004; Nuzhdin et al. 2004; Khaitovich et al. 2005). This method is limited to closely related species as it is based on the hybridization of target RNA and gene probes designed for other species (Oshlack et al. 2007), and when the probe and target RNA sequences are inconsistent to some extent, this method fails. Even in related species, several studies (Gilad et al. 2005; Bar-Or et al. 2006) found that this approach may be problematic.

Using microarray data, some theories on gene expression evolution across genomes have been suggested. Yanai et al. (2004) found that no expression conservation exists in human and mouse orthologous gene pairs because the evolution of expression profiles of orthologous gene pairs is comparable to that of randomly paired genes. Khaitovich et al. (2004) suggested that the majority of expression divergences between species are selectively neutral and are of no functional significance. The above two studies deviated from the ideas that genes should be expressed properly to conduct their functions and that basic biological processes are often conserved between related species. Jordan et al. (2005) suggested that gene expression

divergence among mammalian species is subject to the effects of purifying selective constraint, and it could also be substantially influenced by positive Darwinian selection. Liao and Zhang (2006) found that the expression profile divergence for the majority of orthologous genes between humans and mice is significantly lower than expected under neutrality and is correlated with the coding sequence divergence.

Another issue that should be addressed on the study of gene expression evolution is the relationship between promoter evolution and gene expression evolution. While the premise that the differences in upstream regulatory sequences represent gene expression divergence is widely accepted by researchers, several studies have shown that the changes in transcription factor binding sequences (TFBSs) have only little effect on gene expression evolution (Oda-Ishii et al. 2005; Fisher et al. 2006; Wang et al. 2007; Tirosh et al. 2008).

The diverse conclusions on gene expression evolution may be due, in part, to the improper comparisons of gene expression patterns across genomes. Expression data should not be compared across probes directly (Liao and Zhang 2006). Some scientists seek indirect methods, which can make the expression data comparable across probes and even across platforms or species. The conservation of gene coexpression patterns across species has been widely surveyed (Stuart et al. 2003; Ihmels et al. 2005; Singer et al. 2005; Oldham et al. 2006). However, coexpression shows little information on the expression conservation or evolution of orthologous genes across species. To overcome these obstacles, Liao and Zhang (2006) introduced the relative mRNA abundance among tissues (RA) and extracted 26 common tissues between humans and mice to make cross-species expression comparisons possible; Dutilh et al. (2006) and Tirosh and Barkai (2007) used either all or most one-to-one orthologs as referred sets for facilitating the gene expression comparisons across genomes.

In this study, we investigated several aspects of gene expression evolution between human and mouse genomes based on olignonucleotide microarray data of humans and mice generated by Su et al. (2004), which is widely used and is one of the largest data sets for humans and mice (Jordan et al. 2005; Yang et al. 2005; Liao and Zhang 2006; Liao and Zhang 2008). Two methods presented by Liao and Zhang (2006) and Dutilh et al. (2006) were adopted and compared for the aim of making reliable conclusions.

Methods

Microarray data and orthology

Human and mouse expression data were downloaded from GNF SymAtlas V1.2.4. (http://symatlas.gnf.org/SymAtlas/) by Su et al. (2004). This data set covers 79 human and 61 mouse tissues using the designed Affymetrix microarray chips (human: U133A&GNF1H; mouse: GNF1M). The expression levels were obtained using the MAS 5.0 procedure (Li and Wong 2001; Hubbell et al. 2002; Irizarry et al. 2003) as an average among replicates. To evaluate the reliability of our results, two additional data sets used by Su et al. (2002) (retrieved from the Gene Expression Omnibus database at the National Center for Biotechnology Information) and a yeast expression dataset by Spellman et al. (1998) (downloaded from http://genome-www.stanford.edu/cellcycle/data/rawdata/) were also analyzed.

The annotation files for GNF1H and GNF1M were downloaded from GNF SymAtlas along with the data files. The annotation file for U133A was downloaded from the Affymetrix website (http://www.affymetrix.com). To assign the Ensembl IDs for each gene, the annotation files (human:uniprot_sprot_human.dat.gz; mouse:uniprot_sprot_rodents.dat.) were downloaded from the Uniprot ftp site at (ftp://us.expasy.org/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/).

The orthologous pairs of human and mouse genes and human and yeast genes were downloaded from the Ensembl ftp site (ftp://ftp.ensembl.org/pub/release-47/mysql/compara_mart_homology_47/).

Only one-to-one orthologs were considered for our analyses. The orthologous genes with multiple probe sets were removed from our analyses. The numbers of human and mouse orthologous gene pairs used for this study were 4110 for the dataset by Su et al. (2004) and 1960 for the dataset by Su et al. (2002). The number of human and yeast orthologgous gene pairs was 577.

Comparison of gene expression patterns between genomes

Two procedures presented by Liao and Zhang (2006) (procedure I) and Dutilh et al. (2006) (procedure II) were used for comparing gene expression patterns between human and mouse genomes. For procedure I, the expression data of 26 common tissues from two species were extracted and normalized by their relative abundance (RA) values calculated by

$$RA_H(i,j) = S_H(i,j) / \sum_{j=1}^{n} S_H(i,j)$$

$$RA_M(i,j) = S_M(i,j) / \sum_{j=1}^{n} S_M(i,j)$$

where n is the number of common tissues, H represents humans, M represents mice, and $S_H(i,j)$ and $S_M(i,j)$ are the expression levels of gene i in human tissue j and mouse tissue j, respectively. Then the similarity of gene expression patterns for human and mouse gene i is calculated by

$$EC(i) = \frac{\sum_{j=1}^{n}[RA_H(i,j)RA_M(i,j)] - \frac{\sum_{j=1}^{n} RA_H(i,j)\sum_{j=1}^{n} RA_M(i,j)}{n}}{\sqrt{(\sum_{j=1}^{n}[RA_H(i,j)]^2 - \frac{[\sum_{j=1}^{n} RA_H(i,j)]^2}{n})(\sum_{j=1}^{n}[RA_M(i,j)]^2 - \frac{[\sum_{j=1}^{n} RA_M(i,j)]^2}{n})}}.$$

For procedure II, all one-to-one orthologs between humans and mice were used as references. Then the similarity (r_i) of gene expression patterns for human and mouse gene *i* is obtained by correlating the expression correlation values of gene *i* from two different species and the corresponding one-to-one orthologs in their species.

Calculating Z-scores for GO Terms

For each GO term or gene family, the Z-scores were calculated as:

$$Z = \frac{\overline{r}_s - \overline{r}_p}{sd_p / \sqrt{n}}$$

where $\overline{r}_s$ is the mean of correlation values of the orthologs in this GO term, $\overline{r}_p$ and sd_p are the mean and standard deviation of correlation values for all available orthologs, and *n* is the number of the available members in this GO term.

The motif-count score in the alignment of promoter regions

We proposed a motif-count score of the pairwise alignment of promoter sequences, which can be easily derived from the local alignment for two sequences. The promoter sequence was defined as −1000 and +200 bp of the TSS for this study. The matrix for local alignment was constructed with no gap or mismatch allowed. The local alignments with lengths >4 were regarded as conserved DNA motifs/sequences. Each conserved DNA motif/sequence was assigned a score that equaled its length minus 4. The motif-count score for a pairwise alignment of promoter sequences was calculated by summing up the scores of all conserved DNA motifs/sequences in the matrix. Although it is true that some of the conserved DNA motifs/sequences are not true transcription factor binding sites, it is reasonable to assume that the motif-count score based on conserved sequences is generally proportional to that based on true DNA motifs. Thus, the motif-count score allows, in general terms, to measure the similarity of

the composition of multiple DNA motifs in two promoter sequences and could help infer biologically the similarity of regulatory patterns for two promoters.

dN/dS ratio

The nonsynonymous substitution rates (dN), synonymous substitution rates (dS), and their ratio (dN/dS) were used to represent the rates of coding sequence evolution, which were retrieved from the Ensembl ftp web site (ftp://ftp.ensembl.org/pub/release-47/mysql/compara_mart_homology_47/). The ratio of dN/dS is an indicator of selective pressures on coding sequence evolution, where $dN/dS > 1$ indicates that genes are under positive selection pressure while $dN/dS < 1$ indicates stabilizing selection.

Results

Identification of gene expression conservation in orthologs

By using procedure I and procedure II, the correlations of expression profiles for 4110 human and mouse orthologous genes pairs and random gene pairs were calculated. The results confirmed the theory of non-random expression conservation of orthologs (data not shown), which has been explored by several studies.22,31,32 At the significance level of 1% of genomic background, procedure I and procedure II identified 727 and 559 orthologous gene pairs of expression conservation, which were used for the following functional enrichment analysis of gene expression evolution.

Analyses of gene expression evolution in terms of biological functions

For orthologous gene pairs that were identified by procedure I and procedure II, we conducted an overrepresented GO term analysis to the human genes using GOstat (Beissbarth and Speed 2004). The P value was set at 0.05. The 727 and 559 orthologous gene pairs with expression conservation identified by procedure I and procedure II resulted in 18 and 10

overrepresented terms, respectively (Table 4.1). The above analysis indicates that the conservation of gene expression has functional significance.

We also investigated whether there are overrepresented GO terms in the human and mouse orthologous genes of fast expression evolution. For that purpose, we retrieved the orthologous genes with the bottom 5% correlation values identified by procedure I and II, respectively. No overrepresented GO terms were returned for these genes. The lack of GO term enrichment in fast evolving genes may be interpreted as evidence for the neutrality of expression evolution. But note that adaptation could involve only few or single genes and does not necessarily require the simultaneous evolution of the expression of the entire GO terms.

To further validate the evolutionary model of gene expression, we investigated how all available GO terms affected gene expression evolution. We took all the orthologous gene pairs as a population and grouped orthologous gene pairs by GO terms. We selected the GO terms with no less than three members and tested 320 terms in all. For each term, we got a Z-score for the mean correlation. Theoretically, these Z-scores should follow a standard normal distribution if no selection exists (note that we removed the GO terms with only one or two members because the means of small size samples may not form the normal distribution if the population does not agree with an exact normal distribution). We plotted the distribution of Z-scores of GO terms against a standard normal distribution (Figure 4.1). Generally, the curves formed by procedure I and II fit the neutral model. The distribution of Z-scores for procedure I or II tends to have a heavier right tail (the part of the Z-score >1.96) compared to the control, suggesting that a small part of GO terms have negative selection on gene expression. However, a left heavier tail (the part of the Z-score <−1.96) is not observed, suggesting that generally GO terms do not have obvious positive selection on gene expression.

Gene expression evolution is slightly correlated with promoter evolution between humans and mice

It is widely accepted by researchers that promoter differences represent regulatory differences, which are reflected by gene expression divergence. However, several studies have indicated that extensively divergent promoters from species may still maintain the same expression patterns (Oda-Ishii et al. 2005; Fisher et al. 2006; Wang et al. 2007), which suggests the neutrality of promoter evolution. Zhang et al. (2004) found that changes in TFBSs were poorly correlated with divergence of gene expression among yeast paralogs. Tirosh et al. (2008) argued that previously identified TFBS of yeasts and mammals had no detectable effect on gene expression. One reason for no detectable or poor correlation may be that an underlying compensatory mechanism allows promoters to rapidly evolve while maintaining a stabilized expression pattern. However, other possibilities should also be considered, e.g. the inherent complexity of promoters, limited data on identified transcription factor binding sites, a suboptimal evolutionary model for promoters, noise of microarray data and improper comparisons of gene expression between species. Thus, it is necessary to reexamine this relationship using new models.

Functional DNA motifs in promoters are often under selection pressure and seem more conserved between species than non functional DNA sequences. Thus, the evolutionary mechanisms of promoters may accommodate different models compared to the model of neutral evolution subject to purifying selection adopted by coding sequences. In addition, it is important to properly designate the similarity between promoters, which will reflect the similarity of gene regulatory patterns besides the sequence similarity. Here we consider three methods for comparing promoter sequences: global alignment, local alignment and our proposed motif-count

score method. The global alignment score tends to reflect more the promoter conservation as a whole sequence. The motif-count score of alignments tends to reflect more the conservation of composite DNA motifs by disregarding their positions in the promoters. The local alignment score is somewhat a compromise of the previous two methods.

Scores based on global alignment, local alignment and motif-count for all orthologs were calculated. Their correlations with gene expression conservations were 0.014 (P value = 0.3772), 0.016 (P value = 0.3087) and 0.055 (P value = 0.0006525), respectively using procedure I; the correlations from procedure II were 0.025 (P value = 0.1218), 0.030 (P value = 0.06608) and 0.040 (P value = 0.01205). With both procedures, the motif-count score method resulted in a slightly positive and significant correlation between promoter conservation and gene expression conservation. The increase of promoter-expression correlation using our proposed motif-count scores suggests it has improved in describing promoter conservation. To reduce the effects of noise in microarray data, we retrieved the most reliable conserved expression (top 10% r_i) and diverged expression (bottom 10% r_i) for analysis. An obvious decrease in motif-count scores from conserved expression to diverged expression is seen in Figure 4.2 (P values of two sample t test are 0.00289 and 0.003665 for procedure I and II, respectively). The promoter-expression correlations based on these reliable expression patterns were 0.103 (P value = 0.004152) and 0.122 (P value = 0.0006919) using procedures I and II, respectively, indicating a reasonable predictive power of motif-count scores to determine the variability in expression conservation.

From this analysis, it is reasonable to infer that there still could be space for detecting larger promoter-expression correlation if optimal models for describing promoter evolution are used. An optimal model for describing promoter evolution should consider the different evolutionary mechanisms within functional DNA motifs and non functional sequences and the

combinational regulatory effects of composite DNA motifs. In this sense, the promoter evolution could indeed affect the gene expression evolution to some extent.

Reanalyzing gene expression evolution by using other datasets and species

To investigate whether the above conclusions were affected by the choice of the used gene expression dataset, we reanalyzed an additional large microarray dataset used by Su et al. (2002). In total, 1960 pairs of human and mouse orthologs were analyzed. At a significance level of 1% of genomic background, procedures I and II identified 306 and 278 orthologous gene pairs of expression conservation, which confirmed the theory of non-random expression conservation of orthologs. These gene pairs with expression conservation identified by both procedures resulted in 19 and 14 overrepresented GO terms at P value < 0.05, respectively while fast evolving genes had no overrepresented GO terms. The correlations of promoter conservation based on global alignment, local alignment and motif-count scores with gene expression conservation were 0.039 (P value = 0.0967), 0.040 (P value = 0.09138) and 0.058 (P value = 0.01388), respectively using procedure I; 0.034 (P value = 0.1507), 0.040 (P value = 0.08712) and 0.065 (P value = 0.00582), respectively using procedure II.

In addition, we investigated whether these conclusions on gene expression evolution are held when comparing distant species such as humans and yeast. For that purpose, human expression data used by Su et al. (2004) and the yeast cell cycle expression data used by Spellman et al. (1998) were analyzed. Note that only procedure II can be employed for this analysis. The number of human-yeast one-to-one orthologous pairs was 577. At the significance level of 1% of genomic background, 94 orthologs were identified as having conserved expression, suggesting that the theory of non-random expression conservation of orthologs should be true between humans and yeast. These 94 orthologs returned 11 overrepresented GO

terms at P value < 0.05 using GOstat (Beissbarth and Speed 2004). Fast evolving genes (bottom 10% r_i) returned no overrepresented GO terms. The above analysis suggests that gene expression conservation has functional significance in both related species and distant species. Finally, we investigated whether gene expression conservation is correlated with promoter conservation. No significant correlation was obtained between r_i and global alignment score (correlation: −0.008, P value = 0.8477), local alignment score (correlation: −0.044, P value = 0.2888) or motif-count score (correlation: −0.067, P value = 0.1056). These results indicate that the weak correlation between promoter conservation and gene expression conservation is not maintained between humans and yeast, which is contrary to the conclusion between humans and mice. The explanation for this finding could be that gene regulatory patterns by DNA motifs may be similar between humans and mice and thus allow their weak correlation with gene expression patterns while gene regulatory patterns are too different between humans and yeast to be correlated with gene expression patterns.

Discussion

In this study, we analyzed gene expression evolution for orthologs based on human and mouse models. Based on our results, it is reasonable to assume some functional significance for orthologs with expression conservation and neutrality for orthologs with fast expression evolution. Thus, a neutral model with negative selection for gene expression evolution may best explain our results. Additionally, we found a weak correlation between promoter conservation and gene expression conservation. These analyses reveal the inherent complexity of gene expression evolution.

Our neutral model for gene expression evolution differs from previous studies in that the functional significance in gene expression evolution is largely neutral except in some conserved

expression patterns; in addition, our model does not mean that gene expression evolution will be well correlated with evolutionary divergence time, evidenced by the fact that determining whether there is a correlation between gene expression divergence and coding sequence divergence is very conflicting in previous studies (Yanai et al. 2004; Jordan et al. 2005; Khaitovich et al. 2005; Dutilh et al. 2006; Liao and Zhang 2006; Tirosh and Barkai 2008). There could be a possibility that different genes may use different tempos of gene expression evolution with unknown determining factors.

Tirosh et al. (2008) tested the changes of DNA motifs in the promoter region to find out if they were correlated with expression divergence and found no detectable correlation. The failure of detecting significant correlations could be due to the limited number of known DNA motifs compared to the unknown true number or/and the lack of proper models for multiple DNA motifs. Zhang et al. (2004) used a regression model of multiple DNA motifs to account for gene expression. Although Zhang et al. (2004) addressed the promoter-expression correlations based on paralogs while our study was based on orthologs, the conclusions are very similar, suggesting that there could be some mechanisms that promoter evolution affects gene expression evolution.

In this study, the correlations between coding sequence evolution and promoter evolution range from −0.034 to 0.210. Although these correlations may be significant, coding sequence evolution cannot fully account for promoter evolution. Thus, there could be other evolutionary mechanisms in promoters besides nucleotide mutations. We hypothesize that one mechanism may involve mainly the duplication and transposition of DNA motifs, which have been suggested by two previous studies (Johnson et al. 2006; Kim et al. 2007). This mechanism may affect gene expression evolution. Our proposed motif-count score reflects some information

relative to this mechanism, which may contribute to the detection of promoter-expression association. In addition, two recent studies (Field et al. 2009; Park and Makova 2009) indicated that the evolution of DNA-encoded nucleosome organization and turnover of transcription start sites in promoters may also affect gene expression evolution. We infer that a proper model of promoter evolution considering all mechanisms may be found strongly associated with gene expression evolution.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* **96**(12): 6745-6750.

Bar-Or C, Bar-Eyal M, Gal TZ, Kapulnik Y, Czosnek H, Koltai H. 2006. Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. *BMC Genomics* **7**: 110.

Bedell MA, Jenkins NA, Copeland NG. 1997. Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes Dev* **11**(1): 1-10.

Beissbarth T, Speed TP. 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**(9): 1464-1465.

Bergmann S, Ihmels J, Barkai N. 2004. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol* **2**(1): E9.

Dutilh BE, Huynen MA, Snel B. 2006. A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics* **7**: 10.

Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E. 2009. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat Genet* **41**(4): 438-445.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**(5771): 276-279.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T *et al.* 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* **2**(7): E207.

Gerlt JA, Babbitt PC. 2000. Can sequence determine function? *Genome Biol* **1**(5): REVIEWS0005.

Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res* **15**(5): 674-680.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.* 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**(5439): 531-537.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18**(12): 1585-1592.

Ihmels J, Bergmann S, Berman J, Barkai N. 2005. Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet* **1**(3): e39.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2): 249-264.

Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, Westhead DR, Wood IC, Jackson RM, Buckley NJ. 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res* **34**(14): 3862-3877.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* **345**(1): 119-126.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**(5742): 1850-1854.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol* **2**(5): E132.

Kim JD, Faulk C, Kim J. 2007. Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1. *Nucleic Acids Res* **35**(10): 3442-3452.

Lelandais G, Vincens P, Badel-Chagnon A, Vialette S, Jacq C, Hazout S. 2006. Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms. *Bioinformatics* **22**(11): 1359-1366.

Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**(1): 31-36.

Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**(19): 6987-6992.

Liao BY, Zhang JZ. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**(3): 530-540.

Marques AC, Vinckenbosch N, Brawand D, Kaessmann H. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol* **9**(3): R54.

Meuwissen R, Berns A. 2005. Mouse models for human lung cancer. *Genes Dev* **19**(6): 643-664.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol* **21**(7): 1308-1317.

Oda-Ishii I, Bertrand V, Matsuo I, Lemaire P, Saiga H. 2005. Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of *Otx* between the ascidians *Halocynthia roretzi* and *Ciona intestinalis*. *Development* **132**(7): 1663-1674.

Oldham MC, Horvath S, Geschwind DH. 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A* **103**(47): 17973-17978.

Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2007. Using DNA microarrays to study gene expression in closely related species. *Bioinformatics* **23**(10): 1235-1242.

Park C, Makova KD. 2009. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. *Genome Biol* **10**(1): R10.

Sell S. 2003. Mouse models to study the interaction of risk factors for human liver cancer. *Cancer Res* **63**(22): 7553-7562.

Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* **22**(3): 767-775.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **9**(12): 3273-3297.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**(5643): 249-255.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A *et al.* 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**(7): 4465-4470.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al.* 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**(16): 6062-6067.

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* **8**(4): R50.

Tirosh I, Barkai N. 2008. Evolution of gene sequence and gene expression are not correlated in yeast. *Trends Genet* **24**(3): 109-113.

Tirosh I, Weinberger A, Bezalel D, Kaganovich M, Barkai N. 2008. On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol* **4**: 159.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al.* 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(6871): 530-536.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc Natl Acad Sci U S A* **97**(12): 6579-6584.

Wang QF, Prabhakar S, Chanan S, Cheng JF, Rubin EM, Boffelli D. 2007. Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol* **8**(1): R1.

Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* **8**(1): 15-24.

Yang J, Su AI, Li WH. 2005. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Mol Biol Evol* **22**(10): 2113-2118.

Zhang ZQ, Gu JY, Gu X. 2004. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet* **20**(9): 403-407.

Zhou XJ, Gibson G. 2004. Cross-species comparison of genome-wide expression patterns. *Genome Biol* **5**(7): 232.

