The Dissertation Committee for Misook Ha certifies that this is the approved version of the following dissertation

Mechanisms of Gene Expression Evolution in Polyploids

Committee:

Z. Jeffrey Chen, Supervisor

Orly Alter

Thomas Juenger

Wen-Hsiung Li

Edward Marcotte

MECHANISMS OF GENE EXPRESSION EVOLUTION IN POLYPLOIDS

by

Misook Ha, B.S., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2008

Acknowledgements

I thank my dissertation advisor, Dr. Z. Jeffrey Chen, who has taught me how to think critically and address interesting questions throughout the entire course of my graduate study. He has patiently worked with me and has been persistently encouraging and supporting me to try new ideas and methods. He is a role model and a great scientist. I will always remember his creativity and deep passion for science, high standards in research, and leadership roles in accomplishing research and educational missions.

I am grateful to my committee members, Dr. Orly Alter, Dr. Thomas Juenger, Dr. Wen-Hsiung Li and Dr. Edward Marcotte for their helpful advice and critical suggestions to improve my dissertation. I especially thank Dr. Wen-Hsiung Li for his insightful and constructive discussions with my work, which led to the publication of my first research paper. He has helped me keep focused on addressing the important questions in research.

I was very fortunate to work with many wonderful lab members. I thank Craig Dupree for his professional management of our CPU clusters, Grace Kim, Jie Lu, and Erika Lackey who have been nice colleagues as well as good friends. I thank former lab members, Drs. Jianlin Wang, Lu Tian, for their hard work and contributions to the preliminary data for my research. I am grateful to Drs. Changqing Zhang, Danny Ng, and Gyuoungju Park, Zhiguo Han, Jane Liu, David Pang for their helpful discussions to improve my work.

I thank Vanitharani Ramachandran and Dr. Xuemei Chen at the University of California Riverside for their contributions to miRNA microarray experiments as part of a collaborating project.

My special gratitude goes to my father, mother, brother, mother-in-law, and father-in-law for their endless support for my love of science.

Finally, I am indebted to my husband, Soondo Hong, and my son, Eric Hong, for their continuous support, encouragement, and love. Without them, it would be impossible for me to complete my dissertation.

MECHANISMS OF GENE EXPRESSION EVOLUTION IN POLYPLOIDS

Misook Ha, Ph.D.

The University of Texas at Austin, 2008

Supervisor: Z. Jeffrey Chen

Polyploidy, or whole genome duplication (WGD), is a fundamental evolutionary mechanism for diverse organisms including many plants and some animals. Duplicate genes from WGD are a major source of expression and functional diversity. However, the biological and evolutionary mechanisms for gene expression changes within and between species following WGD are poorly understood. Using genome-wide gene expression microarrays and high-throughput sequencing technology, I studied the genetic and evolutionary mechanisms for gene expression changes in synthetic and natural allopolyploids that are derived from hybridization between closely related species. To investigate evolutionary fate of duplicate genes, I tested how duplicate genes respond to developmental and environmental changes within species and how ancient duplicate genes contribute to gene expression diversity in resynthesized allopolyploids. We found that expression divergence between gene duplicates was significantly higher in response to environmental stress than to developmental process. Furthermore, duplicate genes

related to external stresses showed higher expression divergence between two closely related species and in resynthesized and natural allotetraploids than single-copy genes. A slow rate of expression divergence of duplicate genes during development may offer dosage-dependent selective advantage, whereas a high rate of expression divergence between gene duplicates in response to external changes may enhance adaptation.

To investigate molecular mechanisms of expression diversity among allopolyploids, I analyzed high-throughput sequencing data of small RNAs in allopolyploids and their progenitors. Small interfering RNAs (siRNAs) induce epigenetic modification and gene silencing of repeats, while microRNAs (miRNAs) and trans-acting siRNAs (ta-siRNAs) induce expression modulation of protein coding genes. Our data showed that siRNA populations in progenitors were highly maintained in allopolyploids, and alteration of miRNA abundance in allopolyploids was significantly correlated with expression changes of miRNA target genes. These results suggest that stable inheritance of parental siRNAs in allopolyploids helps maintain genome stability in response to genome duplication, whereas expression diversity of miRNAs leads to interspecies variation in gene expression, growth, and development.

Results from these research objectives show that genome-wide analysis of high throughput gene expression and small RNAs provides new insights into molecular and evolutionary mechanisms for gene expression diversity and phenotypic variation between closely related species and in the new allopolyploids.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. BACKGROUND AND INTRODUCTION

INTRODUCTION

WGD (whole genome duplication) or polyploidy is an important biological process. There are two kinds of polyploidy: autopolyploidy and allopolyploidy [1]. Autopolyploidy results from doubling of a single genome, and allopolyploidy results from the combination of distinct genomes. Recent genome sequencing of many organisms has revealed that WGD occurred in most organisms during evolution [2-9]. Following polyploidy, many duplicate genes are lost and returned to the single-copy status [4]. Differential loss of duplicate genes can induce genomic incompatibility and reproductive isolation [10]. Clearly, many duplicate genes from polyploidy are conserved. The retained duplicate genes provide dosage-dependent selective advantage and genetic robustness against deleterious mutations on essential genes [11, 12]. Alternatively, duplicate genes diverge in protein sequences or expression patterns by subfunctionalization or neofunctionalization [13]. In subfunctionalization, the functions of duplicate genes are specialized in subsets of their ancestral functions [14]. In neofunctionalization, duplicate genes gain new functions by changing expression patterns or protein sequences. Sequence and expression divergence between duplicate genes is a major source of evolutionary novelty [15]. Moreover, heterozygosity in many stable allopolyploid plants leads to permanent fixation of hybrid vigor. This implies that allopolyploidy is often an instantaneous speciation process [1].