Table 4.1. The overrepresented GO terms in human and mouse orthologs with expression conservation

GO term	Description	Number of genes	P value
Procedure I			
GO:0005856	cytoskeleton	52	0.0324
GO:0005624	membrane fraction	43	0.0434
GO:0015629	actin cytoskeleton	18	0.0434
GO:0005509	calcium ion binding	53	0.0434
GO:0004867	serine-type endopeptidase inhibitor activity	11	0.0434
GO:0044430	cytoskeletal part	34	0.0434
GO:0000267	cell fraction	58	0.0434
GO:0016052	carbohydrate catabolic process	15	0.0434
GO:0009605	response to external stimulus	52	0.0434
GO:0006936	muscle contraction	18	0.0434
GO:0007286	spermatid development	7	0.0434
GO:0016491	oxidoreductase activity	54	0.0434
GO:0006941	striated muscle contraction	6	0.0434
GO:0006006	glucose metabolic process	15	0.0434
GO:0008236	serine-type peptidase activity	18	0.0434
GO:0019318	hexose metabolic process	18	0.0434
GO:0019320	hexose catabolic process	11	0.0445

GO:0048232	male gamete generation	21	0.0469
Procedure II			
GO:0048232	male gamete generation	21	0.00054
GO:0043232	intracellular non-membrane-bounded organelle	74	0.00054
GO:0019953	sexual reproduction	25	0.00068
GO:0006996	organelle organization	52	0.00322
GO:0007276	gamete generation	22	0.00433
GO:0007286	spermatid development	7	0.00913
GO:0051276	chromosome organization	22	0.00917
GO:0006323	DNA packaging	19	0.0161
GO:0016585	chromatin remodeling complex	7	0.02
GO:0016043	cellular component organization	101	0.0464

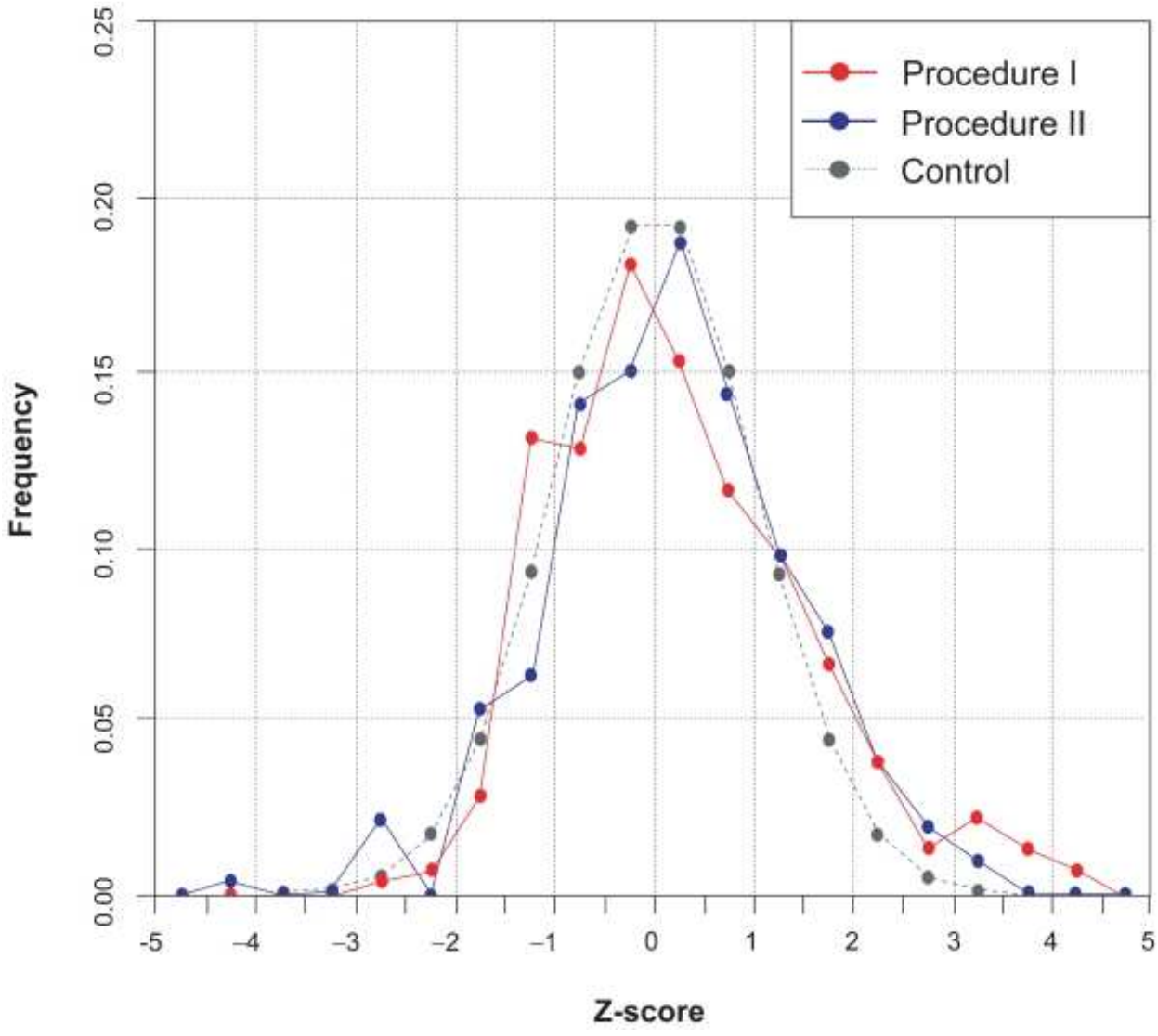

Figure 4.1. Distribution of Z-scores for GO terms.

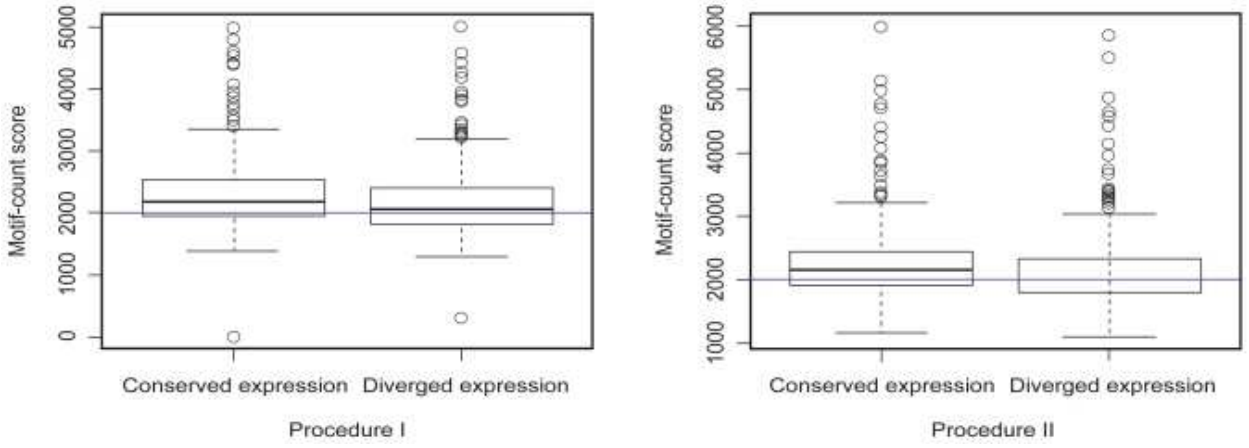

Figure 4.2. Comparison of the motif-count scores between conserved expression and diverged expression.

CHAPTER 5

MCSCANX: A TOOLKIT FOR DETECTION AND EVOLUTIONARY ANALYSIS OF GENE SYNTENY AND COLLINEARITY[1]

[1]Yupeng Wang, Haibao Tang, Jeremy D. DeBarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-ho Lee, Huizhe Jin, Barry Marler, Hui Guo, Jessica C. Kissinger and Andrew H. Paterson. Submitted to *Nucleic Acids Research*, 09/22/11.

Abstract

MCScan is an algorithm able to scan multiple genomes or subgenomes in order to identify putative homologous chromosomal regions, and align these regions using genes as anchors. The *MCScanX* toolkit implements an adjusted MCScan algorithm for detection of synteny and collinearity that extends the original software by incorporating 14 utility programs for visualization of results and additional downstream analyses. Applications of *MCScanX* to several sequenced plant genomes and gene families are shown as examples. *MCScanX* can be used to effectively analyze chromosome structural changes, and reveal the history of gene family expansions that might contribute to the adaptation of lineages and taxa. An integrated view of various modes of gene duplication can supplement the traditional gene tree analysis in specific families. The source code and documentation of *MCScanX* are freely available at http://chibba.pgml.uga.edu/mcscan2/.

Introduction

Comparative genomic studies often rely on the accurate identification of homology (genes that share a common evolutionary origin) within or across genomes. Homologous genes are further classified as either orthologous, if they were separated by a speciation event, or paralogous, if they were separated by a gene duplication event within the same genome. Recently, comparisons between related eukaryotic genomes reveal various degrees to which homologous genes remain on corresponding chromosomes (synteny) and in conserved orders (collinearity) during evolution (Coghlan et al. 2005). Over evolutionary time, genomes have been shaped and dynamically restructured by several forces such as whole-genome duplication (WGD), segmental duplication, inversions and translocations (Dietrich et al. 2004; Dujon et al. 2004; Nakatani et al. 2007; Salse et al. 2009). These forces have acted in various combinations

and to differing degrees to result in taxonomic groups with different modes of genome structure modification and gene family expansion. For example, angiosperm (flowering plants) genomes appear more volatile than mammalian genomes (Coghlan et al. 2005). Angiosperm genomes show remarkable fluctuations in size and organization, even among close relatives, and all examined angiosperms have undergone one or more ancient WGD (Bowers et al. 2003). In contrast, karyotype evolution among major vertebrate lineages appears to have been slower, with a single whole-genome duplication event ~500 million years ago (Knight et al. 2005). However, hundreds of invertebrates are paleopolyploids (Otto and Whitton 2000) and their rates of chromosomal rearrangement have been suggested to be almost twice that of vertebrates (Ranz et al. 2001; Bourque et al. 2004; Coghlan et al. 2005). Further, there is also a remarkable lack of synteny and high rate of rearrangement in the parasitic and pathogenic protistan phylum Apicomplxa compared to what is seen in vertebrates (Debarry and Kissinger 2011).

Traditionally, synteny was identified via the clustering of neighboring matching gene pairs, as implemented in various programs including ADHoRe (Vandepoele et al. 2002), LineUp (Hampson et al. 2003), the Max-gap Clusters by Multiple Sequence Comparison (MCMuSeC) (Ling et al. 2009) and OrthoCluster (Vergara and Chen 2009). However, detection of synteny is often complicated by gene loss, tandem duplications, gene transpositions and chromosomal rearrangements, any of which may produce artifacts. Collinearity, a more specific form of synteny, requires conserved gene order. More recent methods apply dynamic programming to chains of pair-wise collinear genes, and often specify a certain scoring scheme that rewards the adjacent collinear gene pairs (or "anchor genes") and penalizes the distance between anchor genes. This class of methods has been implemented in software tools such as DAGchainer (Haas et al. 2004), ColinearScan (Wang et al. 2006b), MCScan (Tang et al. 2008b), SyMAP (Soderlund

et al. 2006), FISH (Calabrese et al. 2003) and CYNTENATOR (Rodelsperger and Dieterich 2010). In addition to algorithmic differences, synteny and collinearity detection tools often differ in application ranges, inputs, presentation of results and/or computational costs.

Although pair-wise collinear relationships among chromosomal regions have been widely studied, the multi-alignment (alignment of 3 or more regions) of collinear chromosomal regions (referred to as collinear blocks) is more important as it can reveal ancient WGD events (Tang et al. 2008a; Tang et al. 2008b) and complex chromosomal duplication/rearrangement relationships (Abrouk et al. 2010). Collinear blocks are comprised of anchor genes which are located at collinear positions and non-anchor genes which are assumed to have experienced gene gains, losses or transposition. Further, anchor genes are more likely to be homologs (Jun et al. 2009) and tend to be under stronger purifying selection than non-anchor genes (Casneuf et al. 2006). Patterns of synteny and collinearity can provide insight into the evolutionary history of a genome, and inform on potentially useful downstream analyses. However, although graphic interfaces for visualizing synteny and collinearity may be incorporated, many available software packages for synteny and collinearity detection do not directly provide downstream analysis tools. Further, genes may be duplicated by mechanisms other than whole-genome duplication, such as tandem, proximal and/or dispersed duplications, each of which may make different contributions to evolution (Freeling 2009; Debarry and Kissinger 2011). In addition, analysis of gene family evolution may require that it be placed in the context of genome evolution. To analyze the evolution of a genome, it may be helpful to correlate gene family analysis with different duplication modes for a more integrated view. To our knowledge, only the *MicroSyn* package (Cai et al. 2011) provides analysis of collinearity within gene families, but it cannot superimpose such analysis on a context of whole-genome collinearity.

MCScan is able to identify collinear blocks in genomes or subgenomes and then conduct multi-alignments of collinear blocks using collinear genes as anchors (Tang et al. 2008a; Tang et al. 2008b). MCScan is also customizable for genomes of different sizes and with different average intergenic distances. Using MCScan, a Plant Genome Duplication Database (PGDD) has been constructed and is publicly available at http://chibba.pgml.uga.edu/duplication/. The MCScan software package and PGDD database have been applied to a variety of research areas such as genome duplication and evolution (Lyons et al. 2008; Charles et al. 2009; Wang et al. 2009b; Lin et al. 2010; Tang et al. 2010; Debarry and Kissinger 2011; Lin et al. 2011; Wang et al. 2011), annotation of newly sequenced genomes (Paterson et al. 2009) and the evolution of gene families (Watanabe et al. 2008; Kopriva et al. 2009; Li et al. 2009; Okazaki et al. 2009; Wang et al. 2009a; Causier et al. 2010; Higgins et al. 2010; Hyun et al. 2010; Knoller et al. 2010; Li and Zhang 2011; Palmieri et al. 2011).

Building on the MCScan algorithm, here we describe a software package named *MCScanX* for synteny and collinearity detection, visualization and diverse downstream analyses. Compared with MCScan, the usage of *MCScanX* has been greatly simplified. To more clearly show how frequently chromosomal regions are duplicated, multi-alignments of collinear blocks against reference chromosomes can be viewed through a web browser with various highlighted features (e.g. tandem arrays, coverage statistics). The overall pattern of synteny and collinearity between or among genomes can be visualized by up to four types of plots. Compared with existing synteny and collinearity detection tools, a distinct feature of *MCScanX* is that diverse tools for evolutionary analyses of synteny and collinearity are incorporated, aiding efforts to construct gene families using collinearity information, infer gene duplication modes and enrichments, characterize collinear genes with nucleotide substitution rates, detect collinear

tandem arrays, perform statistical analyses of duplication depths and collinear orthologs, and analyze collinearity within gene families.

Materials and methods

Gene set and homology search

Whole-genome protein sequences and gene positions for *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, *Glycine max*, *Oryza sativa,* and *Brachypodium distachyon* were retrieved from Phytozome v7.0 (http://www.phytozome.net/). Whole-genome protein sequences and gene positions for *Sorghum bicolor* and *Zea mays* were retrieved from EnsemblPlants (http://plants.ensembl.org/index.html) and MaizeSequence Release 5b.60 (http://www.maizesequence.org/index.html) respectively. If a gene had more than one transcript, only the first transcript in the annotation was used. To search for homology, the protein-coding genes from each genome was compared against itself and other genomes using BLASTP (Altschul et al. 1990). For a protein sequence, the best five non-self hits in each target genome that met an *E*-value threshold of 10^{-5} were reported.

MCScanX algorithm

The *MCScanX* algorithm is a modified version of MCScan (Tang et al. 2008b). Whole-genome BLASTP results are used to compute collinear blocks for all possible pairs of chromosomes and scaffolds. First, BLASTP matches are sorted according to gene positions. To avoid high numbers of local collinear gene pairs due to tandem arrays, if consecutive BLASTP matches have a common gene and its paired genes are separated by fewer than 5 genes, these matches are collapsed using a representative pair with the smallest BLASTP *E*-value. Then, dynamic programming is employed to find the highest scoring paths (i.e. chains of collinear gene

pairs) using the following scoring schema, assuming that two gene pairs, *u* and *v*, are on the path where *u* precedes *v*,

$$ChainScore(v) = MatchScore(v) + \max\{ChainScore(u) + GapPenalty * NumberofGaps(u,v), 0\}$$

where by default $MatchScore(v) = 50$ for one gene pair, $GapPenalty = -1$, and $NumberofGaps(u,v)$, the maximum number of intervening genes between u and v, should be fewer than 25. Non-overlapping chains with scores over 250 (i.e. involving at least 5 collinear gene pairs) are reported. In a pair of collinear blocks, there are two distinct genomic locations with aligned collinear genes as anchors.

The expected number of occurrences (*E*-value) of a pair of collinear blocks is estimated using the formula introduced by Wang et al. (Wang et al. 2006b),

$$E = 2P_N^m \prod_{i=1}^{m-1} (\frac{l_{1i}}{L_1} \cdot \frac{l_{2i}}{L_2})$$

where *N* is the number of matching gene pairs between the two chromosomal regions defined by the pair of collinear blocks, *m* is the number of anchors in the pair of collinear blocks, L_1 and L_2 are respective lengths of the two chromosomal regions, and l_{1i} and l_{2i} are distances (in terms of nucleotide numbers) between two adjacent anchors in the pair of collinear blocks. The default *E* value cutoff of *MCScanX* is 10^{-5}.

Multiple chromosomal regions threaded by consecutive ancestral loci are progressively aligned against reference chromosomes, where each genome being tested is used as a reference successively, according to the following procedure: 1) any reference chromosome is scanned from start to end, and empty tracks are placed alongside the reference chromosome to hold potential aligned collinear blocks; 2) Collinear blocks are progressively aligned against reference chromosomes pinpointed by anchors and assigned to the nearest empty tracks (once a track region is filled, it cannot be assigned collinear blocks again). In aligned collinear blocks, only

symbols of anchor genes are shown while un-matched positions (gaps) between anchors (regardless of numbers of intervening genes) are denoted by "||"; 3) at each locus of reference chromosomes, the number of tracks occupied by collinear blocks is recorded to reflect the duplication depth.

Classification of duplicate gene origins

Genes within a single genome can be classified as singletons, dispersed duplicates, proximal duplicates, tandem duplicates and segmental / WGD duplicates depending on their copy number and genomic distribution. The following procedure is used to assign gene classes: 1) All genes are initially classified as "singletons" and assigned gene ranks according to their order of appearance along chromosomes; 2) BLASTP results are evaluated and the genes with BLASTP hits to other genes are re-labeled as "dispersed duplicates"; 3) In any BLASTP hit, the two genes are re-labeled as "proximal duplicates" if they have a difference of gene rank<20 (configurable); 4) In any BLASTP hit, the two genes are re-labeled as "tandem duplicates" if they have a difference of gene rank=1; 5) *MCScanX* is executed. The anchor genes in collinear blocks are re-labeled as "WGD/segmental". So, if a gene appears in multiple BLASTP hits, it will be assigned a unique class according to the order of priority: WGD/segmental>tandem>proximal>dispersed.

Detection of orthologous gene pairs using OrthoMCL

Whole-genome protein sequences from *Arabidopsis*, *Populus*, *Vitis*, *Glycine*, *Oryza, Brachypodium*, *Sorghum* and *Zea* were merged and searched against themselves for homology using BLASTP with an *E*-value cutoff of 10^{-5}. Default parameters of OrthoMCL (Li et al. 2003) were used. The combination of OrthoMCL intermediate files "orthologs.txt" and

"coorthologs.txt" (generated by *orthomclDumpPairsFiles*) was used as the whole set of ortholog pairs.

Enrichment analysis

Enrichment analysis is performed using Fisher's exact test. The *P*-value was calculated for the null hypothesis that there is no association between the members of a gene family and a particular gene duplication mode and is corrected with the total number of duplication modes for multiple comparisons (i.e. Bonferroni correction). The *P*-value cutoff of 0.05 is used to suggest putative enrichment of certain gene duplication modes.

Computing Ka and Ks

Non-synonymous (Ka) and synonymous (Ks) substitution rates are estimated by Nei-Gojobori statistics (Nei and Gojobori 1986), available through the "Bio::Align::DNAStatistics" module of the BioPerl package (http://www.bioperl.org/wiki/Module:Bio::Align::DNAStatistics). Note that the "Bio::Align::DNAStatistics" module may generate invalid Ka or Ks (i.e. non-digital output) for some homologous gene pairs due to mis-alignments.

Gene family examples

Lists of published *Arabidopsis* gene families were obtained from TAIR (http://www.arabidopsis.org/browse/genefamily/index.jsp). Only families with more than 9 genes were considered in order to have enough statistical power to detect enrichment of duplication modes. *Arabidopsis* disease resistance gene homologs were downloaded from the NIBLRRS Project website (http://niblrrs.ucdavis.edu).

Execution of the *MCScanX* package

MCScanX is freely available at http://chibba.pgml.uga.edu/mcscan2. All programs in the *MCScanX* package should be executed using command line arguments on Mac OS or Linux systems. On Mac OS, Xcode (http://developer.apple.com/xcode/) should be installed prior to the installation of *MCScanX* package. On Linux systems, the Java SE Development Kit (JDK) and "libpng" should be installed before the installation of *MCScanX* package. To list available command line options, the user can simply type the name of a program without any options.

Results

Structure of the *MCScanX* package

The *MCScanX* package consists of two main components: 1) three core programs that implement an adjusted MCScan algorithm to generate pairwise and multiple alignments of collinear blocks and 2) twelve downstream analysis programs for displaying and analyzing identified synteny and collinearity output by the core programs. The structure of the *MCScanX* package is shown in Figure 5.1. Compared with the previous version (0.8) of MCScan, there are numerous improvements in *MCScanX*. First, preprocessing of BLASTP input has been pipelined into the execution of core programs. Next, in MCScan, each gene was assigned a family ID to identify tandem genes, where the family ID has to be pre-computed using the Markov Clustering Algorithm (MCL) software (Enright et al. 2002). In *MCScanX*, tandem genes are assessed by gene rank according to chromosomal positions and thus, execution of MCL is no longer required. The aforementioned two improvements have made the installation and execution of *MCScanX* easier and more efficient. Furthermore, multi-alignments of collinear blocks, which are output as HTML files in *MCScanX*, can be easily and clearly viewed. In addition, numerous visualization and downstream analysis tools are incorporated into the *MCScanX* package, greatly enhancing

the biological applications of the MCScan algorithm. In the following, we describe in detail each program in the *MCScanX* package.

The first core program, named *MCScanX*, can generate both pair-wise and multiple alignments of collinear blocks, similar to the previous MCScan version (0.8). However, *MCScanX* takes only a simplified GFF format file and a BLASTP tabular file as inputs. The simplified GFF file should contain the gene locations (which include chromosome, gene symbol, start and end) for the genomes to be compared. The BLASTP input file is one BLASTP output or combined multiple BLASTP outputs in tabular format (option "-m8" in BLAST and "–outfmt 6" in BLAST+) for all protein sequences in the species of interest. Note that when *MCScanX* is applied to multiple species, it may be useful to guard against over-enrichment of gene pairs from closely related species and we recommend that the BLASTP input file include the combined BLASTP outputs of pairwise genome comparisons and self-genome comparisons with a cutoff of best hits instead of a single BLASTP output of pooled protein sequences from different species. Alternatively, the BLASTP input can be replaced by a tab delimited file containing pair-wise homologous relationships detected by third party software. In this case, the user needs to implement *MCScanX_h* (the second core program). In addition, *MCScanX_h* can generate statistics on numbers of collinear homolog pairs and their percentages (relative to the numbers of input homolog pairs).

We also adopted an adjusted MCScan algorithm. Matches among genes are first sorted according to chromosomal positions for all possible pairs of chromosomes and scaffolds, and in both transcriptional directions. Adjacent collinear genes are chained using dynamic programming (see Methods), outputting pairwise collinear blocks and tandem gene pairs to ".collinearity" and ".tandem" files respectively. Note that during the chaining of collinear genes,

distances between genes are calculated in terms of differences in gene ranks. Use of differences in gene ranks provides relative gene distances, which can mitigate the effects of different gene densities (per unit physical DNA) among species on collinearity detection. Next, multiple chromosomal regions threaded by consecutive anchor loci are progressively aligned against "reference" genomes. Because there could be many intervening / non-anchor genes between consecutive anchor genes, especially for divergent genomes, the alignment of non-anchor genes is highly flexible and could clutter the view of results. Thus, in *MCScanX*, the alignment among non-anchor genes is discarded in the output and non-anchor genes (mismatches) are simply denoted by "||" in the multi-alignment of gene orders. As a result, the layout of multiple alignments is less affected by alignment parameters and anchor genes and duplication depths can be easily discerned in the resulting multiple alignments.

The results of *MCScanX* multiple alignments are presented in HTML format with variously colored features that can be displayed using a web browser. An example is shown in Figure 5.2. In a reference chromosome, both anchor and non-anchor genes are shown, while in aligned collinear blocks only anchor genes are shown. Along the reference chromosome, duplication depth (i.e. number of aligned collinear blocks) is shown at each locus to indicate how frequently chromosomal regions are duplicated, and tandem genes are highlighted in red. In principle, all aligned collinear blocks can be also references. Note that in certain cases, in a specific alignment (e.g. A-B-C), an anchor locus is lost in the reference chromosome (A) and in turn cannot be shown in aligned collinear blocks (B and C) due to the non-reciprocity of the employed algorithm. To study differential gene loss, the user is suggested to analyze the results using the gene or genome of interest as the reference (i.e. the alignments B-A-C and C-A-B can

show that the anchor locus exists between B and C but is lost in A) to ensure that complete chromosomal neighborhoods and matching segments are observed.

The third core program, named *duplicate gene classifier*, can classify the duplicate genes of a single species into WGD/segmental, tandem, proximal and dispersed duplicates. WGD/segmental duplicates are inferred by the anchor genes in collinear blocks. Tandem duplicates are defined as paralogs that are adjacent to each other on chromosomes, which are suggested to arise from illegitimate chromosomal recombination (Freeling 2009). Proximal duplicates are paralogs near each other, but interrupted by several other genes (e.g. separated by fewer than 20 genes, configurable). Proximal duplicates are inferred to result from localized transposon activities (Zhao et al. 1998), or ancient tandem arrays interrupted by more recent gene insertions. Dispersed duplicates are paralogs that are neither near each other on chromosomes, nor do they show conserved synteny (Ganko et al. 2007). Distant single-gene translocations mediated by transposons may explain the wide spread of dispersed duplicates (Freeling 2009), often via pack-MULEs (Jiang et al. 2004), helitrons (Yang and Bennetzen 2009), or CACTA elements (Paterson et al. 2009) in plant genomes, or through "retropositions" (Wang et al. 2006a). Inferences about the mechanism(s) responsible for duplication of genes may reveal unusual evolutionary characteristics for particular lineages. *Duplicate gene classifier,* incorporating the *MCScanX* procedure, takes in the same input files as *MCScanX*, and returns statistics of duplicate gene origins and a file showing the likely origin of each gene.