Duplicate genes from WGD provide excellent genetic materials for studying changes in regulatory networks. First, remodeling gene expression regulatory network is

a key mechanism for phenotypic variation [16]. As more and more complete genome sequences become available, it became clear that most species contain a large number of gene duplicates or similar sets of orthologs. Moreover, among closely related species, difference in genome sequences is surprisingly small. For example, the difference in genome sequences between human and chimpanzee is less than 5% [17, 18]. This implies that morphological differences among related species are mainly due to changes in gene expression regulation and gene copy numbers. Indeed, there is evidence supporting major effects of expression and copy number differentiation on species-specific morphogenesis. In animals, spatial expression variation of *Hox* genes expression in mouse, chick and python is correlated with axial morphological differences among these species [19, 20]. In plants, expression of *FLC* at different times and different levels determine flowering time variation and reproductive isolation in allopolyploids, *Arabidopsis thaliana*, and *A. arenosa* [21].

Second, duplicate genes tend to diverge in expression regulation [22]. Especially, duplicate genes tend to change expression patterns rather than to change their biochemical functions or amino-acid sequences [23]. This expression divergence between duplicate genes expands regulatory networks within species [24]. Moreover, duplicate genes diverge in expression regulation among species [25]. Therefore, duplicate genes derived from WGD may evolve to gain species-specific expressed patterns. Moreover, successful polyploids remodel gene expression in response to multiplication of whole genome. Conservation and expression divergence of duplicate genes in polyploids can provide advantageous morphological variation in polyploids [26].

Third, WGD duplicate genes have evolved concurrently from shared ancestors. Therefore, investigating expression changes between duplicate genes within species or among species provides new insights into expression changes during evolution.

MECHANISMS OF GENE EXPRESSION DIVERGENCE BETWEEN DUPLICATE GENES

The molecular mechanisms for expression divergence between duplicate genes are poorly understood. At least four possible mechanisms are available to explain expression divergence between duplicate genes (Figure 1.1). A classical hypothesis suggests gene expression regulation mainly resulting from interactions among gene specific transcription activators and repressors and their target genes. Thus, (1) changes in *cis*-regulatory sequences of duplicate genes and/or (2) changes in target recognition of transcription factors may result in expression divergence between gene duplicates. Other mechanisms such as chromatin modifications and RNA-mediated pathways may affect the expression fate of duplicate genes. (3) Duplicate genes may diverge in expression by chromatin modifications and homology-dependent gene silencing [27]. (4) At posttranscriptional levels, duplicate genes may produce sense and antisense RNAs and promote RNA-mediated gene expression divergence [28]. This part reviews the current knowledge as well as my understanding of the four possible mechanisms of gene expression change among duplicate genes.

Change of cis-regulatory sequences

Changes in upstream transcription binding sites alter the recognition of transcription regulators or binding affinity of transcription factors. *Cis*-regulatory motifs recognized by transcription factors and transcription machineries become degenerate [29]. Promoter sequences are less constrained in sequence change than coding sequence. Thus, in the upstream region, nucleotide substitution, deletion or insertion rates are high. In yeast, Zhang *et al* [30] showed that divergence of *cis*-regulatory motifs are significantly correlated with expression difference between duplicate genes. However, difference of *cis*-regulatory motifs explains only 4% of expression divergence in yeast [30]. This suggests that there are other factors influencing expression differentiation among duplicate genes.

Evolution of duplicate transcription factors

Sequence changes in the genes encoding duplicate transcription factors may expand regulatory networks [31]. With rapid sequence divergence in DNA binding domains, duplicate transcription factors may evolve to have different target gene sets. Specifically, duplicate transcription factors may evolve to diversify their target gene sets, and divergent duplicate transcription factors have expanded gene regulatory networks in yeasts [24].

Epigenetic changes

Epigenetics refers to phenotypic or gene expression variation that is independent of changes in primary DNA sequence. Changes in chromatin structure and RNA

regulation are responsible for many epigenetic phenomena. Epigenetic states are often heritable and reversible. Changes in chromatin structure activate or repress gene expression. The resulting changes in gene expression without alteration in genomic sequences are inheritable. Chromatin structure affects gene expression through controlling accessibility of transcriptional machineries or recognition by other factors [32]. These chromatin modifications include DNA methylation, covalent modifications of histones, and siRNA generation. For example, in toadflax flowers, heritable silencing of the gene *Lcyc* induces asymmetric flower phenotypes by inheritance of DNA methylation in the locus without any mutation in the coding sequences [33].

DNA methylation occurs at cytosine residues and often in CG dinucleotide or CNG trinucleotide. DNA methylation at CG sites (CG mehtylation) is regulated by maintenance DNA METHYLTRANSFERASE 1, DNMT1 in animals and MET1 in plants [34]. CG methylation by MET1 in the promoter of FWA and FERTILIZATION INDEPENDENT SEED2 (FLS2) induces gene silencing during male gametogenesis and endosperm development [35]. During female gametogenesis, FWA and FLS2 are expressed because DEMETER (DME) is expressed, which inhibits MET1 activity [36]. The DOMAINS REARRANGED METHYLTRANSFERASE 1 and 2 (DRM1 and 2) are considered to be *de novo* DNA methylases in *Arabidopsis thaliana*. DRM1 and 2 have been shown to be essential for the establishment of DNA methylation and silencing of FWA and SUPERMAN (SUP) [37]. In plants, DNA methylation is also detected at CNG and CHH sites (non-CG methylation, H = A, C, or T not G). DRM1/2 and CHROMOMETHYLASE3 (CMT3) is involved in non-CG methylation [38]. Recent

genome-wide analysis of DNA methylation in *Arabidopsis thaliana* show that DNA methylation occupies mainly in heterochromatic regions enriched with repeat elements [39]. About 2,000 coding genes were also methylated in the genic regions, and the majority of them are located adjacent to transposable elements [32]. Many transposable elements in heterochromaitic regions escape transcriptional silencing and are transcribed in *met1* and *ddm1* mutants. In *drm1drm2cmt3* mutants, up-regulation of protein coding genes as well as transcription of transposons were observed [40]. These studies suggest that CG methylation is involved in silencing of repeats, whereas non-CG methylation is involved in down-regulation of protein-coding genes, in addition to silencing repeats.

Post-translational modifications of histones include acetylation, methylation, phosphorylation, ubiquitylation and SUMOylation at lysine residues (K) of core histones including histone 3 (H3) or histone 4 (H4). It has been shown that H3K27me3 is associated with repression of several genes such as MEA [41, 42] and PHE1 [43]. Genome-wide profiling of H3K27me3 in *Arabidopsis thaliana* demonstrated that H3K27me3 is enriched in euchromatic region and the targets that are usually expressed at low levels. H3K27me3 is also an important gene repressor for the normal development of mammals and *Drosophila*. In animals, the Polycomb-group (PcG) protein complexes PhoRC, PRC1, while PRC2 are required to establish and maintain H3K27me3 [44, 45]. In Drosophila, PRC2 catalyze H3K27me3 and PRC1 is involved in spreading H3K27me3 bidirectionally [46]. However, Pc homolog was not found in plants. H3K27me3 co-localizes with the only Arabidopsis Heterochromatin Protein 1 LHP1/TFL2 [47]. This suggests that even though different orthologs interact with H3K27me3, their action

mechanisms are very similar among plants and other animal species. H3K9 methylation is also a repressive marker. H3K9me2 is mainly localized in heterochromatin. H3K9me3 is mainly localized in euchromatin, promoters and coding sequences [47].

Histone acetylation occurs in various sites including H3K9, H3K14, H3K18, H3K56, H4K5, H4K8, H4K13, and H4K16 residues. The acetylation of conserved lysine residues neutralizes the positive charge of the histone tails and decreases their affinity for negatively charged DNA, thereby promoting the accessibility of chromatin to transcriptional regulators [48]. Alternatively, the "histone code" hypothesis proposes that the combination of different covalent modification states of lysine and/or arginine residues on histone tails, including histone acetylation and methylation, provide signals for recruitment of specific chromatin-associated proteins, which in turn alter chromatin states and affect transcriptional regulation [49]. It may be too simple to divide hisone modifications into active or repressive. Accumulating data have resulted in conflicting observations of the binary viewpoint. This may be related to fast turnover of histone modifications and specific interactions of histone modifications with other protein factors. Nonetheless, histone modifications are important to gene expression, from constitutive repression to subtle and dynamic changes. For example, H3K9methylation at the *GLABRA2* (*GL2*) promoter promotes cell-cycle specific expression of *GL2* [50, 51]. Histone modifications regulate gene expression in response to environmental cues. Histone deacetylase (HDAC) HDA19 regulates pathogene-responsive genes promoting pathogene resistance in Arabidopsis [52]. Epigenetic changes are responsible for gene expression in developmental process and cell differentiation. Specifically, many histone

modifications change dynamically in response to developmental singals and environmental cues. Differentiating chromatin modifications of duplicate genes, epigenetic changes may cause expression divergence between duplicate genes [53-55]. However, it is not very well known how epigenetic modifications mediate gene expression changes during developmental process. Moreover, little is known about how chromatin modifications are responsible for expression variation between duplicate genes and how chromatin modifications affect expression divergence between closely related species. Among *Arabidopsis thaliana* ecotypes, the DNA methylation landscape is very conserved [56, 57], DNA methylation is probably not a significant factor for expression changes among different ecotypes in *A. thaliana* [57]. However, histone modifications are dynamic and may be related with locus-specific expression patterns between closely related species. For example, in *A. thaliana* and *A. arenosa*, higher expression of *A. arenosa FLC* is associated with more H3K4 dimethylation and H3K9 acetylation but less H3K9 dimethylation than *A. thaliana FLC* [21].

Roles of small RNA

Plants have various small RNA pathways including (1) small interfering RNAs (siRNAs), (2) microRNA (miRNA), (3) trans-acting siRNA (ta-siRNA), (4) natural antisense-derived siRNA (NAT-siRNA), and (5) heterochromatin-associated siRNAs. miRNA, ta-siRNA and NAT-siRNA mainly target protein coding genes and are directly associated with expression of protein-coding genes, while siRNAs mainly target non-coding genes or repeat elements and are involved in the maintenance of genome stability.