Once the outputs of the core programs are generated, various visualization and downstream analysis tools can be applied. To display synteny and collinearity, four types of plots can be generated: dual synteny plot (Figure 5.3A), circle plot (Figure 5.3B), dot plot (Figure 5.3C) and bar plot (Figure 5.3D) using the Java programs: *dot plotter, circle plotter, dual synteny*

plotter, and *bar plotter*, respectively. The ".collinearity" file generated by *MCScanX* can be annotated with non-synonymous (Ka) and synonymous (Ks) substitution rates using the Perl program *add ka and ks to collinearity.pl*. Gene families constructed based on collinear relationships (instead of BLAST hits) can be generated based on the ".collinearity" file using the Perl program *group collinear genes.* It may be interesting to see how frequently chromosomal regions are duplicated within or across species for understanding species-specific or shared evolutionary events, and the program *dissect multiple alignment* can compute the number of intra- and inter- species collinear blocks at each locus of reference genomes and show statistics on gene numbers at different duplication depths. To avoid high numbers of local collinear gene pairs generated by *MCScanX* due to tandem arrays, tandem matches are collapsed using a representative pair with the smallest BLASTP *E*-value during *MCScanX* execution. However, a tandem array at an ancestral locus may imply positional gene family expansion (Vergara and Chen 2010). Thus, a tool named *detect collinear tandem arrays* is provided for detection of collinear tandem arrays.

The *MCScanX* package provides a variety of tools for analyzing gene family evolution based on the synteny and collinearity identified by *MCScanX*. *Origin enrichment analysis* can detect potential enrichment of duplicate gene origins for gene families, based on the classification of whole-genome duplicate genes (the output of *duplicate gene classifier*). *Detect collinearity within gene families* outputs all collinear gene pairs among gene family members. *Family circle plotter* can detect all collinear gene pairs within a gene family and plot them using a genomic circle *Family tree plotter*, with a Newick-format tree (direct results from most phylogenetic software) and ".collinearity" and ".tandem" files (generated by *MCScanX*) as inputs, can graphically annotate a phylogenetic tree with collinear and tandem relationships.

Application examples

Estimation of the number of WGD events

MCScan version 0.8 was implemented to estimate the number of WGD events of *Arabidopsis, Carica, Populus* and *Vitis*, through analysis of the duplication depths of their collinear blocks using *Vitis* as the reference genome (Tang et al. 2008a; Tang et al. 2010). To facilitate this analysis using the output of *MCScanX*, the tool *dissect multiple alignment* is provided. When the user applies the *MCScanX* package, the BLASTP and GFF inputs should be restricted to a single genome for self-genome comparison or between two genomes for cross-genome comparison. Alternatively, a BLASTP of self-genome comparison and cross-genome comparison may be merged for both comparisons. However, self-genome comparison may not be as sensitive as cross-genome comparison due to the differential loss of functionally redundant genes, sometimes in a complementary fashion (Tang et al. 2008b). Although the determination of an exact number of WGD events may be heuristic, the output of *dissect multiple alignment* can give a reasonable estimate. Note that a duplication depth x indicates that there are x and x+1 aligned collinear blocks in the target genome using cross-genome and self-genome comparisons respectively. For example, *dissect multiple alignment* was applied to both self-genome and cross-genome comparisons between *Arabidopsis* and *Vitis*. Using *Arabidopsis* and *Vitis* as references, the maximum duplication depths of *Arabidopsis* collinear blocks are 7 (self-genome comparison, so the maximum number of aligned *Arabidopsis* collinear blocks is 8) and 11 (cross-genome comparison, so the maximum number of aligned *Arabidopsis* collinear blocks is 11) respectively, suggesting that the lineage experienced at least three WGD events to achieve this duplication depth, i.e. a triplication WGD event γ × two duplication WGD events α and β (Bowers et al.

2003; Tang et al. 2008a; Tang et al. 2008b). By applying *dissect multiple alignment* to self-genome comparison of *Vitis*, the maximum duplication depth of *Vitis* collinear blocks is 4. However, the gene numbers at levels 3 and 4 (297 and 6 respectively) are much smaller than at level 2 (6993). A whole-genome triplication (WGT) plus small-scale chromosomal duplications is the simplest explanation for this duplication pattern (Tang et al. 2008a; Tang et al. 2008b). Note that analysis of duplication depths of collinear blocks can generate good estimates on relatively recent WGD events. Very ancient WGD events often do not result in discernable collinear blocks in extant species due to extensive chromosome rearrangement, loss or gain of chromosomal segments, loss or transposition of duplicate genes, horizontal gene transfers, etc. A recent study, through analyzing the phylogenetic trees of cross-species gene families, reported two ancestral WGD events for seed plants and angiosperms respectively (Jiao et al. 2011).

Detection of collinear orthologs

Detection of collinear orthologs is important for understanding gene evolution. The comparison between collinear orthologs and all orthologs can reveal how gene orders are conserved (or inversely, how frequently chromosomes are rearranged) between species. Limited only by the state of a genome's annotation and the assumption that sufficient sequence similarity is present for detection, a complete set of orthologs for a set of species can be generated by third-party software such as OrthoMCL (45). We implemented OrthoMCL to find ortholog pairs among *Arabidopsis*, *Populus*, *Vitis*, *Glycine*, *Oryza, Brachypodium*, *Sorghum* and *Zea.* The ortholog pairs identified by OrthoMCL were regarded as the whole set of orthologs, and were then used as the input of *MCScanX_h*. Besides standard *MCScanX* output, *MCScanX_h* generated statistics on the numbers of collinear ortholog pairs and all ortholog pairs, and percentages of collinear ortholog pairs between any two of the selected angiosperm genomes

(Table 5.1). As expected, gene order is better conserved within monocots and within eudicots than between monocots and eudicots. Within eudicots, *Vitis* shows the highest level of collinearity with the other 3 species, suggesting that *Vitis* most closely resemble the gene order of the eudicot ancestral genome, due in part to the lack of recent WGDs (Jaillon et al. 2007).

Differences in duplicate gene origins among angiosperms

Using self-genome BLASTP outputs and the tool *duplicate gene classifier*, we classified the origins of duplicate genes for *Arabidopsis*, *Populus*, *Vitis*, *Glycine*, *Oryza, Brachypodium*, *Sorghum* and *Zea* respectively. The results are shown in Table 5.2. The collinear blocks in the self-genome comparisons result from segmental or whole-genome duplications. Most collinear blocks within these flowering plant genomes were derived from WGDs because of their high coverage throughout the genome as well as supporting Ks evidence (Tang et al. 2008b).

WGDs have had different impacts on the gene repertoires of the investigated taxa. Strikingly, ~ 76.0% of *Glycine* genes were duplicated and retained from WGD events, versus only 14.5% of *Oryza* genes. The proportions of genes involved in WGD events may reflect the relative timing of the most recent WGD event, as well as the level of gene retention following the WGD. For example, *Vitis*, with only 15.0% of genes created by WGD (actually WGT), was inferred to have undergone the γ WGT event, which likely predated the divergence of most eudicots more than 100 million years ago (Tang et al. 2008a; Tang et al. 2008b). Other eudicot lineages have experienced lineage-specific WGDs in addition to the shared γ event. 27.0% of *Arabidopsis* appear to have been created through WGD, having experienced α and β WGD events since its divergence from other members of the Brassicales clade (Bowers et al. 2003; Tang et al. 2008a). *Populus*, with 51.6% of genes created by WGD, was inferred to have undergone an additional WGD event in the Salicoid lineage (Tang et al. 2008a). *Glycine*, with the highest proportion of

WGD genes, was reported to have experienced two additional WGD events, with the most recent occurring 13 million years ago (Schmutz et al. 2010). A total of 29.2% of *Zea* genes were created through WGD, which experienced a lineage specific WGD after its divergence from *Sorghum* (15.2% genes created by WGD) (Wei et al. 2007; Salse et al. 2008). Although tandem genes are volatile after gene duplication, those retained may indicate functional significance. We find that tandem genes account for about 1~3% of genes in each genome, smaller than ~10% reported by Rizzon et al. (Rizzon et al. 2006). This difference is due to the algorithm of *duplicate gene classifier*, which treats the tandem duplicates located at ancestral loci as WGD duplicates. Proximal duplicates account for larger proportions of genes in the genomes with fewer WGD duplicates, e.g. there are 5.4% of *Oryza* genes and 6.7% of *Vitis* genes created by proximal duplications, while in other genomes, the numbers of proximal duplicates are comparable to those of tandem duplicates.

Detection of collinear tandem arrays

In the *MCScanX* package, tandem arrays are defined as clusters of consecutive tandem duplicates. Via *detect collinear tandem arrays*, tandem arrays are first determined according to successive gene ranks in all chromosomes. Collinear gene pairs are then searched against these tandem arrays. If any gene of a collinear pair is located within a tandem array, the gene is replaced by the tandem array and then reported. If a tandem array is located at an anchor locus of a collinear block, it is termed a collinear tandem array. Collinear tandem arrays can indicate positional gene family expansions (Vergara and Chen 2010), which could be important for forming large gene families, or adopted as an alternative path to increasing gene copy number in the genomes that experienced fewer WGD events. For example, we applied the tool *detect collinear tandem arrays* to a comparison of the *Arabidopsis* and *Vitis* genomes. A total of 1,160

pairs of collinear tandem arrays were detected between *Arabidopsis* and *Vitis*, of which only 68 (5.9%) pairs have equal numbers of tandem duplicates in each species, while 54.3% of pairs have more tandem duplicates in *Vitis* than *Arabidopsis*. In conjunction with the finding above that *Vitis* has more proximal duplicates than other species, we suggest that tandem and proximal duplications contribute relatively more to the expansion of the *Vitis* genome than to other eudicots that experienced more WGDs in their evolutionary histories.

Analysis of gene family evolution

While *MCScanX* can detect synteny and collinearity using whole-genome homology and gene positional information, it is also of interest to analyze collinearity within a gene family, toward clarifying gene family evolution (Sampedro et al. 2005). We used the *Arabidopsis* MADS-box gene family as an example to illustrate the usefulness of *MCScanX* for analyzing the history of gene family expansion. Using the tool *detect collinearity within gene families*, we detected 14 collinear gene pairs from the members of the MADS box gene family. The inferred collinear relationships of the MADS box gene family members can be displayed and placed within the context of whole-genome collinearity using a genomic circle generated by *family circle plotter* (Figure 5.4). Next, a phylogenetic tree was constructed for the MADS box gene family using PhyML package (Guindon et al. 2010). The Newick tree was then used as the input of *family tree plotter*. A plot that showed the phylogenetic tree, collinear and tandem relationships for the MADS box gene family was generated (Figure 5.5). The overlay of positional history over the gene clades reveals interesting characteristics of the MADS-box gene family. We note that the clade with many collinear relationships (WGD or segmentally duplicated) appears to be the MIKC^{c}-type (Becker and Theissen 2003). In contrast, the remaining

clades of MADS-box genes appear to favor dispersed duplications (Freeling et al. 2008; Freeling 2009).

The tool *origin enrichment analysis*, which is able to detect potential enrichments of duplicate gene origins, was applied to 126 published *Arabidopsis* gene families of 10 or more genes, available at TAIR (http://www.arabidopsis.org/). We found that 46 (36.5%) gene families were enriched for at least one of the four types of origins at α=0.05. For example, disease resistance gene homologs and the cytochrome P450 gene family are enriched for dispersed and proximal duplicates, while the cytoplasmic ribosomal protein gene family and C2H2 zinc-finger proteins are enriched for WGD duplicates, as previously noted (Freeling et al. 2008).

Discussion

Existing tools for synteny and collinearity detection mainly include i-ADHoRe (Vandepoele et al. 2002), LineUp (Hampson et al. 2003), MCMuSeC (Ling et al. 2009), OrthoCluster (Vergara and Chen 2009), DiagHunter (Cannon et al. 2003), DAGChainer (Haas et al. 2004), ColinearScan (Wang et al. 2006b), MCScan (Tang et al. 2008b), SyMAP (Soderlund et al. 2006), FISH (Calabrese et al. 2003), Cyntenator (Rodelsperger and Dieterich 2010), MicroSyn (Cai et al. 2011) and Cinteny (Sinha and Meller 2007), of which i-ADHoRe and SyMAP are currently on their 3 (Fostier et al. 2011) and 3.4 (Soderlund et al. 2011) versions respectively. We summarized the functions of synteny and collinearity detection tools regarding five elements: visualization, operation on multiple (>2) genomes, multi-alignments, evolutionary analyses of synteny and collinearity (e.g. estimating WGD events, gene order conservation and duplicate gene origins, constructing collinear gene groups/families, etc), and analyses of gene families. Functional comparison of different synteny and collinearity detection tools is shown in Table 5.3. Seven tools output synteny or collinearity information as plain texts, while the other

tools provide visualization options, though types and numbers of plots vary among different tools. As for the data scale, most tools published in the past four years can operate on multiple genomes. Four tools can perform multi-alignments of collinear blocks. MicroSyn is focused on collinearity analysis within gene families. i-ADHoRe 3 has provided several post-processing programs for dissecting multi-alignments of collinear blocks, in addition to detecting and visualizing synteny and collinearity. Among these synteny and collinearity detection tools, 11 tools cover fewer than two functions, and MicroSyn and i-ADHoRe 3 cover three functions. *MCScanX*, with all five functions, can perform more biological analyses than any other synteny or collinearity detection tool.

MCScanX is unique in providing multiple programs for evolutionary analysis of synteny and collinearity, which are a necessary step towards biological discovery. Further, *MCScanX* has connected collinearity analyses between whole-genome and gene family scales. To our knowledge, the following biological analyses implemented in *MCScanX* are not yet available in other synteny and collinearity detection tools: constructing gene families using collinearity information, inferring gene duplication modes and enrichments, detecting collinear tandem arrays, performing statistical analyses of duplication depths and collinear orthologs, and annotating phylogenetic trees with collinearity and tandems.

For synteny and collinearity detection tools, effective identification of collinear gene pairs is the basis for collinear block construction and downstream analyses. It is informative to perform a quantitative evaluation of *MCScanX* on the identification of collinear gene pairs. Two widely implemented tools, MCScan and i-ADHoRe 3 were chosen as competitors. Because a benchmark for assessing synteny and collinearity tools has not been established (Fostier et al. 2011), we compared their performances by applying them to the *Arabidopsis thaliana* genome.

Note that a higher number of detected collinear gene pairs does not simply indicate better performance, as true and false positives must be simultaneously considered and well balanced (Cannon et al. 2003). A total of 5,794 collinear gene pairs (i.e. WGD duplicate gene pairs) in the *Arabidopsis* genome including 3,822 α, 1,451 β and 521 γ pairs profiled using an integrated phylogenomic approach in the study from Bowers et al. (Bowers et al. 2003), were regarded as the whole set of collinear gene pairs. The performances of MCScan, *MCScanX* and i-ADHoRe 3 were evaluated by power (i.e. sensitivity), defined as the ratio between numbers of true positives and all collinear gene pairs; and precision, defined as the ratio between numbers of true positives and all positives (i.e. true positives + false positives). When MCScan and *MCScanX* were compared, the same parameters were used. Based on the default parameters of *MCScanX* (match size=5, max gaps=25), MCScan and *MCScanX* identified 4,134 and 4,225 collinear gene pairs, of which 3,375 and 3,407 were true positives respectively. Power was 0.58 and 0.59, and precision was 0.82 and 0.81 for MCScan and *MCScanX*, respectively. The above statistics suggest that MCScan and *MCScanX* are generally comparable in detecting collinear gene pairs, while *MCScanX* has a slightly higher power and a slightly lower precision. Based on its default parameters, i-ADHoRe 3 identified 6,233 non-overlapping collinear gene pairs, of which 3,459 were true positives. Its power and precision was 0.60 and 0.55. However, direct comparison between *MCScanX* and i-ADHoRe 3 using their respective default parameters was not reasonable because i-ADHoRe 3 output many more positives. To this end, we executed MCScan and *MCScanX* using a more relaxed set of parameters (match size=3, max gaps=50), which output 5,554 and 6,110 positives respectively. Based on the new parameters, power was 0.65 and 0.67, and precision was 0.68 and 0.64 for MCScan and *MCScanX* respectively. The new statistics suggest that in terms of identification of collinear gene pairs, MCScan and *MCScanX*

each perform better than i-ADHoRe 3 and remain comparable to one another, with MCScan having higher precision and *MCScanX* having higher power. The small difference between MCScan and *MCScanX* is because in order to make *MCScanX* more easily and efficiently implemented, preprocessing of BLASTP input was pipelined into the execution of the main programs and the dependency of MCL was dropped. In MCScan, cross-family BLASTP hits are removed based on MCL output, while in *MCScanX*, all non-self BLASTP hits are considered, leading to an enlarged pool of BLASTP hits. MCL may generate 5-20% incorrect families and its performance is affected by inflation value (a parameter of the MCL algorithm used to control the granularity/tightness of protein clusters) (Enright et al. 2002). So the cross-family BLASTP hits based on MCL gene families indeed contain some collinear gene pairs, though the proportion of collinear gene pairs is smaller in cross-family BLASTP hits than in within-family BLASTP hits. This results in marginally higher power and lower precision for *MCScanX* than MCScan, though their performances on identifying collinear gene pairs are very similar. Since MCScan was successfully applied to the distantly related apicomplexans (Debarry and Kissinger 2011), we believe that *MCScanX* is also applicable over a wide range of organisms besides angiosperms.

In conclusion, *MCScanX* is a toolkit that implements an adjusted MCScan algorithm for detection of synteny and collinearity and incorporates 14 computer programs for visualizing and analyzing identified synteny and collinearity. The usefulness of the *MCScanX* toolkit has been demonstrated through a series of real data applications and comparison with other synteny and collinearity detection tools. *MCScanX* is freely available at http://chibba.pgml.uga.edu/mcscan2/.

References

Abrouk M, Murat F, Pont C, Messing J, Jackson S, Faraut T, Tannier E, Plomion C, Cooke R, Feuillet C *et al.* 2010. Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci* **15**(9): 479-487.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Becker A, Theissen G. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol Phylogenet Evol* **29**(3): 464-489.

Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res* **14**(4): 507-516.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930): 433-438.

Cai B, Yang X, Tuskan GA, Cheng ZM. 2011. MicroSyn: a user friendly tool for detection of microsynteny in a gene family. *BMC Bioinformatics* **12**: 79.

Calabrese PP, Chakravarty S, Vision TJ. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19 Suppl 1**: i74-80.

Cannon SB, Kozik A, Chan B, Michelmore R, Young ND. 2003. DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol* **4**(10): R68.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol* **7**(2): R13.

Causier B, Castillo R, Xue YB, Schwarz-Sommer Z, Davies B. 2010. Tracing the evolution of the floral homeotic B- and C-function genes through genome synteny. *Mol Biol Evol* **27**(11): 2651-2664.

Charles M, Tang HB, Belcram H, Paterson A, Gornicki P, Chalhoub B. 2009. Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of *Pooideae* and *Ehrhartoideae*, after their divergence from *Panicoideae*. *Mol Biol Evol* **26**(7): 1651-1661.

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet* **21**(12): 673-682.

Debarry JD, Kissinger JC. 2011. Jumbled genomes: missing Apicomplexan synteny. *Mol Biol Evol* **28** (10): 2855-2871.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi SD *et al.* 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**(5668): 304-307.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E *et al.* 2004. Genome evolution in yeasts. *Nature* **430**(6995): 35-44.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**(7): 1575-1584.

Fostier J, Proost S, Dhoedt B, Saeys Y, Demeester P, Van de Peer Y, Vandepoele K. 2011. A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* **27**(6): 749-756.

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

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res* **18**(12): 1924-1937.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol* **24**(10): 2298-2309.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**(3): 307-321.

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

Hampson S, McLysaght A, Gaut B, Baldi P. 2003. LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res* **13**(5): 999-1010.

Higgins JA, Bailey PC, Laurie DA. 2010. Comparative genomics of flowering time pathways using *Brachypodium distachyon* as a model for the temperate grasses. *PLoS One* **5**(4): e10065.

Hyun TK, Kim JS, Kwon SY, Kim SH. 2010. Comparative genomic analysis of mitogen activated protein kinase gene family in grapevine. *Genes Genom* **32**(3): 275-281.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**(7161): 463-467.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**(7008): 569-573.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**(7345): 97-100.

Jun J, Mandoiu II, Nelson CE. 2009. Identification of mammalian orthologs using local synteny. *BMC Genomics* **10**: 630.

Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann Bot* **95**(1): 177-190.

Knoller AS, Blakeslee JJ, Richards EL, Peer WA, Murphy AS. 2010. Brachytic2/ZmABCB1 functions in IAA export from intercalary meristems. *J Exp Bot* **61**(13): 3689-3696.

Kopriva S, Mugford SG, Matthewman C, Koprivova A. 2009. Plant sulfate assimilation genes: redundancy versus specialization. *Plant Cell Rep* **28**(12): 1769-1780.

Li C, Zhang YM. 2011. Molecular evolution of glycinin and beta-conglycinin gene families in soybean (*Glycine max* L. Merr.). *Heredity* **106**(4): 633-641.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**(9): 2178-2189.

Li W, Liu B, Yu L, Feng D, Wang H, Wang J. 2009. Phylogenetic analysis, structural evolution and functional divergence of the 12-oxo-phytodienoate acid reductase gene family in plants. *BMC Evol Biol* **9**: 90.

Lin L, Pierce GJ, Bowers JE, Estill JC, Compton RO, Rainville LK, Kim C, Lemke C, Rong J, Tang H *et al.* 2010. A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics* **11**: 395.

Lin L, Tang H, Compton RO, Lemke C, Rainville LK, Wang X, Rong J, Rana MK, Paterson AH. 2011. Comparative analysis of *Gossypium* and *Vitis* genomes indicates genome duplication specific to the *Gossypium* lineage. *Genomics* **97**(5): 313-320.

Ling X, He X, Xin D. 2009. Detecting gene clusters under evolutionary constraint in a large number of genomes. *Bioinformatics* **25**(5): 571-577.

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang HB, Wang XY, Bowers J, Paterson A, Lisch D *et al.* 2008. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: coge with rosids. *Plant Physiol* **148**(4): 1772-1781.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**(9): 1254-1265.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**(5): 418-426.

Okazaki Y, Shimojima M, Sawada Y, Toyooka K, Narisawa T, Mochida K, Tanaka H, Matsuda F, Hirai A, Hirai MY *et al.* 2009. A chloroplastic UDP-glucose pyrophosphorylase from *Arabidopsis* is the committed enzyme for the first step of sulfolipid biosynthesis. *Plant Cell* **21**(3): 892-909.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401-437.

Palmieri F, Pierri CL, De Grassi A, Nunes-Nesi A, Fernie AR. 2011. Evolution, structure and function of mitochondrial carriers: a review with new insights. *Plant J* **66**(1): 161-181.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al.* 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**(7229): 551-556.

Ranz JM, Casals F, Ruiz A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res* **11**(2): 230-239.

Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput Biol* **2**(9): e115.

Rodelsperger C, Dieterich C. 2010. CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS One* **5**(1): e8861.

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C. 2009. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci U S A* **106**(35): 14908-14913.

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**(1): 11-24.

Sampedro J, Lee Y, Carey RE, dePamphilis C, Cosgrove DJ. 2005. Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. *Plant J* **44**(3): 409-419.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**(7278): 178-183.

Sinha AU, Meller J. 2007. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics* **8**: 82.

Soderlund C, Bomhoff M, Nelson WM. 2011. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* **39**(10): e68.

Soderlund C, Nelson W, Shoemaker A, Paterson A. 2006. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res* **16**(9): 1159-1168.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008a. Synteny and collinearity in plant genomes. *Science* **320**(5875): 486-488.

Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci U S A* **107**(1): 472-477.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**(12): 1944-1954.

Vandepoele K, Saeys Y, Simillion C, Raes J, Van De Peer Y. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res* **12**(11): 1792-1801.

Vergara IA, Chen N. 2009. Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Curr Protoc Bioinformatics* **Chapter 6**: Unit 6 10 16 10 11-18.

Vergara IA, Chen N. 2010. Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics* **11**: 516.

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S *et al.* 2006a. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**(8): 1791-1802.

Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. 2009a. Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol* **10**(6): R68.

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. 2006b. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics* **7**: 447.

Wang X, Tang H, Paterson AH. 2011. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell* **23**(1): 27-37.

Wang X, Tang H, Bowers JE, Paterson AH. 2009b. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res* **19**(6): 1026-1032.

Watanabe M, Mochida K, Kato T, Tabata S, Yoshimoto N, Noji M, Saito K. 2008. Comparative genomics and reverse genetics analysis reveal indispensable functions of the serine acetyltransferase gene family in *Arabidopsis*. *Plant Cell* **20**(9): 2484-2496.

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S *et al.* 2007. Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* **3**(7): e123.

Yang L, Bennetzen JL. 2009. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc Natl Acad Sci U S A* **106**(47): 19922-19927.

Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**(5): 479-492.

Table 5.1. Numbers of collinear ortholog pairs and total ortholog pairs and percentage of collinear ortholog pairs in selected angiosperm genomes.

Species1	# of collinear ortholog pairs, # of total ortholog pairs and percentage of collinear ortholog pairs						
	Pt	Gm	Vv	Os	Bd	Sb	Zm
At	14278, 46944, 30.4%	17498, 58038, 30.1%	7378, 24086, 30.6%	319, 24992, 1.3%	202, 22719, 0.9%	350, 24120, 1.5%	142, 24689, 0.6%
Pt	-	34545, 92901, 37.2%	15734, 38727, 40.6%	2121, 37575, 5.6%	1632, 32790, 5.0%	1523, 36059, 4.2%	687, 35596, 1.9%
Gm	-	-	18310, 47652, 38.4%	1437, 46916, 3.1%	1308, 43130, 3.0%	1263, 46631, 2.7%	501, 47326, 1.1%
Vv	-	-	-	1315, 19678, 6.7%	981, 18080, 5.4%	1194, 19137, 6.2%	293, 19501, 1.5%
Os	-	-	-	-	15492, 34413, 45.0%	15664, 39695, 39.5%	14112, 35206, 40.1%
Bd	-	-	-	-	-	14070, 32701, 43.0%	13111, 30841, 42.5%
Sb	-	-	-	-	-	-	18084, 36826, 49.1%

1 Abbreviations: At: *Arabidopsis thaliana*; Pt: *Populus trichocarpa*; Gm: *Glycine max*; Vv: *Vitis vinifera*; Os: *Oryza sativa*; Bd: *Brachypodium distachyon*; Sb: *Sorghum bicolor*; Zm: *Zea mays*.

Table 5.2. Numbers of genes from different origins as classified by duplicate gene classifier in eight angiosperm genomes

Species	# of genes	# of genes from different origins (percentage)				
		Singletons	WGD	Tandem	Proximal	Dispersed
Arabidopsis	27105	5272 (19.5)	7321 (27.0)	769 (2.8)	892 (3.3)	12851 (47.4)
Populus	40650	5014 (12.3)	20989 (51.6)	713 (1.8)	999 (2.5)	12935 (31.8)
Glycine	46360	1459 (3.1)	35233 (76.0)	582 (1.3)	670 (1.4)	8416 (18.2)
Vitis	23647	6275 (26.5)	3539 (15.0)	688 (2.9)	1590 (6.7)	11555 (48.9)
Oryza	40634	12720 (31.3)	5896 (14.5)	960 (2.4)	2184 (5.4)	18874 (46.4)
Brachypodium	25524	4842 (19.0)	4575 (17.9)	697 (2.7)	827 (3.2)	14583 (57.1)
Sorghum	34564	5839 (16.9)	5260 (15.2)	895 (2.6)	1283 (3.7)	21287 (61.6)
Zea	39365	8212 (20.9)	11506 (29.2)	774 (2.0)	1175 (3.0)	17698 (45.0)

Table 5.3. Functional comparison of different synteny and collinearity detection tools

Tool	Year published	Visualization	Multiple genomes	Multi-alignments	Evolutionary analyses of synteny and collinearity	Analyses of gene families
i-ADHoRe 3	2011	+	+	+	-	-
LineUp	2003	-	-	-	-	-
MCMuSeC	2009	-	+	-	-	-
OrthoCluster	2009	-	+	-	-	-
DiagHunter	2003	+	-	-	-	-
DAGChainer	2004	+	-	-	-	-
ColinearScan	2006	-	-	-	-	-
MCScan	2008	-	+	+	-	-
SyMAP 3.4	2011	+	+	-	-	-
FISH	2003	-	-	-	-	-
Cyntenator	2010	-	+	+	-	-
MicroSyn	2011	+	+	-	-	+
Cinteny	2007	+	+	-	-	-
MCScanX		+	+	+	+	+

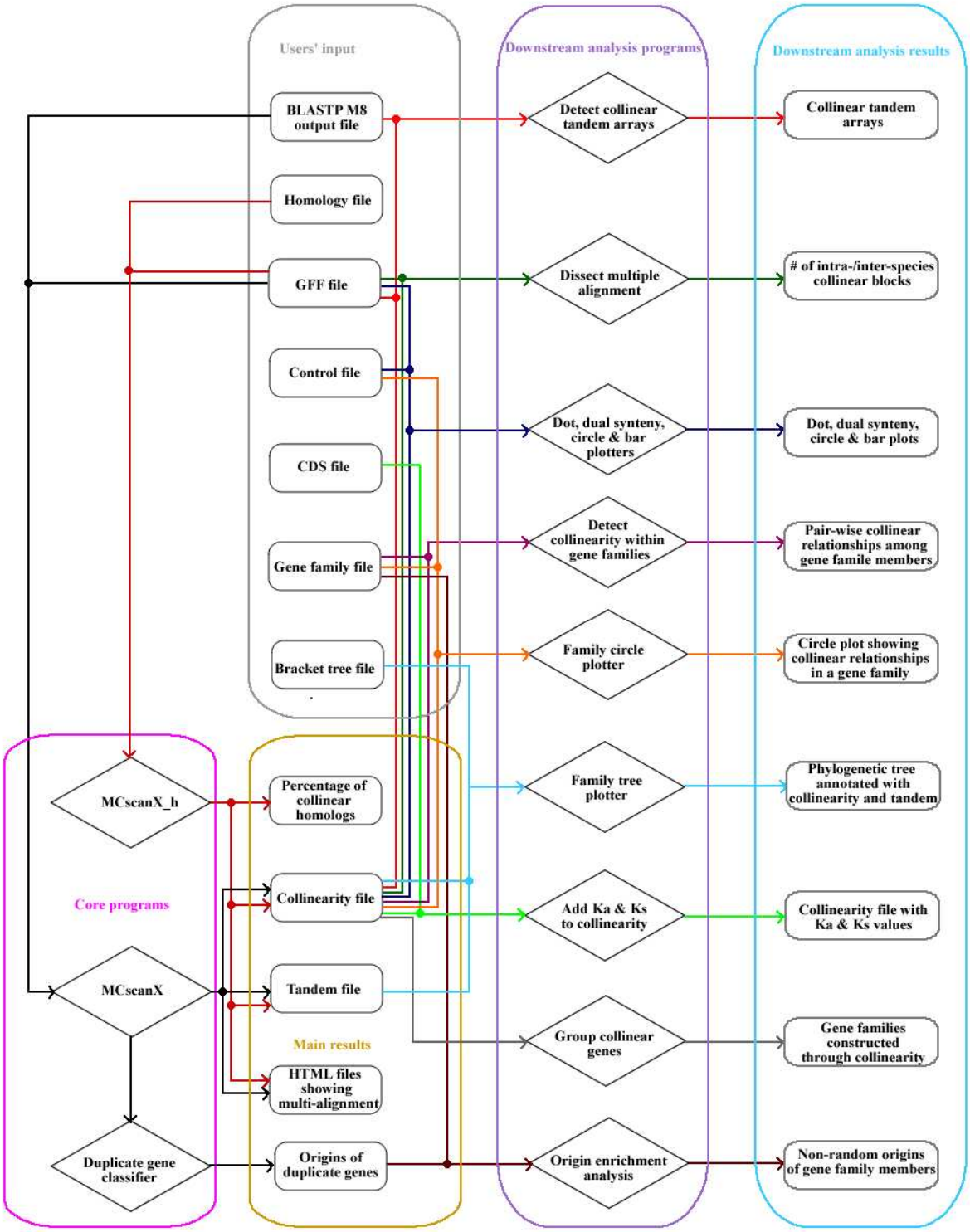

Figure 5.1. The structure of the *MCScanX* package illustrating major components and their dependencies.

Duplication depth	Reference chromosome	Collinear blocks				
2	AT1G01010	GSVIVT01019702001	GSVIVT01027477001			
2	AT1G01020	||	GSVIVT01027464001			
4	AT1G01030	GSVIVT01019699001	GSVIVT01027463001	AT3G61970	AT2G46870	
4	AT1G01040	||	GSVIVT01027462001	||	||	
4	AT1G01050	GSVIVT01019697001	GSVIVT01027459001	||	AT2G46860	
4	AT1G01060	||	GSVIVT01027456001	||	AT2G46830	
4	AT1G01070	||	||	||	||	
4	AT1G01073	||	||	||	||	
4	AT1G01080	||	GSVIVT01027441001	||	||	
4	AT1G01090	GSVIVT01019683001	GSVIVT01027439001	||	||	
5	AT1G01100	GSVIVT01019679001	GSVIVT01027436001	||	||	AT4G00810
5	AT1G01110	GSVIVT01019668001	GSVIVT01027429001	||	||	AT4G00820
5	AT1G01115	||	||	||	||	||
5	AT1G01120	||	GSVIVT01027424001	||	AT2G46720	||
5	AT1G01130	||	GSVIVT01027421001	||	||	||
5	AT1G01140	||	||	||	||	||
5	AT1G01150	||	||	||	||	||
5	AT1G01160	||	GSVIVT01027418001	||	||	AT4G00850
5	AT1G01170	||	||	||	||	AT4G00860
5	AT1G01180	||	||	||	||	||
5	AT1G01190	GSVIVT01019653001	GSVIVT01027404001	AT3G61880	AT2G46660	||
4	AT1G01200		GSVIVT01027396001	||	||	||
4	AT1G01210		GSVIVT01027190001	||	||	||
4	AT1G01220		GSVIVT01027188001	||	||	||
4	AT1G01225		GSVIVT01027187001	||	||	AT4G00905

Figure 5.2. Sample HTML output displaying multiple alignments of collinear blocks by *MCScanX*. The first and second columns show duplication depth and gene symbol at each locus of reference chromosome, where tandems are marked in red. The remaining columns show aligned collinear blocks, where only the symbols of anchor genes are shown.

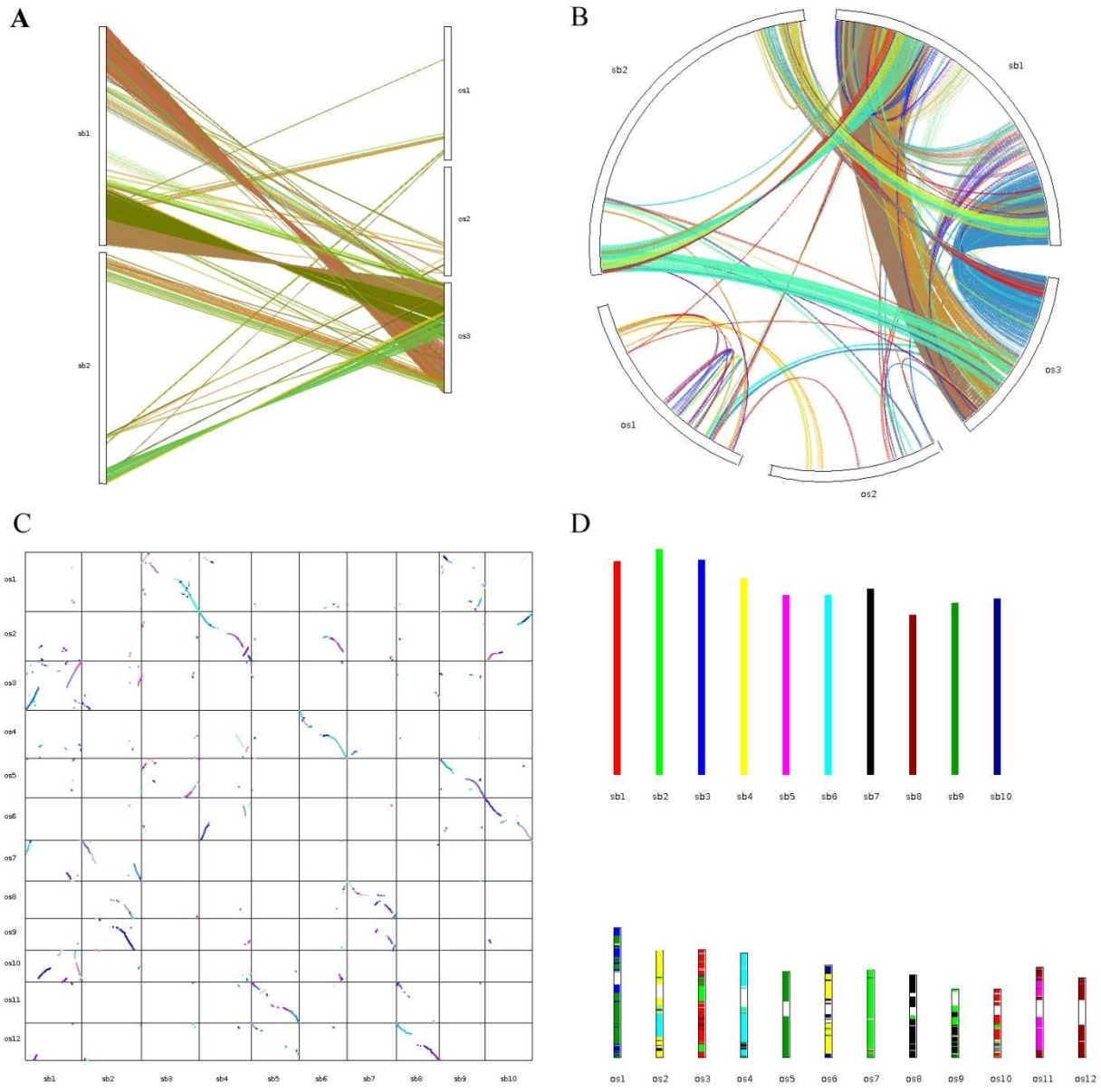

Figure 5.3. Different types of plots showing patterns of synteny and collinearity: A) dual synteny plot, B) circle plot, C) dot plot and D) bar plot, generated by *dual synteny plotter*, *circle plotter*, *dot plotter* and *bar plotter* respectively. Chromosomes are labeled in the format "species abbreviation"+ "chromosome ID". Abbreviations: os: *Oryza sativa*; sb: *Sorghum bicolor*.

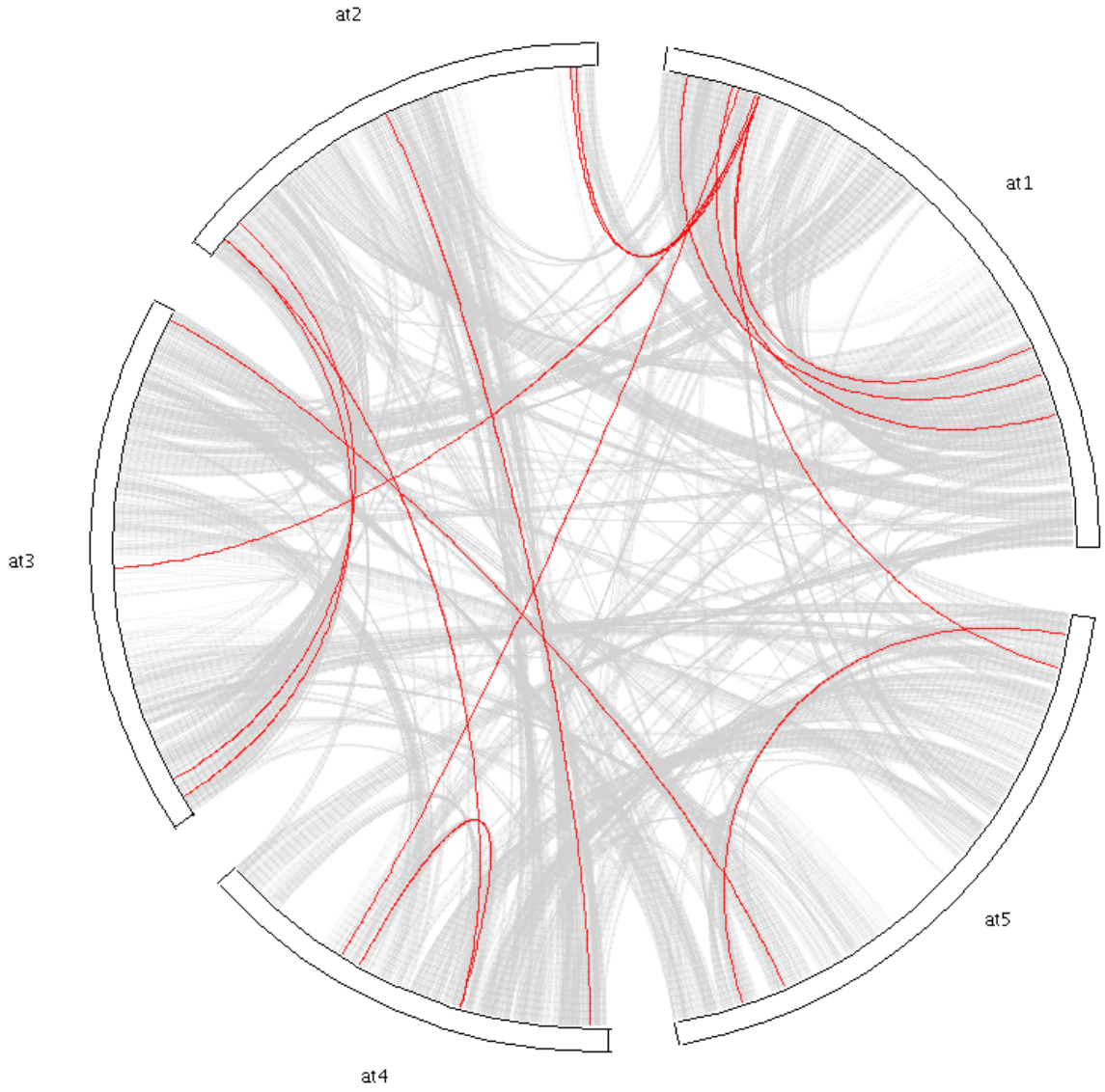

Figure 5.4. Circle plot showing collinearity in the MADS box gene family over the gray background of collinearity in *Arabidopsis* (the collinear blocks in *Arabidopsis*). The circle plot can be generated by *family tree plotter*. Chromosomes are labeled in the format "species abbreviation"+ "chromosome ID". Abbreviations: at: *Arabidopsis thaliana*.

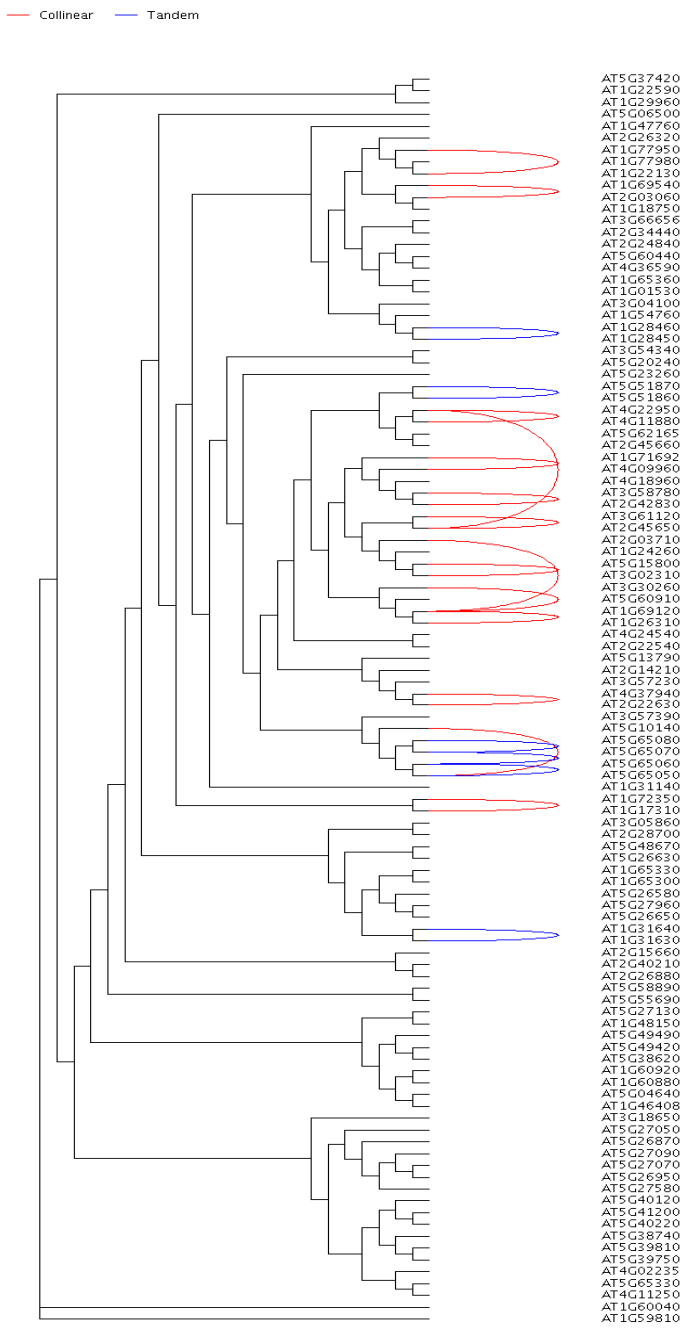

Figure 5.5. Phylogenetic tree of the MADS box gene family in *Arabidopsis* annotated with collinear and tandem relationships. Curves connecting pairs of gene names suggest either the collinear relationship (red) or tandem relationship (blue). This annotated tree is output from *family tree plotter*.

CHAPTER 6

MODES OF GENE DUPLICATION CONTRIBUTE DIFFERENTLY TO GENETIC NOVELTY AND REDUNDANCY, BUT SHOW PARALLELS ACROSS DIVERGENT ANGIOSPERMS[1]

[1]Yupeng Wang, Xiyin Wang, Haibao Tang, Xu Tan, Stephen P. Ficklin, F. Alex Feltus and Andrew H. Paterson. Accepted by *PLoS One*.

Reprinted here with permission from the publisher.

Abstract

Both single-gene and whole-genome duplications (WGD) have recurred in angiosperm evolution. However, the evolutionary effects of different modes of gene duplication, especially regarding their contributions to genetic novelty or redundancy, have been inadequately explored.

In *Arabidopsis thaliana* and *Oryza sativa* (rice), species that deeply sample botanical diversity and for which expression data are available from a wide range of tissues and physiological conditions, we have compared expression divergence between genes duplicated by six different mechanisms (whole-genome, tandem, proximal, DNA-based transposed, retrotransposed and dispersed duplication), and between positional orthologs. Both neo-functionalization and genetic redundancy appear to contribute to retention of duplicate genes. Genes resulting from WGD and tandem duplications diverge the slowest in both coding sequences and gene expression, and contribute most to genetic redundancy, while other duplication modes contribute more to evolutionary novelty. WGD duplicates may more frequently be retained due to dosage amplification, while inferred transposon-mediated gene duplications tend to reduce gene expression levels. The extent of expression divergence between duplicates is discernibly related to duplication mode, different WGD events, amino acid divergence, and putatively neutral divergence (time), but the contribution of each factor is heterogeneous among duplication modes. Gene loss may retard inter-species expression divergence. Members of different gene families may have non-random patterns of origin that are similar in *Arabidopsis* and rice, suggesting the action of pan-taxon principles of molecular evolution.

Gene duplication modes differ in contribution to genetic novelty and redundancy, but show some parallels in taxa separated by hundreds of millions of years of evolution.

Introduction

Whole-genome duplications (WGDs) have occurred in the lineages of plants (Paterson et al. 2010), animals (Jaillon et al. 2004; Aury et al. 2006) and fungi (Wolfe and Shields 1997; Kellis et al. 2004), with possible consequences including evolution of novel or modified gene functions (Ohno 1970; Lynch and Conery 2000; Zhang and Cohn 2008; Kassahn et al. 2009), and/or provision of "buffer capacity" (Chapman et al. 2006; VanderSluis et al. 2010) or genetic redundancy that increases genetic robustness (Gu et al. 2003; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; DeLuna et al. 2010). Genome duplication may also increase opportunities for nonreciprocal recombination (Wang et al. 2007; Wang et al. 2009; Wang et al. 2011), permitting or causing duplicated genes to evolve in concert for a period of time. Rapid DNA loss and restructuring of low-copy DNA (Song et al. 1995; Ozkan et al. 2001; Shaked et al. 2001; Kashkush et al. 2002), retrotransposon activation (O'Neill et al. 1998; Kashkush et al. 2003; Paterson et al. 2009) and epigenetic changes (Chen and Pikaard 1997; Comai et al. 2000; Lee and Chen 2001; Rodin and Riggs 2003; Adams and Wendel 2005; Rapp and Wendel 2005) following WGD may further provide materials for evolutionary change.