What mechanisms might be responsible for recognizing genomic sequences and altering chromatin modifications? There is an increasing amount of data to support that siRNA can confer sequence specificity for epigenetic modifications. Expression of a double-stranded RNA (dsRNA) containing promoter sequences resulted in transcriptional gene silencing accompanied by *de novo* methylation of the target promoter and siRNA generation in plants [58]. Only the DNA sequences complementary to the guide RNA were methylated, suggesting RNA-directed DNA methylation. siRNAs direct DRM2 to catalyze DNA methylation *in de novo* in siRNA complementary genomic region in *A. thaliana* [37, 38]. The generation of siRNAs involved in RNA-directed DNA methylation requires RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) and the plant-specific protein NUCLEAR RNA POLYMERASE IV (also called NUCLEAR RNA POLYMERASE D 1a, NRPD1A) [59-62]. RDR2 and NRPD1A mediate the production of dsRNAs. DCL3 cuts the dsRNAs into ~24 nucleotide siRNAs. siRNAs are loaded into AGO4 that forms a posttranscriptional gene silencing complex (PTGS). The siRNA-loaded AGO4 complex is localized in the nucleus with DRM2 to guide RNA-directed DNA methylation. siRNAs turn off expression of target genes by guiding cleavage of complementary mRNAs as well as DNA methylation. siRNAs diced by DCL2, DCL3, or DCL4 are loaded into ARGONAUTE1 (AGO1) after methylated at the 3' ends by HUA ENHANCER1 (HEN1) [63, 64]. siRNA-loaded AGO1 complexes recognize target mRNAs by complementary base pairing and degrade target mRNAs by RNase III activity of AGO1 [65, 66]. Therefore, siRNAs play a role in maintaining silencing of repeat elements, exogenous dsRNAs, and some genes.

MicroRNAs (miRNAs) are a class of small RNAs that serve as posttranscriptional negative regulators of gene expression in plants and animals. The 20–24 nucleotide single-stranded miRNAs repress the target genes by mRNA degradation or translational repression [67-69]. *MIRNA* genes are transcribed by RNA polymerase II from intergenic and/or coding sequences [70] that are independent of their target genes, generating primary miRNA (pri-miRNA) that is processed by nuclear RNaseIII-like enzymes, such as Dicer and Drosha in animals [71] and DICER-LIKE proteins (e.g., DCL1) in plants [72]. The resulting pre-miRNA contains miRNA:miRNA* intermediate duplex formed by self-complementary foldback structure. HEN1 methylates the 2′ hydroxy of the 3′ terminal nucleotide of plant miRNAs, which protects the 3′ end from uridylation and degradation [73]. The pre-miRNAs are transported into cytoplasm by HASTY (HST), a homolog of Exportin-5 that is involved in transport of pre-miRNAs and tRNAs in animals [74]. HYL1, a double-stranded RNA binding protein, is also required for miRNA accumulation [75, 76]. The double-stranded miRNA that is unwound by a helicase-like enzyme, and the miRNA strand whose 5′-end is less tightly paired are usually incorporated into the effector RNA-induced silencing complex (RISC) [77]. One or more ARGONAUTE proteins such as AGO1 in the active miRNA-containing RISC complex help guide the targets by complementary sequences [65]. As a result, most plant miRNAs function as negative regulators to guide miRNAs to mRNA targets for degradation. Furthermore, some miRNAs may play a role in chromatin modifications and gene transcription. The genes encoding two *Arabidopsis* miRNA targets (*PHABULOSA* and *PHAVOLUTA*) are heavily methylated downstream of the miRNA complementary sites,

and the methylation is reduced in *phb-1d* and *phv-1d* mutants [78]. Although the cause of reduced DNA methylation is unknown, the data suggest a link between miRNAs and transcriptional regulation.

Since the first miRNA (lin-4) was discovered in *Caenorhabditis elegans* [79], thousands of miRNAs have been identified in plants and animals. Estimates indicate that 1–5% of the transcribed genes in animals contain miRNAs, making them one of most abundant and dynamic classes of genetic regulators [80]. The collection of miRNAs is growing in the miRNA Registry (http://microrna.sanger.ac.uk/sequences/). Release 10.1 (December 2007) listed 5,395 miRNA locus entries, including 564 in human, 461 in mouse, 193 in zebra fish, 137 in nematode, 147 in fruit fly, 199 in Arabidopsis, 215 in poplar, 243 in rice, 96 in maize, 263 in moss, and 72 in alga. In plants many miRNAs have relatively few targets because target recognition requires near-perfect complementarity, whereas target recognition for the animal miRNAs requires a relatively low level of sequence complementarity [81]. Complementarity to the core region (positions 1–10) of miRNA is often sufficient for effective regulation in animals. Therefore, one miRNA can affect transcript and protein levels of hundreds of targets in animals [82]. Consequently, these miRNAs control a wide range of physiological and developmental processes in animals and in plants as well. In plants, miRNAs mediate leaf development including radial patterning in shoots, organ identity and flowering [69, 83]. MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs, phytohormone signaling

[84, 85] and responses to biotic and abiotic stresses [86] that are also mediated by natural cis-antisense siRNAs (nat-siRNA) [87] and 30–40-nt small RNAs.