Genes may be duplicated by several mechanisms in addition to WGDs, which have been collectively referred to as small-scale duplications (Maere et al. 2005) or single-gene duplications (Cusack and Wolfe 2007; Freeling 2009). Tandem duplicates are consecutive in the genome while proximal duplicates are near one another but separated by a few genes. These two gene duplication modes are presumed to arise through unequal crossing over (Freeling 2009) or localized transposon activities (Zhao et al. 1998). Dispersed duplicates are neither adjacent to each other in the genome nor within homeologous chromosome segments (Ganko et al. 2007). Distant single-gene transposition may explain the widespread existence of dispersed duplicates

within and among genomes (Freeling 2009). Distant single-gene transposition duplication (referred to as distantly-transposed duplication) may occur by DNA-based or RNA-based mechanisms (Cusack and Wolfe 2007). DNA transposons such as packmules (rice) (Jiang et al. 2004), helitrons (maize) (Brunner et al. 2005), and CACTA elements (sorghum) (Paterson et al. 2009) may relocate duplicated genes or gene segments to new chromosomal positions (referred to as DNA-based transposed duplication). RNA-based transposed duplication, often referred to as retrotransposition, typically creates a single-exon retrocopy from a multi-exon parental gene, by reverse transcription of a spliced messenger RNA. It is presumed that the retrocopy duplicates only the transcribed sequence of the parental gene, detached from the parental promoter. The new retrogene is often deposited in a novel chromosomal environment with new (i.e. non-ancestral) neighboring genes and, having lost its native promoter, is only likely to survive as a functional gene if a new promoter is acquired (Brosius 1991; Kaessmann et al. 2009).

Classical population genetic theory suggests that a likely consequence of gene duplication is reversion to single copy (singleton), unless at least one gene copy evolves new function (Ohno 1970). More recently, the subfunctionalization model, which proposes that duplicated gene copies might both be retained if they partition the functions of the ancestral gene between them, has described an important modification of the classical model (Force et al. 1999; Lynch and Conery 2000). Some studies also show evidence to support the value of genetic redundancy *per se* (Hughes and Hughes 1993; Hughes 1994; Gu et al. 2003; Chapman et al. 2006; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; DeLuna et al. 2010) or dosage balance (Papp et al. 2003; Veitia 2003; Maere et al. 2005; Paterson et al. 2010).

The angiosperms (flowering plants) are an outstanding model in which to elucidate the consequences of gene duplication. All angiosperms are now thought to be paleopolyploids

(Bowers et al. 2003), many of which underwent multiple WGDs (Paterson et al. 2004; Tang et al. 2008a). Traces of past WGDs can often be detected from pairwise syntenic alignments through software such as ColinearScan (Wang et al. 2006) and multiple alignments using MCScan (Tang et al. 2008b). *Arabidopsis*, selected as the first angiosperm genome to be sequenced due to its small genome size and minimal DNA sequence duplication, has experienced two 'recent' WGDs, i.e. since its divergence from other members of the *Brassicales* clade (α and β), and a more ancient triplication (γ) shared with most if not all eudicots (Bowers et al. 2003; Tang et al. 2008a; Tang et al. 2008b). Likewise, rice appears to have experienced at least two WGDs, one shared with most if not all cereals (ρ), and another more ancient event (σ) (Tang et al. 2010). Single-gene duplications in angiosperms are also widespread (Freeling et al. 2008; Freeling 2009; Woodhouse et al. 2010).

One avenue for systematic investigation of functional divergence between duplicate genes is comparison of their spatiotemporal expression profiles, comparing degrees of divergence with proxies of duplication age such as synonymous substitution rates (Ks) between duplicate genes. In *Arabidopsis*, the rate of protein sequence evolution is asymmetric in >20% of duplicate pairs and functional diversification of surviving duplicate genes has been proposed to be a major feature of the long-term evolution of polyploids (Blanc and Wolfe 2004). *Arabidopsis* genes created by large-scale duplication events are more evolutionarily conserved in gene expression than those created by small-scale duplication or those that do not lie in duplicate segments, and the time since duplication is correlated with functional divergence of genes (Casneuf et al. 2006). Further, there may be also a strong positive correlation between expression divergence and non-synonymous mutation (Ka) in *Arabidopsis*, and the different modes (segmental, tandem and dispersed) of duplication may affect patterns of expression divergence

(Ganko et al. 2007). *Arabidopsis* duplicated genes show greater expression diversity than singleton genes across closely related species and allopolyploids (Ha et al. 2009). In rice, expression correlation is significantly higher for gene pairs from WGDs or tandem duplications than dispersed duplications, and expression divergence is closely related to divergence time (Li et al. 2009).

Though many studies have investigated the functional divergence and retention of duplicate genes, conclusions are often contradictory, e.g. gene retention has been attributed to either neofunctionalization (Zhang and Cohn 2008; Kassahn et al. 2009) or genetic redundancy (Gu et al. 2003; Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; DeLuna et al. 2010), and expression divergence between duplicate genes has been suggested to be either time dependent (Casneuf et al. 2006; Li et al. 2009) or selection dependent (Ganko et al. 2007). The fates of duplicate genes may be influenced by different modes of gene duplication, which have been suggested to retain genes in a biased manner (Freeling 2009). With much richer expression and annotation data available now than for most prior studies, and improved ability to discern various mechanisms of gene duplication, we find merit in re-examining some existing hypotheses and exploring some new hypotheses regarding the consequences of gene duplication. Here, we related multiple types of genomic data to gene expression divergence in two angiosperm species, *Arabidopsis* and *Oryza* (rice), to formally test possible evolutionary patterns (hypotheses). A far richer volume of analyzed microarray data than was available in prior studies improves the robustness of statistical analyses.

Results

A total of 4,566 Affymetrix *Arabidopsis* Genome ATH1 Arrays and 508 Affymetrix GeneChip Rice Genome Arrays were used to generate the expression profiles of 22,810

Arabidopsis genes and 27,910 rice genes. We classified gene duplications into six modes: WGD, tandem, proximal, DNA-based transposed, retrotransposed and dispersed duplication, according to the procedure shown in Figure 6.1 and described in methods. Note that in this study, a gene may have up to five potential duplication relationships, depending on the number of BLASTP hits. For WGD duplicates, redundant duplication relationships were removed using co-linearity restrictions. If a gene was created by single-gene duplications, all possible duplication relationships were considered. However, redundant duplication relationships in single-gene duplications did not enlarge the gene set created by each duplication mode. In a distantly transposed duplication, one duplicate gene is the parental (ancestral) copy while the other is the transposed (derived) copy, at a novel locus. Dispersed duplications, which we cannot attribute to specific mechanisms, are regarded as a control group. The number of pairs of duplicate genes and number of unique genes (i.e. number of created genes) in each mode of duplication is summarized in Table 6.1. A total of 2,981 α, 1,161 β and 417 γ WGD duplicate pairs in *Arabidopsis*; and 1,712 ρ and 568 γ WGD duplicate pairs in rice, have expression profiles. In this study, the degree of similarity between the expression profiles of a pair of genes across all experiments is measured by the Pearson's correlation coefficient (r). To express in positive values the evolution of gene expression between duplicates or orthologs, we use the term "expression divergence", measured by $1-r$ (Liao and Zhang 2008; Liao et al. 2010).

Gene duplication modes contribute differentially to genetic novelty and redundancy

Expression divergence between duplicate genes was compared across modes of duplication (Figure 6.2). The trends of expression divergence between duplicates in *Arabidopsis* and rice are very similar: DNA-based transposed duplication ≈ retrotransposed duplication > dispersed duplication > proximal duplication > WGD ≈ tandem duplication (both ANOVA

model involving all duplication modes and Tukey's HSD test between adjacent duplication modes are significant at α=0.05). Although retrotransposed duplications have a little higher average expression divergence than DNA-based transposed duplications, the difference is not significant (P-value>0.05). WGDs result in a little higher expression divergence than tandem duplications in *Arabidopsis* but the difference is not significant in rice.

Despite the relatively fast evolution of gene expression shown by distantly transposed duplications, a tendency toward coexpression between genes duplicated by all modes can be observed by comparison with 10,000 randomly selected gene pairs (Figure 6.2). Furthermore, we used r<0.371 and r<0.621 (95% quantile of the *r* values obtained from random gene pairs) as criteria for determining that two duplicate genes have diverged in expression in *Arabidopsis* and rice respectively (Gu et al. 2002; Blanc and Wolfe 2004). The proportions of divergent expression between genes duplicated by different modes are shown in Table 6.2. All these data suggest that the extent of expression divergence of retained duplicates is affected by the duplication mechanism: WGD and tandem duplicates are more likely to maintain their original expression patterns, proximal duplications show intermediate divergence, and distantly transposed duplications tend to have the biggest changes of gene expression profiles.

Computationally, genetic redundancy may be inferred from simultaneous conservation in protein sequences that determine molecular functions, and expression patterns which determine biological processes (Liljegren et al. 2000; Briggs et al. 2006). WGD and tandem duplicates tend to be simultaneously conserved in protein sequences (using 25% quartile of Ka of all duplicate pairs, i.e. <0.329 in *Arabidopsis* and <0.383 in rice, as criteria) and in gene expression (using $r \geq 0.371$ in *Arabidopsis* and $r \geq 0.621$ in rice as criteria), while distantly transposed and dispersed duplicates have a random association (assuming that conservation in protein sequences

and gene expression were independent in the pooled duplicate genes) between these parameters, and proximal duplicates fall in between (Table 6.3).

Expression levels differ between the genes created by different duplication modes (Figure 6.3). WGD and dispersed duplicates have higher gene expression levels than tandem, proximal and distantly transposed duplications (2-sample *t*-tests are significant at α=0.05). The higher expression of WGD duplicates is consistent with their retention due to dosage amplification, a theory which has been proven in yeast (Papp et al. 2003; Vitkup et al. 2006; Conant and Wolfe 2007). Potentially transposon mediated gene duplications including tandem, proximal and distantly transposed duplications tend to be associated with lower gene expression levels than other duplication modes (Figure 6.3). Dispersed duplication, with unclear genetic mechanisms so far, is associated with gene expression levels comparable to WGD.

Expression divergence following polyploidy

Since its divergence from other Brassicales, *Arabidopsis* experienced two WGDs (α and β), while sharing a more ancient genome triplication (γ) with all rosids and perhaps all eudicots (Bowers et al. 2003; Tang et al. 2008a; Tang et al. 2008b). Rice has experienced two WGDs: the ρ event shared with all Poaceae, and the more ancient σ event (Tang et al. 2010). Although expression divergence has been compared between WGD and single-gene duplications (Casneuf et al. 2006; Ganko et al. 2007; Li et al. 2009), the combinational effects of different WGD events on expression divergence have not been addressed. We propose that WGD events themselves, together with the subsequent 'adaptation' of the resulting genome to the newly-duplicated state, may accelerate evolution, contributing to variation in expression divergence sometimes attributed to time (usually measured by Ks) alone (Casneuf et al. 2006; Li et al. 2009).

To further investigate the combinational effects of multiple WGD events, we compared the expression divergence of duplicates from different WGD events (Figure 6.4). Not surprisingly, expression divergence between the WGD duplicates of more ancient events tends to be larger: γ duplicates > β duplicates > α duplicates in *Arabidopsis*, and σ duplicates > ρ duplicates in rice (both ANOVA model involving all WGD events and Tukey's HSD test between adjacent WGD events are significant at α=0.05). Next, we fitted a curve between expression divergence and Ks for each WGD event using a smooth spline with 10 degrees of freedom available in R packages (Figure 6.4). We found no significant correlation between expression divergence and Ks within the more ancient *Arabidopsis* β duplicates (r=0.036, P-value=0.241) or γ duplicates (r=-0.008, P=0.883), or rice σ duplicates (r=0.045, P=0.307) but correlations are significant within the most recent *Arabidopsis* α duplicates (r=0.126, P= 1.364×10^{-11}) and rice ρ duplicates (r=0.105, P= 2.054×10^{-5}). Further, we conducted a power analysis for the non-significant correlations. We found that at α=0.05, the non-significant correlations (β, γ and σ duplicates) did not have higher power than conventionally desired (>0.8) while significant correlations (α and ρ duplicates) had power greater than 0.98, confirming that the relationship between expression divergence and Ks differs among different WGD events.

WGD events themselves influence gene expression divergence, with more ancient WGD duplicated genes likely to have greater expression divergence than more recent duplications, even if both have similar Ks (Figure 6.5). To support this hypothesis statistically, we coded the α, β and γ events by 1, 2 and 3 in *Arabidopsis* and the ρ and σ events by 1 and 2 in rice. Then different linear regression models of expression divergence on Ks and/or WGD codes were fit in *Arabidopsis* and rice respectively. All regression models and their coefficients were statistically significant. For both *Arabidopsis* and rice, the model which counts both Ks and the number of

WGD events that duplicate genes underwent results in the highest adjusted R^2 and lowest Akaike information criterion (AIC) (Table 6.4) with significant nonzero slopes of all coefficients, supporting the hypothesis that WGD events themselves, in addition to Ks, can lead to increased expression divergence between duplicates.

Selection after WGD events may constrain expression divergence of some duplicates. To examine this question, we studied the 25% of WGD duplicate pairs with most conserved expression at each WGD event. At a P-value threshold of 0.05 by Fisher's exact test (corrected for multiple tests), specific GO terms / Pfam domains were associated with conserved expression at each WGD event, and some recurred across different WGD events, e.g. transcription factor activity (GO:0003700) and ribosome (GO:0005840) for *Arabidopsis* α and γ and rice ρ events; protein biosynthesis (GO:0006412) for *Arabidopsis* α and β and rice ρ events. In contrast, WGD duplicates with divergent expression (25% of pairs with highest *d* values at each event) showed little or no enrichment of specific GO terms / Pfam domains and functional terms did not recur between different WGD events.

Expression divergence between *Arabidopsis* and rice

In that most angiosperms share most genes, changes in expression may be fundamental to angiosperm biodiversity. Previous studies have associated duplicated genes with greater expression diversity than singletons in closely related species of both animals (Gu et al. 2004) and plants (Ha et al. 2009). However, it has been difficult to extend such comparisons to more distant species such as *Arabidopsis*, a eudicot, and rice, a monocot, due to greater difficulty discerning orthology or paralogy. To facilitate the comparison of gene expression data generated by different microarray platforms, we adopted a conceptual framework of comparing coexpression patterns across species (Ihmels et al. 2005) (see Methods). Further, we restricted

our study to 2,012 gene pairs suggested both by DNA sequence similarity and by synteny/collinearity to be orthologs between *Arabidopsis* and rice, downloaded from the PGDD database (Tang et al. 2008a; Tang et al. 2008b). The comparison of expression divergence between different types of orthologs shows the following trend: duplicate-duplicate>singleton-duplicate>singleton-singleton (Figure 6.6), with P-values of 0.049 between duplicate-duplicate and singleton-duplicate and 0.010 between singleton-duplicate and singleton-singleton using two-sample *t*-tests. This finding supports that singletons are more conserved in expression than duplicated genes, consistent with the hypothesis that one consequence of gene duplication is increased expression diversity.

Expression divergence may be correlated with both Ks and Ka

Divergence in coding sequences can be denoted by Ks, which indicates putatively-neutral mutations that are synonymous at the amino acid level, or by Ka, which indicates altered amino acids suggestive of the action of selection on gene function. The correlations between expression divergence and coding sequence divergence in angiosperms have been widely discussed (Casneuf et al. 2006; Ganko et al. 2007; Li et al. 2009) but conclusions were inconsistent: Casneuf et al. and Li et al. suggested that Ks is closely correlated with gene expression divergence, while Ganko et al. found little correlation. Since microarray data contain a high level of noise and previous studies often relied on small sets of microarray data or only one species, our analysis of "all arrays" and two highly-divergent species may have broader inference space.

The distributions of Ka or Ks differ markedly for different gene duplication modes, but are relatively consistent in *Arabidopsis* and rice (Figure 6.7). Tandem/proximal and WGD duplicates have qualitatively lower Ks (putatively reflecting younger age) than distantly transposed (DNA and RNA) or dispersed duplicates, the distinction being much clearer in the

small genome of *Arabidopsis* (Figure 6.7A) than the 3x larger and more repeat-rich genome of rice (Figure 6.7B). Within these qualitative distinctions, quantitative differences among the categories are also evident and largely consistent, with relative Ks (putatively age) of duplications following the trend of: dispersed > distantly transposed > WGD > proximal > tandem (both ANOVA model involving all duplication modes and Tukey's HSD test between adjacent duplication modes are significant at α=0.05). Retrotransposed duplicates differ slightly in the two taxa, being similar to DNA-based transposed duplicates in *Arabidopsis*, and to dispersed duplicates in rice. The trend of Ka shows the same qualitative distinction as that of Ks (Figure 6.7C and D), but differing in the quantitative trend with amino-acid altering mutation frequencies being retrotransposed > dispersed > DNA-based transposed > proximal ≈ WGD ≈ tandem (both ANOVA model involving all duplication modes and Tukey's HSD test between adjacent duplication modes are significant at α=0.05). WGD duplicates are more functionally constrained, with higher Ks but equal or lower Ka than proximal duplicates. These data do not show the conventional L-shaped distribution for dispersed and distantly transposed duplicates, because the filters employed in gene selection focus this analysis only on genes that have survived a long time, implying that the genes serve important functions.

Relationships between coding sequence divergence and expression divergence are heterogeneous, and differ among gene duplication modes. For WGD duplicates, expression divergence is significantly correlated with both Ka and Ks in both *Arabidopsis* and rice, although the strength of the correlations is progressively weaker for more ancient duplications and in some cases reaches non-significance (Table 6.5). Expression divergence is also significantly correlated with both Ka and Ks among proximal duplicates. Tandem duplicates differ in the two taxa, with those of rice resembling WGD genes with expression divergence significantly correlated with

both Ka and Ks, and those of *Arabidopsis* resembling distantly transposed duplications with marginal and sometimes non-significant correlation.

While age and functional divergence are more closely related to expression divergence in WGD genes than those resulting from other duplication modes, this does not reflect a lack of expression divergence among other gene duplicates. Indeed, proximal duplication is associated with higher expression divergence than WGD, despite its smaller average Ks. Likewise, DNA-based transposed duplication is associated with higher expression divergence than dispersed duplication, despite smaller Ks (Table 6.6).

In partial summary, expression divergence between duplicate genes may be affected by duplication modes, as well as by the 'age' (Ks) of the duplicated genes, i.e. gene expression divergence may differ among duplication modes at the same Ks or Ka levels. To further validate this claim, we fit a smooth spline curve between expression divergence and Ks or Ka for each duplication mode (Figure 6.8). While these curves fluctuate markedly, at fixed Ks or Ka levels distantly transposed duplications (for example) are generally associated with higher expression divergence between duplicates than WGD or tandem duplications.

DNA methylation of the promoter regions has little impact on expression divergence

Epigenetic mechanisms such as DNA methylation have been suggested to potentially differentiate newly arisen duplicate genes (Rapp and Wendel 2005; Chen and Ni 2006) as well as orthologous genes across closely related species (Ha et al. 2009). Transcriptional silencing has often been associated with DNA methylation in promoter regions (Zhang et al. 2006; Zilberman et al. 2007). Using data on genome-wide DNA methylation status for both *Arabidopsis* and rice (Feng et al. 2010), we examined whether DNA methylation status in promoter regions is related to expression divergence between duplicates or between orthologs. This comparison carries an

inherent assumption that methylation patterns are relatively static and generally apply to all of the microarray studies. A gene promoter region was considered to be methylated if 2 or more adjacent probes are methylated within the region (Zilberman et al. 2007). Proportions of pairs of duplicates that differ in DNA methylation status in promoter regions, separated by gene duplication modes, are summarized in Table 6.7. Distantly transposed duplications appear somewhat more likely to differ in DNA methylation status than other duplication modes. However, the duplicate genes that differ in DNA methylation status in promoter regions do not have more divergent expression than those that have the same DNA methylation status, within any duplication mode (negative data are not shown). Likewise, different methylation status among orthologs also showed no significant relationship to expression divergence, although we confirmed that singletons are a little more likely to be methylated in promoter regions than duplicates (Table 6.8), as proposed by others (Ha et al. 2009). These analyses suggest that the mechanisms by which DNA methylation status affects expression divergence between homologous genes may be complicated, and direct association may not be informative for unraveling such mechanisms.

Gene family members may have non-random patterns of origin

The diversity of gene duplication mechanisms and patterns of gene expression divergence raise questions about how gene families expand and how their members have been retained in the history of evolution. WGD duplicates are differentially retained across different gene functional classifications (Blanc and Wolfe 2004; Maere et al. 2005; Chapman et al. 2006; Paterson et al. 2006). However, we suggest that gene families may be more informative units than functional terms for investigating patterns of gene origin, as duplication relationships in gene families are clearer. Based on our findings above, both functional divergence and

redundancy may contribute to retention of duplicate genes. Furthermore, because the degrees of functional diversification are not equal across gene families and gene duplication modes add additional heterogeneity to patterns of functional divergence, it is possible that gene family members may have non-random patterns of origin, e.g. the gene families with high functional diversification may be enriched with distantly transposed duplications while those families contributing to genetic redundancy are likely to be enriched with WGD duplications.

To examine these questions, we investigated the gene duplication modes of 126 *Arabidopsis* and 24 rice published gene families of 10 or more genes, available at TAIR (http://www.arabidopsis.org/) and Michigan State University (http://rice.plantbiology.msu.edu/) respectively. By using Bonferroni-corrected Fisher's exact test, we found that 64 (50.8%) *Arabidopsis* gene families and 19 (79.2%) rice gene families are enriched for at least one gene duplication mode at α=0.05. For example, DNA-based transposed duplications are enriched in disease resistance gene homologs and the cytochrome P450 gene family (Figure 6.9A-C). Disease resistance gene homologs, most of which have nucleotide binding site-leucine rich repeat (NBS-LRR) domains, express at different levels and tissue specificities, and function in diverse biological processes in *Arabidopsis* (Tan et al. 2007). P450s also express in many tissues in a tissue specific manner and are involved in diverse metabolic processes (Mizutani et al. 1998; Xu et al. 2001). The cytochrome P450 family also shows enrichment for DNA-based transposed duplications in rice. Thus, these two gene families may have achieved functional and expression diversity through some combination of transposition activity and retention of distantly transposed duplicates. Interestingly, these two families are also enriched with proximal duplications, again often associated with greater expression divergence than WGD despite generally similar coding sequence divergence.

WGD duplicates are enriched in other gene families, such as the cytoplasmic ribosomal protein gene family, and C2H2 zinc finger proteins (Figure 6.9D-F). In *Arabidopsis*, a large number of ribosomal genes are co-regulated (Jen et al. 2006). C2H2 zinc finger proteins have been shown to be involved in some basic biological processes such as transcriptional regulation, RNA metabolism and chromatin-remodeling (Englbrecht et al. 2004). Furthermore, C2H2 zinc finger proteins are enriched with retained WGD duplicates in both *Arabidopsis* and rice. Our analyses suggest that gene family members may have common non-random patterns of origin, that recur independently in different evolutionary lineages (such as monocots, and dicots, studied here), and that such patterns may result from specific biological functions and evolutionary needs.

Discussion

In two species that sample a wide range of tissues and physiological conditions in major angiosperm lineages diverged by about 140-170 million years (Hedges et al. 2006) and affected by at least 5 different genome duplication events, we have compared expression divergence between positional orthologs and between genes duplicated by several additional mechanisms. Both neo-functionalization and genetic redundancy can result in retention of duplicate genes. WGD duplicates generally are more frequently associated with genetic redundancy than genes resulting from other duplication modes, partly due to dosage amplification. Tandem duplications also contribute to genetic redundancy, while other duplication modes are more frequently associated with evolutionary novelty. Potentially transposon mediated gene duplications tend to reduce gene expression levels. Expression divergence between duplicates is discernibly related to duplication modes, WGD events, Ka, Ks, and possibly the DNA methylation status of their promoter regions. However, the contribution of each factor is heterogeneous among duplication

modes, and new factors as well as combinatorial effects of different factors are worth further investigation. Gene loss may retard inter-species expression divergence, as singletons are generally more conserved in gene expression than duplicates. Members of different gene families have non-random patterns of origin, and such patterns may be similar between *Arabidopsis* and rice.

The use of large volumes of data and inclusion of as many genes as possible may help to mitigate factors specific to particular developmental states, noise associated with microarray data, and bias reflecting features specific to particular gene families. For example, we have found that the correlations between expression divergence and Ks are not consistent within gene duplication modes (Figure 6.5 and 6.8). For WGD duplicates, significant correlations only exist in those generated by recent WGD events - if only relatively 'young' WGD duplicates are studied, the correlations may be overestimated. Moreover, such correlations are not uniformly distributed among Ks levels - at low Ks levels (<1), all duplication modes may show correlations.