Some miRNAs target the production of phased small RNAs as ta-siRNAs. From the cleavage site of miRNA targets, phased siRNAs of 21~24 nucleotide in length are generated by SGS3, RDR6, and DCL4 [88]. By detecting phased processing small RNAs from miRNA target transcripts, eight ta-siRNA loci (TAS1a, TAS1b, TAS1c, TAS2, TAS3a, TAS3b, TAS3c, and TAS4) have been identified in Arabidopsis. Ta-siRNAs are derived from a ta-siRNA locus and target other mRNAs in the same or related gene family. Ta-siRNAs are involved in gene expression regulation and development. miR390 induces first cleavage of TAS3 primary transcripts. The cleaved TAS3 primary transcripts are processed into dsRNA by RDR6 and diced by DCL4 into 22-24 nucleotide ta-siRNAs. TAS3 ta-siRNAs are loaded in AGO7, which induces down-regulation of AUXIN REPONSE FACTORs such as ARF3 and ARF4 that pay a role in timing and patterning of meristemtic cells and development in Arabidopsis [89].

NAT-siRNAs are produced from natural anti-sense transcripts. A representative example of NAT-siRNAs is derived from the gene encoding Δ^1-pyrroline-5-carboxylate dehydrogenase (P5CDH) and *SRO5*, a gene of unknown function. siRNAs (24nt) are generated from the pair of NATs in response to salt stress, leading to the expression of *SRO5* but silencing of *P5CDH* transcribed in the opposite strand [87]. Although the role of miRNAs in animal and plant development has been extensively studied, little is known about how miRNAs are conserved between species, how the conserved miRNAs play similar or different roles in distinct and related species, how spatial and temporal

regulations of conserved miRNAs change among the related species, and whether the expression patterns of miRNAs and their targets in the progenitors are also maintained in the interspecific hybrids and new allopolyploid species that are derived from two or more divergent species. There is evidence for sequence variation and expression divergence of miRNAs between species as well as co-evolution of miRNA loci and their targets within species. Ason et al. indicated that the timing and location of miRNA expression is not strictly conserved [90]. Several conserved miRNAs such as miR-454a, miR-145, and miR-205 clearly displayed spatial expression differences between two closely related species, medaka and zebrafish. It is conceivable that the spatial and temporal regulation of conserved miRNAs may also play an important role in shaping developmental and physiological changes during animal evolution [91].

In plants, many conserved miRNAs are expressed in diverse species. In a study using miRNA microarrays and RNA blot analysis, Axtell et al. (2005) found that out of 23 miRNAs examined, 19 were expressed in A. thaliana rosette leaves, 13 were expressed in tobacco leaves, 12 were expressed in wheat germ lysate, 13 accumulated in rice seedlings, 13 accumulated in magnolia leaves, 11 accumulated in pine leaves, eight were detected in fern leaves and stems, three were expressed in lycopod leaves and stems, and two were expressed in moss leaf gametophytes [92]. Even the most conserved miRNAs, such as miR160 and miR390, exhibited expression differences between species. Expression of miR390 was not detected in lycopod, pine, or tobacco, but it was expressed in moss. These expression variations may suggest that miR390 expression was lost in some lineages during evolution. Alternatively, expression levels of miR390 could

be below the detection level in some species. The expression variation of conserved miRNAs in plants and animals may be underestimated because of several reasons. RNA blot analysis and miRNA microarrays using pooled tissues may not detect real-time changes in cell types, tissues, or organs. Moreover, developmental variation may exist among different species. Alternative techniques such as miRNA in situ hybridization may reveal subtle changes in spatial and temporal expression among different organs and between different species.

In summary, small RNAs play big roles in gene expression and development in closely related species and polyploids. Changes in miRNA and ta-siRNA accumulation may differentiate target gene expression within and between species [81, 93], and leading to developmental variation. Repetitive DNA sequences and transposons evolve rapidly among different species, which may coincide with diverse siRNA populations between closely related species.

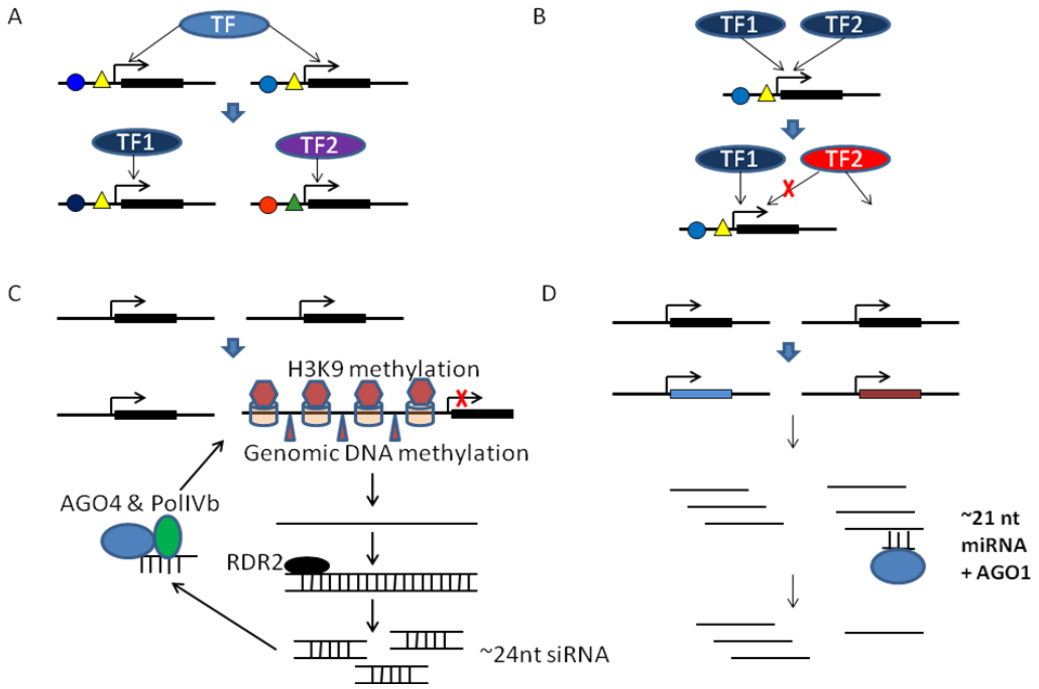

Figure 1.1 Possible mechanisms of gene expression divergence between duplicate genes