We find evidence for duplicate gene retention by both neo-functionalization and genetic redundancy, seemingly at opposite ends of the spectrum of possible fates of duplicated gene pairs. Genetic redundancy has clear biological significance, i.e. provision of buffering capacity (Chapman et al. 2006; VanderSluis et al. 2010) and/or dosage balance (Papp et al. 2003; Veitia 2003; Maere et al. 2005; Paterson et al. 2010), and seems most closely related to WGD or tandem duplicates. The origins of genetic novelty, of clear biological significance in occupation of new niches or adaptation to new environments, may lie more with the greater expression divergence and more independent evolution of distantly transposed and dispersed duplications.

Proximal duplication is more balanced in its contributions to genetic novelty and redundancy than other gene duplication modes.

Detailed delineation of gene duplication modes reveals some new trends. Prior studies classified genes into as few as two types (anchors generated by polyploidy, and non-anchors generated by single-gene duplication (Casneuf et al. 2006)), or as many as three types (segmental, tandem and dispersed: (Ganko et al. 2007)). In this study, we have attempted to distinguish DNA/RNA-based transposed from dispersed duplication, and proximal from tandem duplication. DNA-based transposed duplications tend to evolve faster in expression while having smaller Ks than dispersed duplicates. Tandem duplicates diverge slower in gene expression than proximal duplicates. Proximal duplicates tend to diverge faster in expression than WGD duplicates, though concerted evolution (Wang et al. 2007) may homogenize their coding sequences.

The factors that affect expression divergence are complex

Our analyses suggest that it may be inappropriate to make generalizations about levels and patterns of expression divergence across gene duplication modes. Ks, putatively a proxy for age, seems to be related to expression divergence only within a subset of duplication modes and largely only among younger duplicates. Ka, putatively a proxy for functional change, also shows statistically significant and heterogeneous relationships to expression divergence. The level of these correlations is very low, even in recent WGD duplicates.

Although expression divergence between duplicates is often significantly correlated with coding sequence divergence, it is well known that gene expression is also regulated by other genomic regions such as promoters, 5'UTRs, and 3'UTRs. The correlations between expression divergence and nucleotide substitution rates (μ) of different genomic regions for pairs of

duplicates are summarized in Table 6.9. WGD duplicates show significant correlations between expression divergence and nucleotide substitution rates in all three regions. These correlations become marginal and often non-significant among tandem duplicates. Expression divergence of proximal duplicates is more closely associated with divergence in promoters, 5'UTRs and 3'UTRs than coding sequences. Expression divergence of DNA-based transposed duplicates seem to be most related to Ka and μ of 3'UTRs. Expression divergence of dispersed duplicates is very slightly correlated with Ka but not with other substitution rates. Retrotransposed duplication is least related to any type of sequence divergence, consistent with its general separation of a gene from its native regulatory elements.

In partial summary, expression divergence between duplicate genes may be affected by different and multiple genetic factors depending on the causal duplication mechanism. For pairs of orthologs between *Arabidopsis* and rice, expression divergence seems only correlated with Ka (Table 6.5 and Table 6.9). Single-gene duplications including translocated and tandem/proximal duplications have been suggested to be much more prone to promoter disruption than WGD (Casneuf et al. 2006). We examined this hypothesis using >45% sequence identity as criterion for determining duplicated (non-disrupted) promoter regions, finding proximal duplicates to have higher proportions of duplicated promoter regions than WGD duplicates (Table 6.10). This finding seems to contradict the greater expression divergence of proximal duplicates than WGD duplicates. Thus, we note that each of the investigated genetic/epi-genetic factors may only explain a small portion of the variation of expression divergence between duplicate genes, and perhaps only for certain duplication modes. New factors that may affect expression divergence and how different factors work together are worth investigation.

Possible non-random associations between duplication mode and population size

WGD is often associated with speciation in plants (Stebbins 1982; Wood et al. 2009). If ancestral polyploidy was attendant with speciation, new species would have likely initially faced very small N_e (i.e. effective population size), weak selection, high drift and high mutational load. This could put a premium on buffering, but allow little chance for beneficial mutations. On the other hand, small-scale duplications may have been only infrequently associated with speciation, if at all. Thus they might be more likely to arise in established populations with larger N_e and more efficient selection, all putting a greater premium on evolutionary novelty to attain fixation. A hypothesis worthy of further investigation is that non-random associations between duplication mode and population size have shaped which specific genes and functional variations are retained.

Methods

Genome annotation

Genome annotations were obtained from TAIR (http://www.arabidopsis.org) for *Arabidopsis*, and from the Rice Genome Annotation Project data (http://rice.plantbiology.msu.edu) for rice. Gene structures were retrieved using ENSEMBL Biomart (http://plants.ensembl.org/biomart/martview).

Gene expression data

To reliably assess the expression divergence between duplicates or between orthologs, we used as many publicly available microarray datasets as possible, all of which were obtained from NCBI's GEO (http://www.ncbi.nlm.nih.gov/geo/). At the time of retrieval, 6,009 samples existed for the Affymetrix *Arabidopsis* ATH1 Genome Array (GEO platform GPL198), of which 800 were not available and a total of 5,209 CEL files were downloaded. 550 CEL files for the

Affymetrix GeneChip Rice Genome Array (GEO platform GPL2020) were downloaded, of which 13 were removed due to incorrect array types. For both *Arabidopsis* and rice raw expression data, RMA normalization was performed using the RMAExpress software (http://rmaexpress.bmbolstad.com) across the entire dataset. Outliers were detected using the arrayQualityMetrics (Kauffmann et al. 2009) Bioconductor package, which implements three different statistical tests to identify outliers. A total of 443 and 29 samples were detected as outliers and removed in *Arabidopsis* and rice respectively. Thus, 4,566 and 508 samples remained for *Arabidopsis* and rice, respectively. The annotation files (Release 30) of these two arrays were downloaded from the Affymetrix website (http://www.affymetrix.com), containing 22,810 *Arabidopsis* genes and 27,910 rice genes. For a gene, there may be multiple probe sets or multiple types of probe sets available on the array. However, a general rule for selection of a probe set that best represents the gene's expression profile has not been resolved yet (Elbez et al. 2006; Liao and Zhang 2006). In this study, inclusion or exclusion of "sub-optimal" probe sets with suffix "_s_at" or "_x_at" that are suspected of potential cross-hybridization (may be not sub-optimal in practice according to ref. (Elbez et al. 2006; Liao and Zhang 2006)) had only trivial effects. Thus, to survey as many genes as possible, all types of probe sets were considered, and for a gene with multiple probe sets, we used the first probe set according to alphabetic sorting to represent its expression profile.

Analysis of expression data

Similarity between the expression profiles of two duplicate genes within species was initially measured by either Pearson's (denoted by PCC or *r*) or Spearman's correlation coefficient. Note that all replicate chips were retained and correlations were computed across all individual chips. These two measures generated highly consistent results, and thus we only

showed the statistics measured by Pearson's correlation coefficient. The expression divergence between two duplicate genes or orthologs was measured by 1-*r* (Liao and Zhang 2008; Liao et al. 2010).

Orthologous gene pairs compared between *Arabidopsis* and rice were restricted to 2,012 pairs of orthologs located at corresponding loci in paired syntenic blocks between *Arabidopsis* and rice as identified by MCScan (Tang et al. 2008b), and having expression profiles on the arrays. To assess the expression conservation (EC) for a pair of *Arabidopsis*-rice orthologs, we adopted a conceptual framework of comparing coexpression patterns across species (Ihmels et al. 2005) implemented in several other studies similar to ours (Dutilh et al. 2006; Tirosh and Barkai 2007; Essien et al. 2008; Wang and Rekaya 2009; Wang et al. 2010). In this study, the framework can be described as:

1) The expression matrices, **A** and **B,** in *Arabidopsis* and rice respectively, are restricted to genes for which orthology relationships have been identified and ordered accordingly (i.e., equivalent rows of the two matrices correspond to the expression profiles of a pair of orthologs):

$$\mathbf{A} = [\mathbf{a}_i]_{i=1,\ldots,k}$$
$$\mathbf{B} = [\mathbf{b}_i]_{i=1,\ldots,k}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the vectors of expression profiles for any pair *i* of orthologs for *Arabidopsis* and rice, respectively, and *k* is the number of orthologous gene pairs.

2) **A** and **B** are then converted into two pair-wise correlation matrices, $\mathbf{R}^A$ and $\mathbf{R}^B$, by computing the PCCs between the expression profile of each gene and that of any other gene in each species separately:

$$\mathbf{R}^A = [PCC(\mathbf{a}_i, \mathbf{a}_g)]_{i=1,\ldots,k;g=1,\ldots,k}$$
$$\mathbf{R}^B = [PCC(\mathbf{b}_i, \mathbf{b}_g)]_{i=1,\ldots,k;g=1,\ldots,k}$$

3) The expression conservation for an orthologous gene pair *i* is computed as:

$$EC(i) = PCC(R^A_{i,g}, R^B_{i,g}),\ g = 1,...,k$$

Its corresponding expression divergence is 1-*EC(i)*.

Identification of different modes of gene duplications

The populations of potential gene duplications in *Arabidopsis* or rice were identified using BLASTP. Only the top five non-self protein matches that met a threshold of $E < 10^{-10}$ were considered. Genes without BLASTP hits that met a threshold of $E < 10^{-10}$ were deemed singletons. Pairs of WGD duplicates were downloaded from the PGDD database (Tang et al. 2008a; Tang et al. 2008b). Pairs of α, β, γ duplicates in *Arabidopsis* and pairs of ρ, σ duplicates in rice were obtained from published lists (Bowers et al. 2003; Tang et al. 2010). Single-gene duplications were derived by excluding pairs of WGD duplicates from the population of gene duplications. Tandem duplications were defined as being adjacent to each other on the same chromosome. Proximal duplications were defined as non-tandem genes within 20 annotated genes of each other on the same chromosome (Ganko et al. 2007).

The remaining single-gene duplications (after deducting tandem and proximal duplications) were searched for distant single-gene transposed duplications. To accomplish this aim, genes at ancestral chromosomal positions need to be discerned by aligning syntenic blocks within and between species (Freeling et al. 2008; Tang et al. 2008b). Angiosperm syntenic blocks were downloaded from the Plant Genome Duplication Database (PGDD), available at http:// chibba.agtec.uga.edu/duplication. At the time of retrieval, PGDD provided syntenic blocks within and between 10 species including *Arabidopsis thaliana, Carica papaya, Prunus persica, Populus trichocarpa, Medicago truncatula, Glycine max, Vitis vinifera, Brachypodium distachyon, Oryza sativa, Sorghum bicolor, Zea mays* (Tang et al. 2008a; Tang et al. 2008b). An

Arabidopsis or rice gene locus was regarded as ancestral if the resident gene along with any of its homologous genes (paralogs/orthologs) occur at corresponding loci within any pair of syntenic blocks in PGDD. Using this criterion, the population of *Arabidopsis*/rice genes was divided into two subsets: genes at ancestral loci and genes that were transposed. For a pair of distantly transposed duplicate genes, we required that one copy was at its ancestral locus and the other was at a non-ancestral locus, named the parental copy and transposed copy respectively. If the parental copy has more than two exons and the transposed copy is intronless, we inferred that this pair of duplicate genes occurred by retrotransposition (RNA-based transposition). If both copies have a single exon, the pair of duplicates was unclassified. For other cases of a pair of distantly transposed duplicate genes, we inferred that the duplication occurred by DNA-based transposition. The remaining single-gene duplications in the population, i.e. after deducting WGD, tandem, proximal, DNA-based transposed and retrotransposed duplications from the BLASTP output, were classified as dispersed duplications. After pairs of duplicate genes in each duplication mode were identified, we assigned a unique origin to each duplicated gene, according to the following order of priority: WGD>tandem>proximal>retrotransposed>DNA-based transposed>dispersed.

GO/Pfam enrichment analysis

GO/Pfam enrichment analysis was performed using Fisher's exact test. The P-value was calculated for the null hypothesis that there is no association between a subset of genes and a particular functional/domain category and was corrected with the total number of terms to account for multiple comparisons.

Assessing DNA sequence divergence

Coding sequence divergence between a pair of genes was denoted by either non-synonymous (Ka) or synonymous (Ks) substitution rates. Protein sequences were aligned using Clustalw (Thompson et al. 1994) with default parameters. The protein alignment was then converted to DNA alignment using the "Bio::Align::Utilities" module of the BioPerl package (http://www.bioperl.org/). Ka and Ks were estimated by Nei-Gojobori statistics (Nei and Gojobori 1986), available through the "Bio::Align::DNAStatistics" module of the BioPerl package. Note that the "Bio::Align::DNAStatistics" module may generate invalid Ka or Ks for some duplicate gene pairs due to mis-alignments, which were ruled out from related analysis. All levels of valid Ka or Ks values were considered in related statistical analyses. Because distributions of Ka or Ks were centered at low levels (~1.0), in related figures, to improve their clarity, we only displayed Ka or Ks values between 0 and 2.0.

The promoter region of a gene was restricted to a maximum of 1,000 bp upstream of the transcription start site (TSS) or less if the nearest adjacent upstream gene is closer than 1,000 bp. For a pair of genes, the divergence of promoter sequences was indicated by their Jukes-Cantor nucleotide substitution rate (μ) (Jukes and Cantor 1969), which is available through the "Bio::Align::DNAStatistics" module of the BioPerl package. The divergence in 5'UTR and 3'UTR is also measured by nucleotide substitution rates (μ). Note that the "Bio::Align::DNAStatistics" module may not output μ if the distance between two input nucleotide sequences is too near or too far. Duplicate gene pairs lacking estimation of μ in the promoter region, 5'UTR or 3'UTR were removed from related analysis.

DNA methylation data and its analysis

Arabidopsis and rice genome-wide DNA methylation data were obtained from GEO (accession number: GSE21152) (Feng et al. 2010). We chose this study, which provided DNA methylation for both *Arabidopsis* and rice, because the systematic errors between species should be smaller than in data from separate studies. A gene methylated in the promoter region is defined by the presence of two or more adjacent methylated probes within the promoter DNA sequence (Zilberman et al. 2007; Ha et al. 2009).

Gene families

Lists of published gene families were obtained from TAIR (http://www.arabidopsis.org/browse/genefamily/index.jsp) for *Arabidopsis*, and from the Rice Genome Annotation Project data (http://rice.plantbiology.msu.edu/annotation_community_families.shtml) for rice. Only families with more than nine genes were considered. *Arabidopsis* disease resistance gene homologs were downloaded from the NIBLRRS Project website (http://niblrrs.ucdavis.edu/). The Rice Cytochrome P450 gene family was downloaded from the Cytochrome P450 homepage (Nelson 2009).

References

Adams KL, Wendel JF. 2005. Novel patterns of gene expression in polyploid plants. *Trends Genet* **21**(10): 539-543.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N *et al.* 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**(7116): 171-178.

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

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930): 433-438.

Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS. 2006. Unequal genetic redundancies in *Arabidopsis* - a neglected phenomenon? *Trends Plant Sci* **11**(10): 492-498.

Brosius J. 1991. Retroposons - Seeds of Evolution. *Science* **251**(4995): 753-753.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**(2): 343-360.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol* **7**(2): R13.

Chapman BA, Bowers JE, Feltus FA, Paterson AH. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc Natl Acad Sci U S A* **103**(8): 2730-2735.

Chen ZJ, Ni ZF. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* **28**(3): 240-252.

Chen ZJ, Pikaard CS. 1997. Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proc Natl Acad Sci U S A* **94**(7): 3442-3447.

Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, Byers B. 2000. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell* **12**(9): 1551-1568.

Conant GC, Wolfe KH. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol* **3**: 129.

Cusack BP, Wolfe KH. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* **24**(3): 679-686.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet* **4**(7): e1000113.

DeLuna A, Springer M, Kirschner MW, Kishony R. 2010. Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS Biol* **8**(3): e1000347.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. *Nat Genet* **40**(5): 676-681.

Dutilh BE, Huynen MA, Snel B. 2006. A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. *BMC Genomics* **7**: 10.

Elbez Y, Farkash-Amar S, Simon I. 2006. An analysis of intra array repeats: the good, the bad and the non informative. *BMC Genomics* **7**: 136.

Englbrecht CC, Schoof H, Bohm S. 2004. Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome. *BMC Genomics* **5**(1): 39.

Essien K, Hannenhalli S, Stoeckert CJ, Jr. 2008. Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to *Plasmodium* phenotypic diversity. *PLoS One* **3**(9): e3122.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME *et al.* 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**(19): 8689-8694.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-1545.

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

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res* **18**(12): 1924-1937.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol* **24**(10): 2298-2309.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**(6918): 63-66.

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

Gu ZL, Rifkin SA, White KP, Li WH. 2004. Duplicate genes increase gene expression diversity within and between species. *Nat Genet* **36**(6): 577-579.

Ha M, Kim ED, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A* **106**(7): 2295-2300.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23): 2971-2972.

Hughes AL. 1994. The Evolution of Functionally Novel Proteins after Gene Duplication. *P Roy Soc Lond B Bio* **256**(1346): 119-124.

Hughes MK, Hughes AL. 1993. Evolution of Duplicate Genes in a Tetraploid Animal, *Xenopus-Laevis*. *Mol Biol Evol* **10**(6): 1360-1369.

Ihmels J, Bergmann S, Berman J, Barkai N. 2005. Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genet* **1**(3): e39.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A *et al.* 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**(7011): 946-957.

Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, Gilmartin PM, Westhead DR. 2006. The *Arabidopsis* co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. *Plant J* **46**(2): 336-348.

Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**(7008): 569-573.

Jukes TH, Cantor CR. 1969. *Evolution of Protein Molecules*. Academic Press, New York.

Kaessmann H, Vinckenbosch N, Long MY. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**(1): 19-31.

Kafri R, Dahan O, Levy J, Pilpel Y. 2008. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A* **105**(4): 1243-1248.

Kashkush K, Feldman M, Levy AA. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**(4): 1651-1659.

Kashkush K, Feldman M, Levy AA. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* **33**(1): 102-106.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res* **19**(8): 1404-1418.

Kauffmann A, Gentleman R, Huber W. 2009. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**(3): 415-416.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**(6983): 617-624.

Lee HS, Chen ZJ. 2001. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc Natl Acad Sci U S A* **98**(12): 6753-6758.

Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J. 2009. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics* **10 Suppl 6**: S8.

Liao BY, Weng MP, Zhang J. 2010. Contrasting genetic paths to morphological and physiological evolution. *Proc Natl Acad Sci U S A* **107**(16): 7353-7358.

Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**(19): 6987-6992.

Liao BY, Zhang JZ. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**(3): 530-540.

Liljegren SJ, Ditta GS, Eshed HY, Savidge B, Bowman JL, Yanofsky MF. 2000. SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* **404**(6779): 766-770.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494): 1151-1155.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102**(15): 5454-5459.

Mizutani M, Ward E, Ohta D. 1998. Cytochrome P450 superfamily in *Arabidopsis thaliana*: isolation of cDNAs, differential expression, and RFLP mapping of multiple cytochromes P450. *Plant Mol Biol* **37**(1): 39-52.

Musso G, Costanzo M, Huangfu MQ, Smith AM, Paw J, Luis BJS, Boone C, Giaever G, Nislow C, Emili A *et al.* 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res* **18**(7): 1092-1099.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**(5): 418-426.

Nelson DR. 2009. The cytochrome p450 homepage. *Hum Genomics* **4**(1): 59-65.

O'Neill RJ, O'Neill MJ, Graves JA. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**(6680): 68-72.

Ohno S. 1970. *Evolution by gene duplication*. Springer Verlag, New York.

Ozkan H, Levy AA, Feldman M. 2001. Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* **13**(8): 1735-1747.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**(6945): 194-197.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al.* 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**(7229): 551-556.

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

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends Genet* **22**(11): 597-602.

Paterson AH, Freeling M, Tang H, Wang X. 2010. Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* **61**: 349-372.

Rapp RA, Wendel JF. 2005. Epigenetics and plant evolution. *New Phytol* **168**(1): 81-91.

Rodin SN, Riggs AD. 2003. Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol* **56**(6): 718-729.

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. 2001. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* **13**(8): 1749-1759.

Song K, Lu P, Tang K, Osborn TC. 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proc Natl Acad Sci U S A* **92**(17): 7719-7723.

Stebbins GL. 1982. Plant speciation. *Prog Clin Biol Res* **96**: 21-39.

Tan XP, Meyers BC, Kozik A, Al West M, Morgante M, St Clair DA, Bent AF, Michelmore RW. 2007. Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in *Arabidopsis*. *BMC Plant Biol* **7**: 56.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008a. Synteny and collinearity in plant genomes. *Science* **320**(5875): 486-488.

Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci U S A* **107**(1): 472-477.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**(12): 1944-1954.

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

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* **8**(4): R50.

VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol* **6**: 429.

Veitia RA. 2003. Nonlinear effects in macromolecular assembly and dosage sensitivity. *J Theor Biol* **220**(1): 19-25.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol* **7**(5): R39.

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J. 2006. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics* **7**: 447.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177**(3): 1753-1763.

Wang X, Tang H, Bowers JE, Paterson AH. 2009. Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res* **19**(6): 1026-1032.

Wang X, Tang H, Paterson AH. 2011. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell* **23**(1):27-37.

Wang Y, Rekaya R. 2009. A comprehensive analysis of gene expression evolution between humans and mice. *Evol Bioinform Online* **5**: 81-90.

Wang Y, Robbins KR, Rekaya R. 2010. Comparison of computational models for assessing conservation of gene expression across species. *PLoS One* **5**(10): e13239.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**(6634): 708-713.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci U S A* **106**(33): 13875-13879.

Woodhouse MR, Pedersen B, Freeling M. 2010. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet* **6**(5): e1000949.

Xu WY, Bak S, Decker A, Paquette SM, Feyereisen R, Galbraith DW. 2001. Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana*. *Gene* **272**(1-2): 61-74.

Zhang G, Cohn MJ. 2008. Genome duplication and the origin of the vertebrate skeleton. *Curr Opin Genet Dev* **18**(4): 387-393.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE *et al.* 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**(6): 1189-1201.

Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**(5): 479-492.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39**(1): 61-69.

Table 6.1. Numbers of pairs of duplicate genes and unique genes in each mode of gene duplication

Mode of duplication	Number of pairs of duplicate genes (number of those having complete expression profiles)		Number of unique genes (number of those having expression profiles)	
	Arabidopsis	Rice	*Arabidopsis*	Rice
WGD	6,572 (4,979)	3,593 (2,530)	9,455 (8,089)	5,723 (4,829)
Tandem	2,055 (1,055)	1,741(947)	1,586 (977)	2,948 (2,116)
Proximal	3,113 (1,456)	3,816 (1,990)	669 (379)	1,038 (714)
DNA-based transposed	6,367 (4,088)	8,061 (5,225)	2,230 (1,572)	2,948 (2,116)
Retro-transposed	497 (300)	940 (681)	271 (1,71)	491 (391)
Dispersed	34,887 (26,127)	30,574 (21,385)	7,411 (6,182)	8,313 (6,960)

Table 6.2. Proportion of divergent gene expression between duplicates in each mode of gene duplication

Species	WGD	Tandem duplication	Proximal duplication	DNA-based transposed duplication	Retrotransposed duplication	Dispersed duplication
Arabidopsis	0.577	0.555	0.644	0.759	0.767	0.759
Rice	0.813	0.780	0.865	0.916	0.921	0.904

Table 6.3. Proportion of conservation in both protein sequences and gene expression between duplicates in each mode of gene duplication

Species	WGD	Tandem	Proximal	DNA-based transposed	Retro-transposed	Dispersed	Expected
Arabidopsis	0.335	0.328	0.231	0.071	0.051	0.038	0.071
Rice	0.140	0.170	0.099	0.027	0.023	0.021	0.041

Table 6.4. Linear regression of expression divergence on Ks and WGD events (W)

Regression model	Coefficient, P-value			Adjusted	AIC
	a	b_1	b_2	R^2	
Arabidopsis					
$d = a + b_1 \cdot Ks$	0.593, $< 2.2\times10^{-16}$	0.079, $< 2.2\times10^{-16}$	-	0.027	-10706.164
$d = a + b_2 \cdot W$	0.577, $< 2.2\times10^{-16}$	-	0.074, $< 2.2\times10^{-16}$	0.027	-10706.330
$d = a + b_1 \cdot Ks + b_2 \cdot W$	0.559, $< 2.2\times10^{-16}$	0.050, 1.15×10^{-8}	0.047, 1.05×10^{-8}	0.034	-10736.930
Rice					
$d = a + b_1 \cdot Ks$	0.624, $< 2.2\times10^{-16}$	0.081, 1.84×10^{-7}	-	0.012	-4913.4477
$d = a + b_2 \cdot W$	0.587, $< 2.2\times10^{-16}$	-	0.079, 8.28×10^{-7}	0.011	-4916.3561
$d = a + b_1 \cdot Ks + b_2 \cdot W$	0.557, $< 2.2\times10^{-16}$	0.063, 1.44×10^{-4}	0.058, 6.82×10^{-4}	0.017	-4925.9138

Table 6.5. Correlations between expression divergence (d) and coding sequence divergence

Types of homologs	Number of valid gene pairs	Pearson correlation (P-value) between d and	
		Ka	Ks
Arabidopsis duplicates			
WGD	4,682	0.238 ($< 2.2\times10^{-16}$)	0.176 ($< 2.2\times10^{-16}$)
α	2,858	0.247 ($< 2.2\times10^{-16}$)	0.126 (1.364×10^{-11})
β	1,068	0.146 (1.791×10^{-6})	0.036 (0.241)
γ	371	0.060 (0.253)	-0.008 (0.883)
Tandem	1,033	0.015 (0.635)	0.115 (2.137×10^{-4})
Proximal	1,426	0.057 (0.032)	0.113 (1.891×10^{-5})
DNA-based transposed	3,662	0.052 (0.002)	0.023 (0.173)
Retrotransposed	257	0.042 (0.504)	0.142 (0.023)
Dispersed	23,360	0.046 (3.243×10^{-12})	0.047 (1.087×10^{-12})
Rice duplicates			
WGD	2,390	0.112 (4.006×10^{-8})	0.112 (3.984×10^{-8})
ρ	1,630	0.099 (6.519×10^{-5})	0.105 (2.054×10^{-5})
σ	521	0.059 (0.177)	0.045 (0.307)
Tandem	919	0.091 (0.006)	0.087 (0.008)
Proximal	1,898	0.084 (2.389×10^{-4})	0.095 (3.604×10^{-5})
DNA-based transposed	4,687	0.056 (1.126×10^{-4})	0.017 (0.255)
Retrotransposed	613	0.008 (0.839)	0.037 (0.361)
Dispersed	19,397	0.037 (2.225×10^{-7})	0.017 (0.021)
Arabidopsis-rice orthologs	1,290	0.108 (9.468×10^{-5})	0.003 (0.901)

Table 6.6. Comparisons of expression divergence and Ks between WGD and proximal duplication, and between dispersed and DNA-based transposed duplication

Duplication modes	*Arabidopsis*		Rice	
	Mean d (P-value by t-test)	Mean Ks (P-value by t-test)	Mean d (P-value by t-test)	Mean Ks (P-value by t-test)
WGD vs Proximal	0.690 vs 0.731 (2.912×10^{-6})	1.162 vs 0.816 ($< 2.2\times10^{-16}$)	0.690 vs 0.758 (1.47×10^{-12})	0.759 vs 0.619 ($< 2.2\times10^{-16}$)
Dispersed vs DNA-based transposed	0.813 vs 0.825 (0.019)	1.710 vs 1.490 ($< 2.2\times10^{-16}$)	0.821 vs 0.825 (0.490)	1.169 vs 1.490 ($< 2.2\times10^{-16}$)

Table 6.7. Proportion of pairs of duplicates that have changed DNA methylation status in promoter regions

Species	WGD	Tandem duplication	Proximal duplication	DNA-based transposed duplication	Retrotransposed duplication	Dispersed duplication
Arabidopsis	0.303	0.290	0.309	0.387	0.347	0.318
Rice	0.357	0.417	0.404	0.416	0.447	0.385

Table 6.8. Proportion of genes that are methylated in promoter regions

Species	Singletons	Duplicate genes
Arabidopsis	0.185	0.157
Rice	0.224	0.217

Table 6.9. Correlations between expression divergence and different types of sequence divergence

Type of homologs	# of valid gene pairs	Pearson correlation[1] and P-value between gene expression divergence and				
		Ka	Ks	μ of promoter region	μ of 3' UTR	μ of 5' UTR
Arabidopsis						
WGD	2,839	**0.252,** p $< 2.2\times10^{-16}$	**0.209,** p $< 2.2\times10^{-16}$	**0.147,** p= 3.997×10^{-15}	**0.159**, p $< 2.2\times10^{-16}$	**0.124**, p= 3.681×10^{-11}
Tandem	383	0.040, p=0.434	0.187, p= 2.289×10^{-4}	0.263, p= 1.717×10^{-7}	0.260, p= 2.451×10^{-7}	0.256, p= 3.992×10^{-7}
Proximal	379	0.075, p=0.144	0.129, p=0.012	**0.342**, p= 8.09×10^{-12}	**0.217**, p= 2.034×10^{-5}	**0.246**, p= 1.196×10^{-6}
DNA-based transposed	1,483	**0.059,** p=0.023	4.86×10^{-4} , p=0.985	-0.007, p=0.795	**0.057**, p=0.028	0.023 , p=0.373
Retro-transposed	112	-0.091, p=0.341	0.167, p=0.079	0.020, p=0.832	0.006, p=0.947	-0.121, p=0.204
Dispersed	12,295	**0.065**, p= 7.565×10^{-13}	0.016, p=0.073	0.022 , p=0.013	0.011, p=0.220	0.015 , p=0.092

Rice						
WGD	1,182	**0.073**, p=0.012	**0.119**, p= 3.86×10^{-5}	**0.159**, p= 3.859×10^{-8}	**0.099**, p= 6.712×10^{-4}	**0.073**, p=0.012
Tandem	203	0.011, p=0.874	0.095, p=0.178	0.087, p=0.219	-0.011, p=0.868	0.052, p=0.460
Proximal	340	0.079, p=0.146	0.041, p=0.445	**0.017**, p=0.049	**0.138**, p=0.011	**0.119**, p=0.029
DNA-based transposed	1,720	**0.059**, p= 0.015	0.027, p=0.259	0.069, p=0.004	**0.108,** p= 6.089×10^{-6}	0.096, p= 6.702×10^{-5}
Retro-transposed	258	-0.054, p=0.386	0.018, p=0.763	-0.027, p=0.668	-0.014, p=0.821	-0.010, p=0.874
Dispersed	8,549	**0.032**, p=0.003	0.022, p=0.046	0.021, p=0.054	0.025, p=0.023	0.059, p= 4.303×10^{-8}
Orthologs	641	0.131, p= 8.827×10^{-4}	-0.082, p=0.037	0.046, p=0.236	0.043, p=0.282	-0.049, p=0.214

[1]Bold values indicate statistical significance in both *Arabidopsis* and rice

Table 6.10. Proportion of copied promoter regions among duplicates

Species	WGD	Tandem duplication	Proximal duplication	DNA-based transposed duplication	Retrotransposed duplication	Dispersed duplication
Arabidopsis	0.899	0.923	0.927	0.885	0.865	0.871
Rice	0.382	0.431	0.407	0.344	0.327	0.330

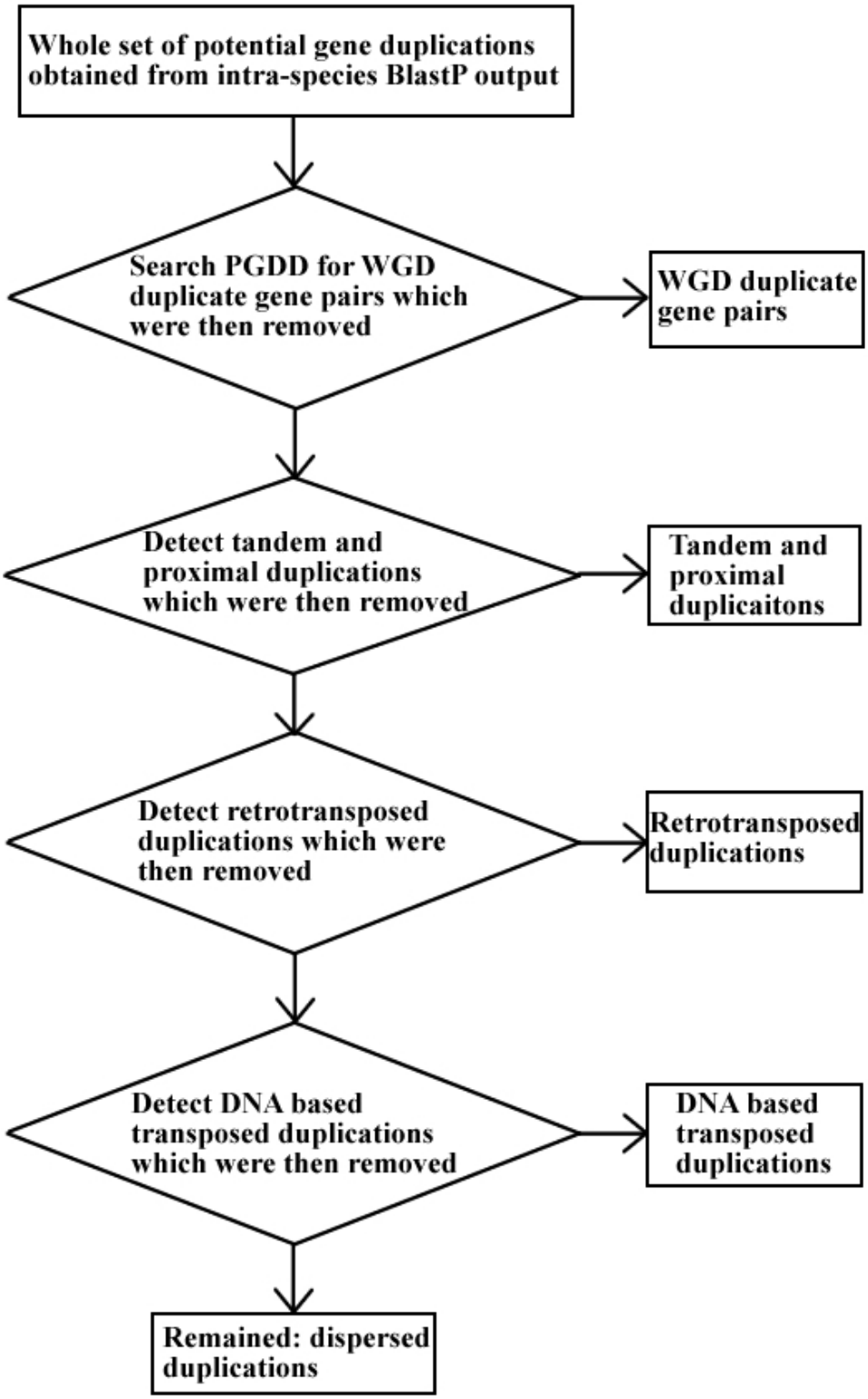

Figure 6.1. Flowchart of the procedure for classifying gene pairs based on mode of duplication

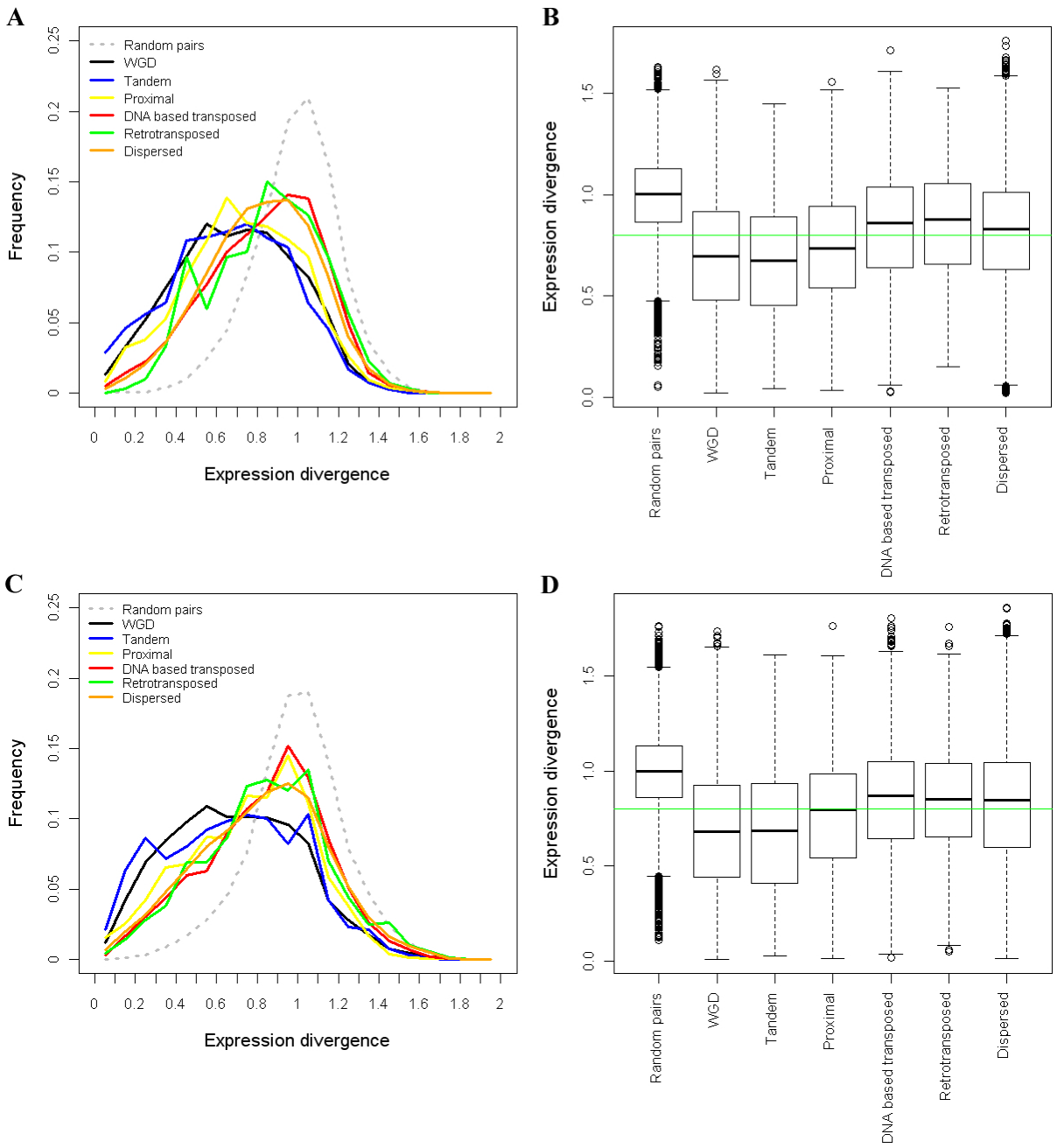

Figure 6.2. Comparison of expression divergence among different modes of gene duplication. (A) Comparison of distributions of expression divergence in Arabidopsis. (B) Comparison of levels of expression divergence in Arabidopsis. (C) Comparison of distributions of expression divergence in rice. (D) Comparison of levels of expression divergence in rice. Green lines in (B, D) indicate average expression divergence across duplication modes.

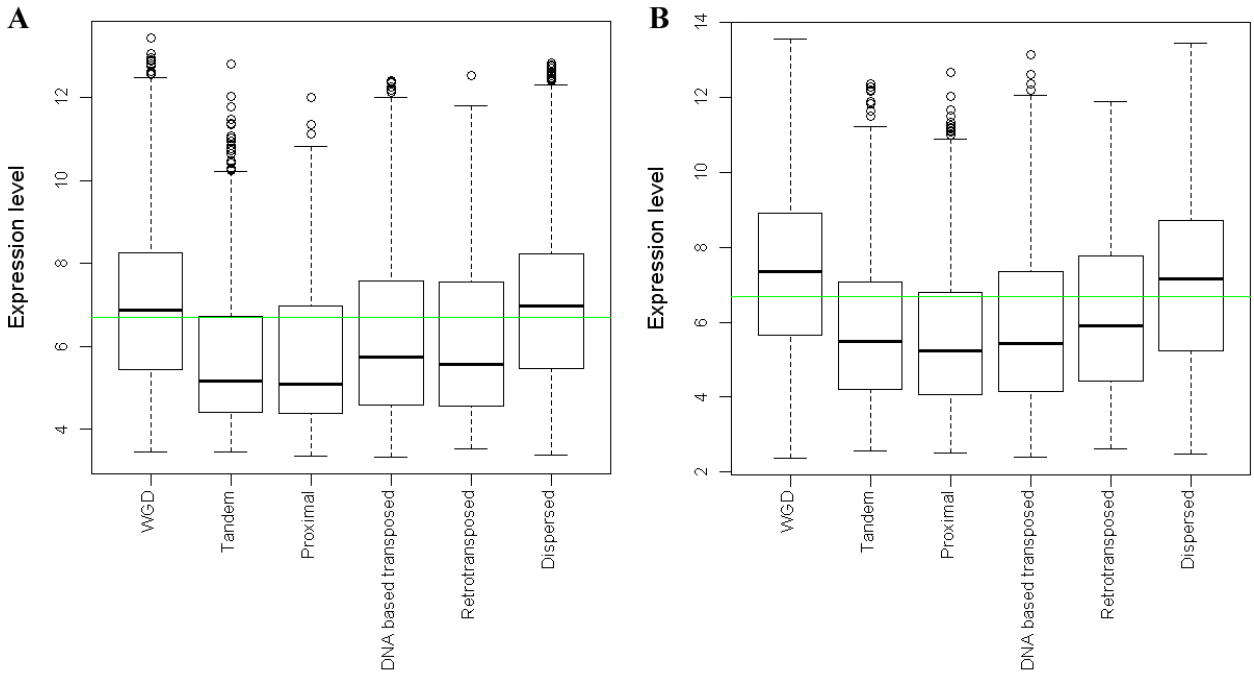

Figure 6.3. Comparison of expression levels between genes created by different duplication modes. (A) Comparison of expression levels between Arabidopsis genes created by different duplication modes. (B) Comparison of expression levels between rice genes created by different duplication modes. Green lines indicate average expression levels**.**

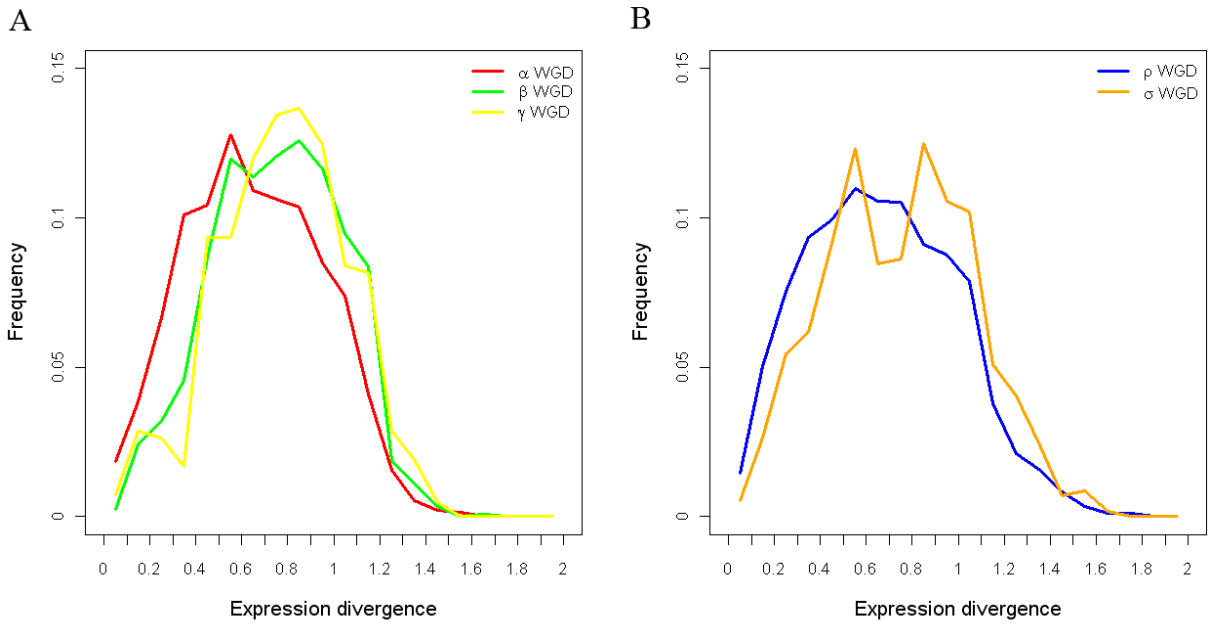

Figure 6.4. Comparison of distributions of expression divergence among different WGD events. (A) Comparison of distributions of expression divergence among different Arabidopsis WGD events. (B) Comparison of distributions of expression divergence among different rice WGD events. α, β and ρ were relatively recent WGD events, while γ and σ were more ancient WGD events.

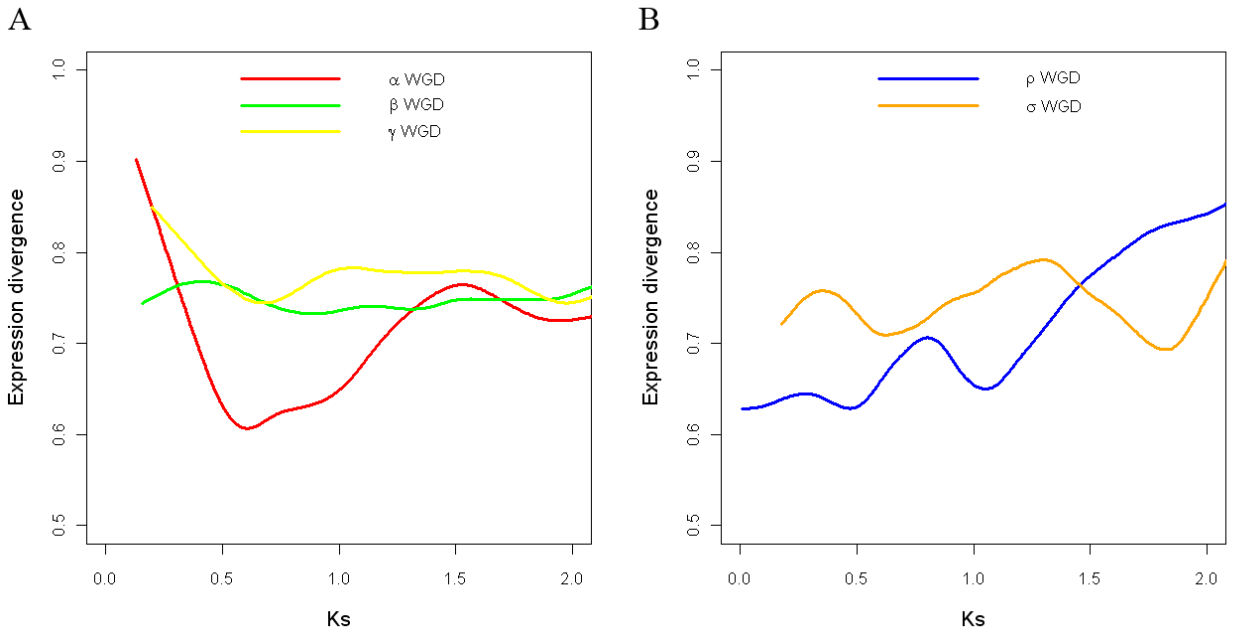

Figure 6.5. Fitted smooth spline curves between expression divergence and Ks for different WGD events. (A) Fitted smooth spline curves between expression divergence and Ks for different Arabidopsis WGD events. (B) Fitted smooth spline curves between expression divergence and Ks for different rice WGD events. α, β and ρ were relatively recent WGD events, while γ and σ were more ancient WGD events.