A. Changes in *cis*-regulatory elements. Sequence variation in regulatory elements can differentiate transcription factor affinities or change interacting transcription factors. **B.** Changes in target specificities of duplicate transcription factors. Changes in protein sequence of duplicated transcription factors differentiate target genes. **C.** Epigenetic changes and chromatin remodeling. siRNAs, DNA methylation and histone modification such as H3K9 methylation repress gene expression. Aberrant double-stranded RNAs are generated by the RNA-dependent RNA polymerase 2 (RDR2) from heterochromatin regions, processed into ~24-nt siRNAs by Dicer-Like (DCL) proteins, and loaded into AGO4. siRNAs in AGO4 guide cleavage of complementary RNAs. PolIVb is involved in siRNA production either transcribing the genomic DNA to produce single-stranded RNAs or transcribing double-stranded RNAs to amplify the single-stranded RNA. **D.** Differential accumulation of miRNA. Mature miRNAs (~21-nt) are loaded into AGO1 protein and induce cleavage or translation inhibition of target mRNAs. Differential accumulation of miRNA changes target gene expression levels in a posttranscriptional manner.

ARABIDOPSIS POLYPLOIDS AS MODEL SYSTEMS

Plants provide a model system for studying mechanisms of gene expression evolution in polyploids. First, the *Arabidopsis thaliana* genome is completely sequenced. Second, whole genome duplication events in *A. thaliana* have been well characterized (Figure 1.2B). At least three rounds of WGD occurred in the evolutionary history of *A. thaliana* (Figure 1.2A) [2, 3]. Comprehensive analysis of the *A. thaliana* genome sequence showed that the most recent whole genome duplication occurred 25-40 million years ago (Mya) and homologous genomic regions from the 3 rounds of whole genome duplication cover 70 to 90% of the present genome of *A. thaliana*. Third, genome-wide measurements of gene expression in various conditions allows investigation of expression divergence between duplicate genes from WGD [94]. Fourth, *Arabidopsis arenosa* and *A. thaliana* are excellent systems to analyze gene expression changes and speciation after WGD. *A. arenosa* and *A. thaliana* split ~6 MYA after whole genome duplication ~20 MYA [95] (Figure 1.2C). These two species share >90% of nucleotide sequence identity in coding regions, and >90% of the ~26,000 70-mer *A. thaliana* oligos cross-hybridize with *A. arenosa* genes [96]. Finally, new synthetic allopolyploids between *A. thaliana* and *A. arenosa* have been generated [97]. Therefore, they are suitable for investigating gene expression differentiation associated with polyploidy (Figure 1.2D).

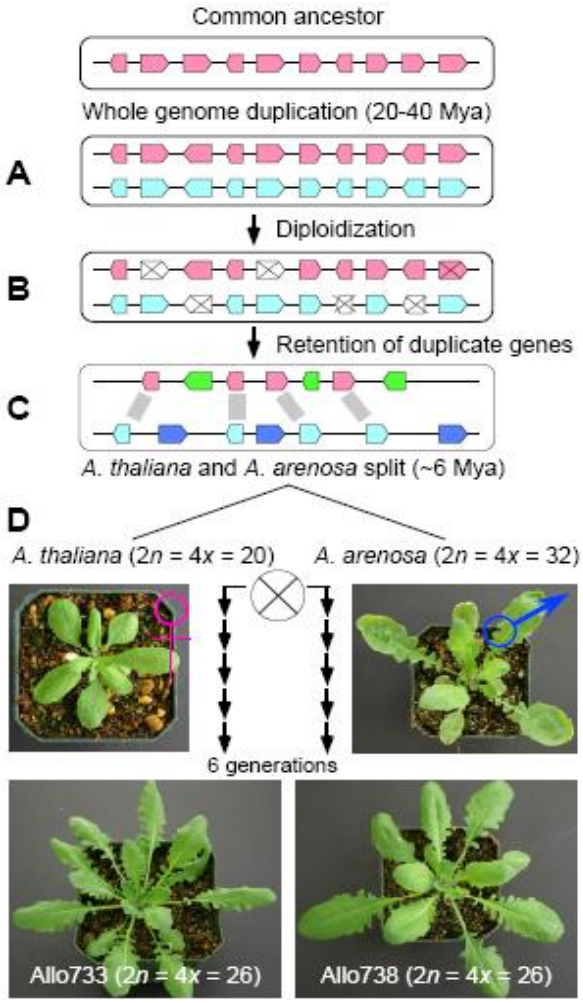

Figure 1.2 Model systems for testing the effects of polyploidy on expression evolution of duplicate genes

2. GENE EXPRESSION EVOLUTION OF DUPLICATE GENES

The majority of this chapter was previously published as one paper:

Misook Ha, Wen-Hsiung Li and Z. Jeffrey Chen. ***Trends in Genetics*** **2007 23:162-166.**

BACKGROUND AND RATIONALE

The genomes of all eukaryotes including yeast, many plants and some animals underwent at least one round of whole genome duplication during their evolutionary history [98, 99]. Duplicate genomes may undergo massive gene loss and genomic rearrangements, leading to a diploidized state, as shown in yeast [100], *Arabidopsis* [101], and rice [64]. Theoretical prediction suggests that one copy of the duplicate genes usually becomes lost by accumulation of deleterious mutations over an evolutionary time-scale [13]. Evidently, many duplicate genes have been retained during evolution as the redundancy conferred by duplicate genes may facilitate species adaptation [98] and genetic robustness [11] against changes in environmental conditions and developmental programs. Both copies may retain if dosage effects are advantageous or one gene duplicate may evolve to possess a novel function via neofunctionalization [102]. Alternatively, both copies retain a different subset of ancestral genes by differential accumulation of mutations via subfunctionalization, leading to the origin of new functional genes. Genome-wide gene expression analyses indicate that duplicate genes offer genetic robustness against null mutations in yeast [11] and tend to cause expression divergence during development and to evolve faster than single-copy genes between *Drosophila* species and within yeast species [103]. However, the hypothesis that duplicate genes have an advantage over single-copy genes in response to adaptive evolution has not been rigorously tested. It is unclear how duplicate genes respond differently to endogenous developmental switches and external

environmental changes during evolution. We employ *Arabidopsis* as a model system for hypothesis-testing because almost every flowering plant went at least one round of whole genome duplication, and over 70% flowering plants are of polyploid origin [104]. Moreover, plants constantly respond to changes in growth environments during development. The *Arabidopsis* genome consists of 60-89% of duplicate segments from at least two "synchronized" events of entire chromosomal (and/or segmental) duplication [101, 105, 106]. We tested a series of hypotheses concerning the expression and sequence divergence of duplicate genes using 2,055 recent duplicate gene pairs and 512 microarray datasets.

MATERIALS AND METHODS

Duplicate genes.

The sequence data and annotation were obtained from the TIGR database (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/) and NCBI (ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana/). We mainly used the well-characterized duplicate genes that arose from the most recent WGD event, which was estimated to be 20-40 million years ago [105, 106]. Older duplicate genes were used only in a few analyses because their duplication dates are unknown. We excluded the genes that were annotated as pseudogenes or had no detectable expression in the experiments. In addition, 544 genes were excluded because more than one gene was assigned to a single array element. According to the analysis of WGD events by Blanc et al. (2003), we detected expression

of 2,055 recent gene duplicates and 1,131 old gene duplicates in a total of 21,298 annotated genes. We also analyzed the expression data of 2,573 recent duplicate gene pairs inferred by Bowers et al. (2003), which included 1,798 duplicate gene pairs that matched both copies and 277 pairs that matched one of the gene duplicate inferred by Blanc et al. (2003) and additional 498 duplicate gene pairs [107]. The two sets of duplicate genes were qualitatively similar [107], and the results of our analyses using the duplicate gene dataset of Blanc et al. (2003) were consistent with those for the dataset of Bowers et al. (2003).

Expression data.

We obtained the Affymetrix ATH1 expression array data from the AtGenExpress expression atlas at TAIR (http://www.arabidopsis.org/info/expression/ATGenExpress.jsp). We compiled 512 microarray datasets for various conditions, including 79 datasets for 79 different developmental stages or organs and 250 datasets for 37 abiotic and biotic stress treatments. We classified the 512 microarray datasets [108] into three groups: environmental (abiotic and biotic) factors, developmental programs, and others (Table 2.1). The first group included environmental factors such as abiotic and biotic stresses, and the second group included developmental signals such as organ differentiation, developmental switches from vegetative to reproductive growth. In the second group, we used expression profiles in the wild-type plants but excluded 16 microarray datasets obtained in the mutants. The data in the third group were not used in the analysis because it consisted of various experimental conditions such as treatment of various chemicals,

hormones, and mutant types, which may include both environmental and developmental factors. We tested gene expression divergence affected by developmental and environmental factors separately, using expression data obtained from 63 different developmental stages or tissue-types and 63 sets of treatment and time-course combinations under abiotic or biotic stress.

We obtained expression estimates using the GC-RMA method [109]. The individual values were used for t-test, and the average values of replicated experiments (triplicates in developmental stages and biotic stresses and duplicates in abiotic stresses) were used for correlation coefficient tests. For comparison, Affymetrix detection algorithms in the MAS5 library implemented in R [110] were used to normalize the data and to estimate expression values, and the background levels and PM/MM ratios were corrected according to the Affymetrix Statistical Algorithms [111]. There was no significant difference in the overall results obtained using the two data normalization methods.

Measurement of expression similarity.

Similar results were obtained using both Pearson's correlation coefficient and Spearman's rank correlation coefficient, and the former results are presented in this study. The Pearson's correlation coefficient (R_{ik}) between gene i and gene k was calculated as

$$R_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k})}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \overline{x_k})^2}}$$

where x_{ji} is the expression value of gene i under condition j; $\overline{x_i}$ is the mean expression value of gene i; s_{ii} is a standard deviation of gene i expression across the conditions (1…n) used in the analysis. R measures the strength of the linear association.

Detection of up-regulated genes.

In each test, the expression data consist of one control (no treatment or a specific tissue) and a series of expression data after the treatments (different stresses or tissues). We used the t-test to determine if the expression of a gene after the treatment (Ga) is greater than that before the treatment (Gb). The null hypothesis was $H_0 = Ga - Gb \leq 0$. A gene is considered to be up-regulated if H_0 is rejected ($P \leq 0.01$) in at least one of the several treatments. The up-regulated genes were used in the statistical tests because the expression patterns of these genes correspond to various stress responses [112, 113].