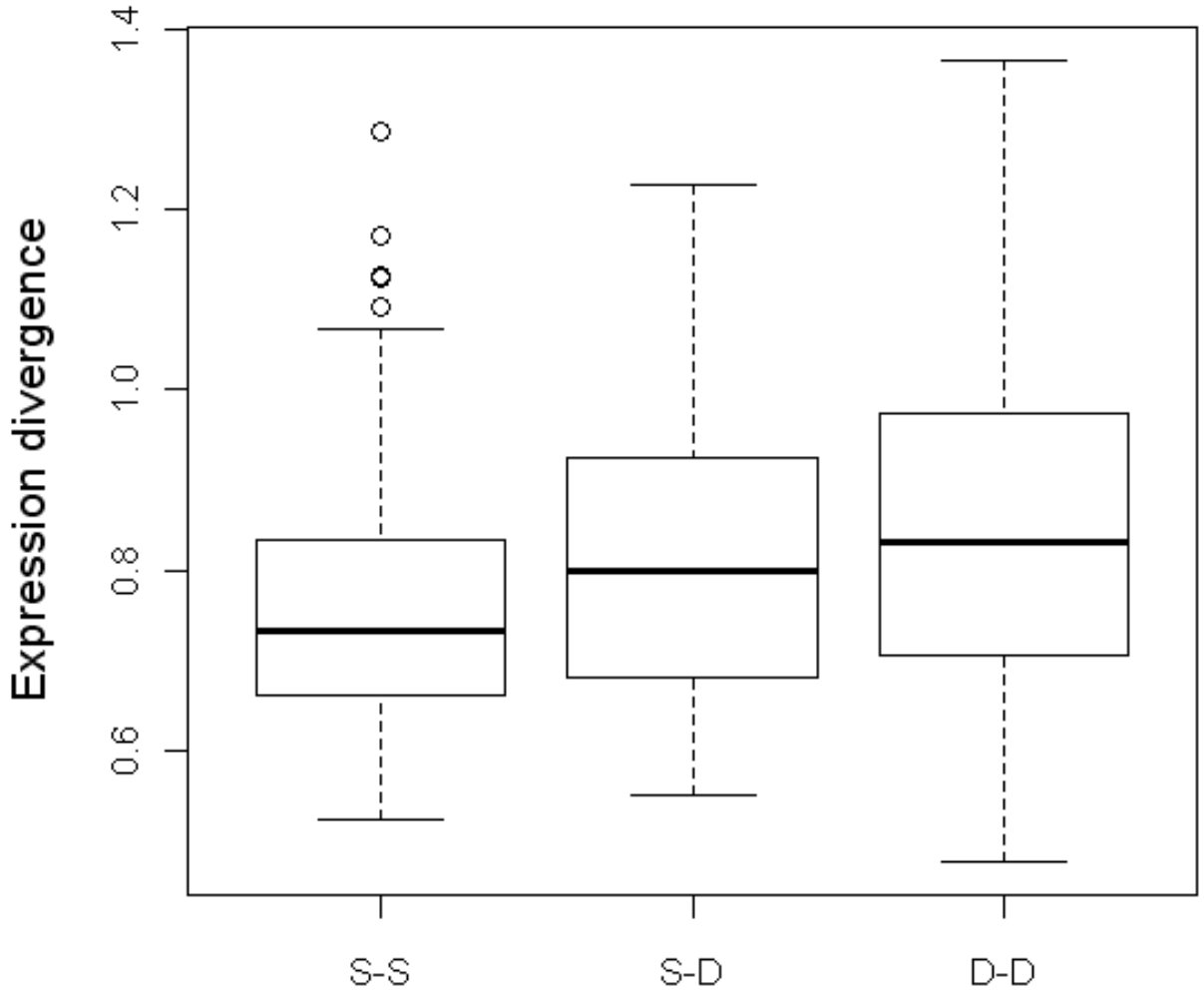

Figure 6.6. Comparison of expression divergence between different types of orthologs: singleton-singleton (S-S), singleton-duplicate (S-D) and duplicate-duplicate (D-D).

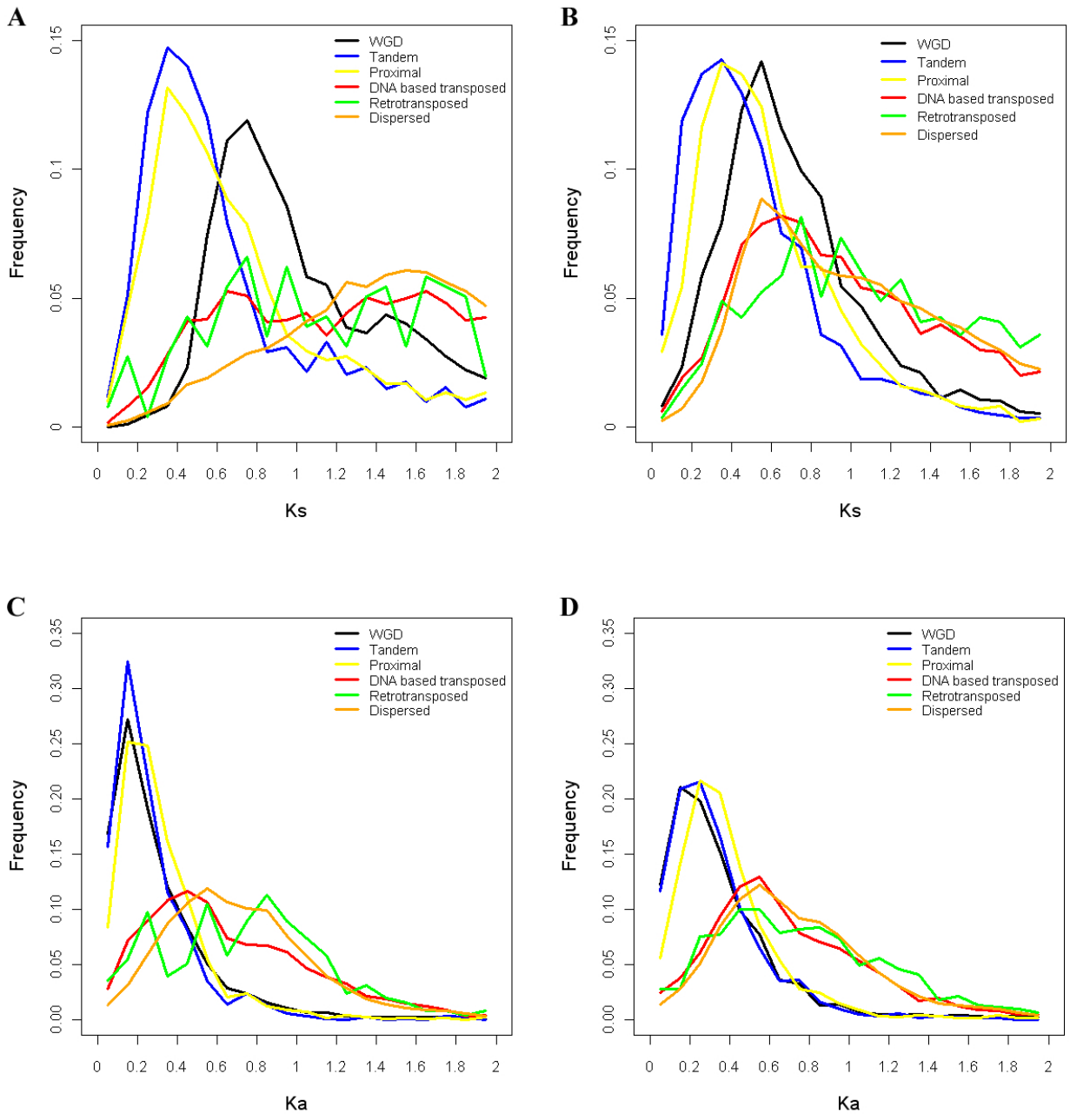

Figure 6.7. Ks and Ka distributions for gene pairs duplicated by different modes, in (A,C) *Arabidopsis* and (B,D) rice

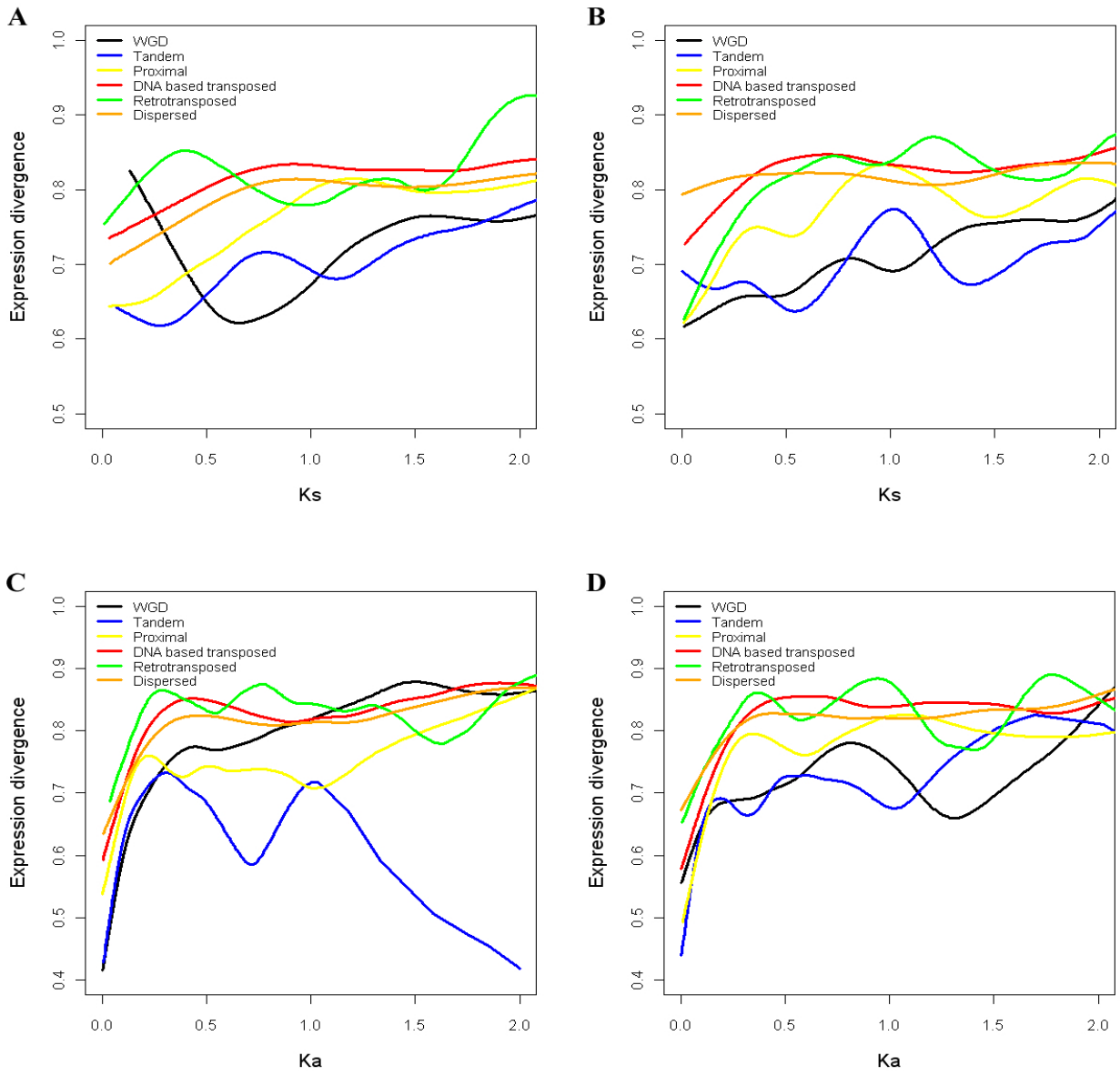

Figure 6.8. Fitted smooth spline curves between expression divergence and Ks or Ka for different modes of gene duplication. (A) Fitted smooth spline curves between expression divergence and Ks in Arabidopsis. (B) Fitted smooth spline curves between expression divergence and Ks in rice. (C) Fitted smooth spline curves between expression divergence and Ka in Arabidopsis. (D) Fitted smooth spline curves between expression divergence and Ka in rice.

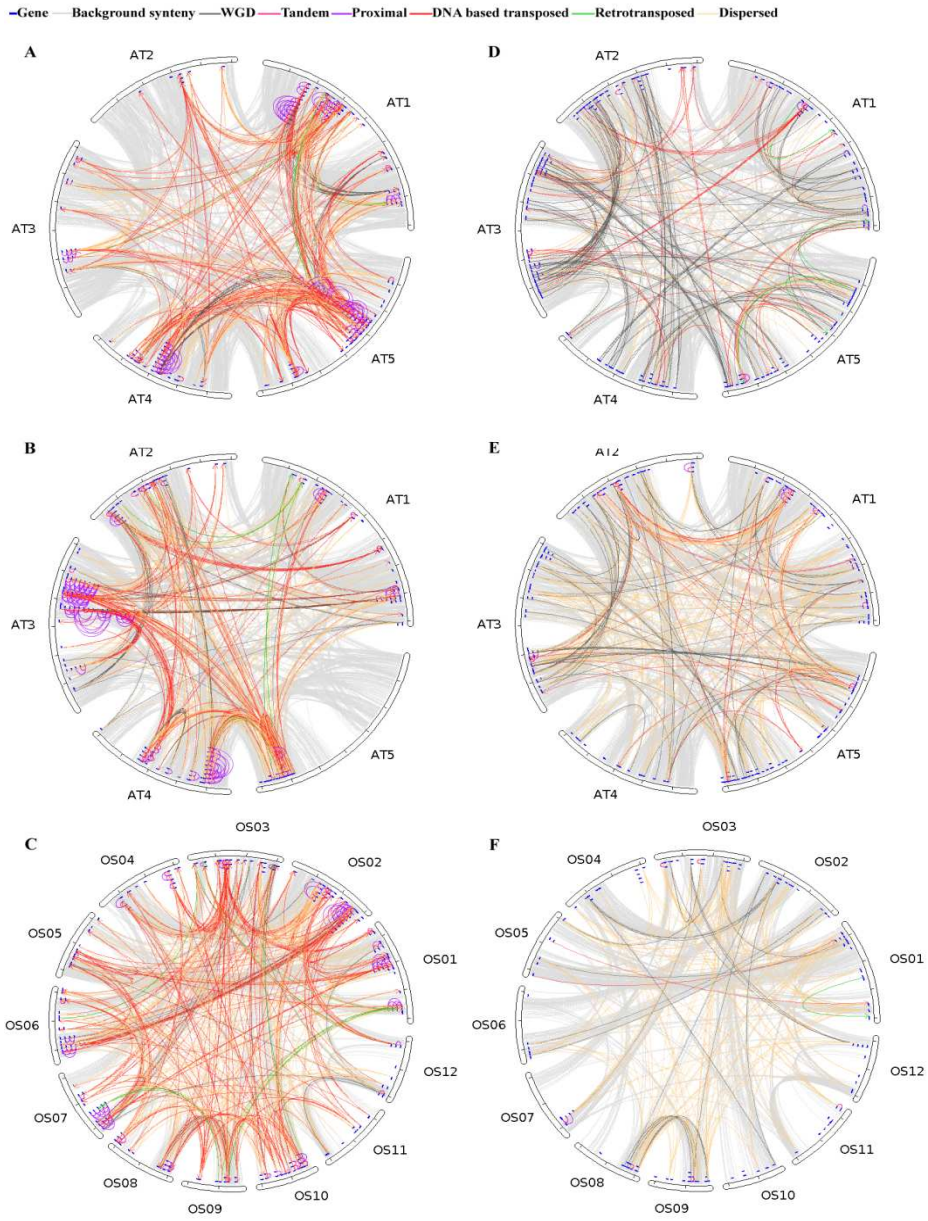

Figure 6.9. Gene duplication modes among the members of selected gene families. (A) Arabidopsis disease resistance gene homologs. (B) Arabidopsis Cytochrome P450 gene family. (C) Rice Cytochrome P450 gene family. (D) Arabidopsis cytoplasmic ribosomal gene family. (E) Arabidopsis C2H2 zinc finger gene family. (F) Rice C2H2 zinc finger gene family. Different gene duplication modes are indicated by different colors.

CHAPTER 7

CONCLUSION

Key findings

Although advances in microarray technology have provided massive high-dimensional gene expression data, the analysis of such data is still challenging. Assessing conservation/divergence of gene expression across species is important for the understanding of gene regulation evolution. However, comparing gene expression levels in different species through cross-species hybridization is often problematic. To date, computationally assessing cross-species conservation of gene expression using microarray data has been mainly based on comparison of expression patterns across corresponding tissues, or comparison of coexpression of a gene with a reference set of genes. Because direct and reliable high-throughput experimental data on conservation of gene expression are often unavailable, the assessment of these two computational models is very challenging and has not been reported yet. In this study, we compared one corresponding tissue-based method and three coexpression-based methods for assessing conservation of gene expression, in terms of their pair-wise agreements, using a frequently used human-mouse tissue expression dataset. We find that 1) the coexpression-based methods are only moderately correlated with the corresponding tissue-based methods, 2) the reliability of coexpression-based methods is affected by the size of the reference ortholog set, and 3) the corresponding tissue-based methods may lose some information for assessing conservation of gene expression. We suggest that the use of either of these two computational models to study the evolution of a gene's expression may be subject to great uncertainty, and the

investigation of changes in both gene expression patterns over corresponding tissues and coexpression of the gene with other genes is necessary.

Gene expression patterns were then compared between human and mouse genomes using two published methods. Specifically, we studied how gene expression evolution was related to GO terms and tried to decode the relationship between promoter evolution and gene expression evolution. The results showed that 1) the significant enrichment of biological processes in orthologs of expression conservation reveals functional significance of gene expression conservation. The more conserved gene expression in some biological processes than is expected in a purely neutral model reveals negative selection on gene expression. However, fast evolving genes mainly support the neutrality of gene expression evolution, and 2) gene expression conservation is positively but only slightly correlated with promoter conservation based on a motif-count score of the promoter alignment. The findings suggest a neutral model with negative selection for gene expression evolution between humans and mice, and promoter evolution could have some effects on gene expression evolution.

Comparisons between related eukaryotic genomes have revealed various degrees to which homologous genes remain syntenic and collinear during evolution. However, detection of conserved synteny and collinearity are often complicated by gene loss, tandem duplications, gene transpositions and chromosomal rearrangements. MCScan is an algorithm able to scan multiple genomes or subgenomes in order to identify putative homologous chromosomal regions, and align these regions using genes as anchors. The *MCScanX* toolkit implements an adjusted MCScan algorithm for detection of synteny and collinearity that extends the original software by incorporating 14 utility programs for visualization of results and additional downstream analyses. Applications of *MCScanX* on several sequenced plant genomes and gene families are shown as

examples. *MCScanX* can be used to effectively analyze chromosome structural changes, and reveal the history of gene family expansions that might contribute to the adaptation of lineages and taxa. An integrated view of various modes of gene duplication can supplement the traditional gene tree analysis in specific families. The source code and documentation of *MCScanX* are freely available at http://chibba.pgml.uga.edu/mcscan2/.

Both single-gene and whole-genome duplications have recurred in angiosperm evolution. We compared expression divergence between genes duplicated by WGD, tandem, proximal, DNA-based transposed, retrotransposed and dispersed duplication modes, and between positional orthologs in taxa diverged by more than 100 million years. Both neo-functionalization and genetic redundancy can result in retention of duplicate genes. WGD duplicates generally are more frequently associated with genetic redundancy than genes resulting from other duplication modes, partly due to dosage amplification. Tandem duplications also contribute to genetic redundancy, while other duplication modes are more frequently associated with evolutionary novelty. Potentially transposon mediated gene duplications tend to reduce gene expression levels. Expression divergence between duplicates is discernibly related to duplication modes, WGD events, Ka, Ks, and possibly the DNA methylation status of their promoter regions. However, the contribution of each factor is heterogeneous among duplication modes, and new factors as well as combinatorial effects of different factors are worth further investigation. Gene loss may retard inter-species expression divergence, as singletons are generally more conserved in gene expression than duplicates. Members of different gene families have non-random patterns of origin, and such patterns may be similar between *Arabidopsis* and rice.

Discussion and future directions

In Chapter 4, we aimed to address two questions related to the evolution of gene expression between humans and mice. First, is spatial expression profile of a gene constrained by natural selection? Second, do *cis*-regulatory changes associate with the expression evolution of a gene? Based on the observation that genes in some GO categories have particularly slow rate of expression evolution, we concluded that gene expression is evolutionarily constrained. Later on we showed a weak correlation between divergence of gene expression and divergence of sequence potentially involved in *cis*-regulation, implying the role *cis*-regulatory changes may have in gene expression evolution.

Several previous studies (Denver et al. 2005; Jordan et al. 2005; Khaitovich et al. 2006; Liao and Zhang 2006b; Xing et al. 2007) have rejected the hypothesis that gene expression is unaffected by natural selection. Two studies for the purely neutral model cited (Khaitovich et al. 2004; Yanai et al. 2005) have been shown to be either technically flawed or could be interpreted alternatively (Liao and Zhang 2006a). Therefore, it seems that whether gene expression profile is a selectively neutral trait is no longer a debated issue. The question remains contentious maybe the extent to which stabilizing selection limits divergence in gene expression. The first section of this chapter only applied similar approaches of former studies and provides no significant progress regarding this issue. Further, simply reporting GO categories enriched with slow-evolving genes (in expression) without a biological explanation is not insightful. From Table 4.1, most GO categories associated with expression conservation are related to male reproduction, but male reproductive proteins have been shown to be fast evolving in coding sequences due to positive selection (Torgerson et al. 2002). The genes with these GO terms are mostly tissue-specific genes exclusively expressed in testis. Tissue-specificity has been shown to be an

important determinant negatively correlated with expression profile evolution (Liao and Zhang 2006a). So whether GO terms directly affect gene expression evolution has not been well solved in this dissertation.

As for the second question about the regulatory sequence evolution, it is not sure that "motifs" used in computing "motif-count score" are regulatory elements. The motif-count score may more reflect the local mutation rate. It could be more reasonable to normalize the score using the "ratio of motif-count score to dS" to support the claim that the divergence of the potential regulatory sequence is truly correlated with the expression divergence. An alternative way is implementing experimentally confirmed *cis*-elements for this analysis.

In Chapter 6, modes of gene duplication in *Arabidopsis* and rice were classified into whole-genome, tandem, proximal, DNA-based transposed, retrotransposed and dispersed duplications. Although it is clear that genes can be duplicated by various genetic mechanisms, the classification of gene duplication modes based on measuring the physical distance between duplicate genes (the distance between collinear positions can be regarded as zero) has limitations. First, a duplicate gene may be simultaneously a WGD duplicate and a tandem duplicate. Next, since transposons may relocate duplicate genes to distant chromosomal positions, it may be true that transposons can also relocate duplicate genes to adjacent or proximal chromosomal positions. So the mechanisms underlying tandem and proximal duplications may be intermingled between unequal chromosomal crossing over and transposon activities. Further, a pair of duplicate genes may have experienced multiple genomic events. For example, two duplicates initially created by a WGD event, might experience further gene movement or transposition. For such duplicates, a simple mode of gene duplication may be inappropriate. In addition, the mechanisms underlying many dispersed duplications are not well

understood. These facts suggest that currently the associations between gene duplication modes and genetic mechanisms are generally rough, necessitating future efforts to depict a whole and clear picture of various modes and mechanisms of gene duplications.

The assumption that computationally, genetic redundancy may be inferred from simultaneous conservation in protein sequences which determine molecular functions, and expression patterns which determine biological processes, may be limited. Showing conservation of protein sequence and expression pattern may be insufficient for claiming genetic redundancy. It can be helpful to incorporate information from functional studies, e.g. phenotypic effects of knockout experiments. In addition, expression divergence may not be simply understood as neofunctionalization, because subfunctionalization can also lead to expression divergence. To exactly attribute expression divergence between duplicate genes to neofunctionalizaion or subfunctionalization, inference of their ancestral expression state and comparisons between current and ancestral expression patterns are an appropriate approach.

We found that expression divergence between duplicate genes is not well associated with the DNA methylation status of their promoter regions. However, many studies have suggested that epigenetic changes such as DNA methylation and histone modification are often associated with gene expression divergence in polyploids (Osborn et al. 2003; Adams and Wendel 2005a; Adams and Wendel 2005b; Chen 2007; Jackson and Chen 2010). In addition, a recent study suggested that transposable elements and small RNAs contributed to gene expression divergence both within and between *Arabidopsis thaliana* and *Arabidopsis lyrata* (Hollister et al. 2011). Findings to date suggest that both genetic and epigenetic factors can be related to gene expression divergence and that gene expression divergence itself might be caused by the

interaction of multiple genetic and epigenetic factors. New factors that may affect expression divergence and how different factors work together warrant further investigation.

References

Adams KL, Wendel JF. 2005a. Novel patterns of gene expression in polyploid plants. *Trends Genet* **21**(10): 539-543.

Adams KL, Wendel JF. 2005b. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**(2): 135-141.

Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol* **58**: 377-406.

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet* **37**(5): 544-548.

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A* **108**(6): 2322-2327.

Jackson S, Chen ZJ. 2010. Genomic and expression plasticity of polyploidy. *Curr Opin Plant Biol* **13**(2): 153-159.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* **345**(1): 119-126.

Khaitovich P, Enard W, Lachmann M, Paabo S. 2006. Evolution of primate gene expression. *Nat Rev Genet* **7**(9): 693-702.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol* **2**(5): E132.

Liao BY, Zhang J. 2006a. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol* **23**(6): 1119-1128.

Liao BY, Zhang JZ. 2006b. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**(3): 530-540.

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V *et al.* 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet* **19**(3): 141-147.

Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol Biol Evol* **19**(11): 1973-1980.

Xing Y, Ouyang ZQ, Kapur K, Scott MP, Wong WH. 2007. Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol Biol Evol* **24**(6): 1283-1285.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E *et al.* 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**(5): 650-659.