Biological classification using Gene Ontology (GO).

The GO for *A. thaliana* was downloaded from TAIR (http://www.arabidopsis.org, released on December 10, 2005) and assigned using published experimental data and/or electronic annotations using INTERPRO [114]. GoSlim was used to classify 14 biological process functional categories. To ensure the accuracy of GO classification, only the GOslim terms with experimental evidence were used for analysis. The evidence

codes are IDA (Inferred from direct assay), IEP (Inferred from expression pattern), IGI (Inferred from genetic interaction), IMP (inferred from mutant phenotype), IPI (Inferred from physical interaction), IEA (inferred from electronic annotation) and ISS (inferred from sequence or structural similarity). A duplicate gene pair was assigned to a GoSlim biological classification if one or both copies of the gene duplicates were annotated. The responses to abiotic, biotic and other (e.g., lights and chemicals) stresses were combined into one category, namely, response to external stresses because the GO terms included in the "response to other stress" and "response to abiotic and biotic stimuli" overlap considerably and both correspond to external stresses. The "transport" was divided into two groups, extracellular (into or out of a cell) and intracellular transport, corresponding to the external and internal processes, respectively. Among the 2,055 gene duplicate pairs, 823 were assigned using GoSlim biological process functional classifications.

Table 2.1 List of ATH1 microarray datasets

Environmental factors (abiotic stress, 2 replicates)

Treatments	Expression measurements	TAIR accession number
Cold stress	Root, Shoot, 0, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00325
Genotixic stress	Root, Shoot, 0, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00326
Osmotic stress	Root, Shoot, 0, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00327
Salt stress	Root, Shoot, 0, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00328
UV-B	Root, Shoot, 0, 0.25h, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00329
Wound	Root, Shoot, 0, 0.25h, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00330
Droought stress	Root, Shoot, 0, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00338
Heat	Root, Shoot, 0, 0.25h, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00339
Oxidative stress	Root, Shoot, 0, 0.5h, 1h, 3h, 6h, 12h, 24h after treatment	ME00340

Environmental factors (biotic stress, 2 replicates)

Treatments	Expression measurement	TAIR accession number
Botrytis cinerea infection	Leaf	ME00341
Pseudomonas syringae ES4326 infection	Leaf	ME00353
Pseudomonas syringae pv. Tomato DC3000	Leaf	ME00331
HrpZ	Leaf	ME00332
GST-NPP1	Leaf	ME00332
Flg22	Leaf	ME00332
LPS	Leaf	ME00332

Treatments	Expression measurement	TAIR accession number
Phytophthora infestans infection	Leaf	ME00342
Erysiphe orontii infection	Leaf	ME00354

Internal factors (developmental stages, 3 replicates)

Tissue	Expression measurements (age)	TAIR accession number
Cotyledons	7 days	ME00319
Hypocotyl	7 days	ME00319
Roots	7 days	ME00319
Shoot apex, vegetative + young leaves	7 days	ME00319
Leaves 1 + 2	7 days	ME00319
Shoot apex, vegetative	7 days	ME00319
Seedling, green parts	7 days	ME00319
Shoot apex, transition (before bolting)	14 days	ME00319
Roots	17 days	ME00319
Rosette leaf #4, 1 cm long	10 days	ME00319
Rosette leaf # 2	17 days	ME00319
Rosette leaf # 4	17 days	ME00319
Rosette leaf # 6	17 days	ME00319
Rosette leaf # 8	17 days	ME00319
Rosette leaf # 10	17 days	ME00319
Rosette leaf # 12	17 days	ME00319
Leaf 7, petiole	17 days	ME00319
Leaf 7, proximal half	17 days	ME00319
Leaf 7, distal half	17 days	ME00319
Developmental drift, entire rosette after transition to flowering, but before bolting	21 days	ME00319
The same as the above	22 days	ME00319
The same as the above	23 days	ME00319

Tissue	Expression measurements (age)	TAIR accession number
Senescing leaves	35 days	ME00319
Cauline leaves	21+ days	ME00319
Stem, 2nd internode	21+ days	ME00319
1st node	21+ days	ME00319
Shoot apex, inflorescence (after bolting)	21 days	ME00319
Flowers stage 9	21+ days	ME00319
Flowers stage 10/11	21+ days	ME00319
Flower stage 12	21+ days	ME00319
Flowers stage 12, sepals	21+ days	ME00319
Flowers stage 12, petals	21+ days	ME00319
Flowers stage 12, stamens	21+ days	ME00319
Flowers stage 12, carpels	21+ days	ME00319
Flowers stage 15	21+ days	ME00319
Stage 15, pedicels	21+ days	ME00319
Flowers stage 15, sepals	21+ days	ME00319
Flowers stage 15, petals	21+ days	ME00319
Flowers stage 15, stamen	21+ days	ME00319
Stage 15, carpels	21+ days	ME00319
Siliques, w/ seeds stage 3; mid globular to early heart embryos	8 wk	ME00319
Siliques, w/ seeds stage 4; early to late heart embryos	8 wk	ME00319
Siliques, w/ seeds stage 5	8 wk	ME00319
Seeds, stage 6, w/o siliques; mid to late torpedo embryos	8 wk	ME00319
Seeds, stage 7, w/o siliques; late torpedo to early walking-stick embryos	8 wk	ME00319

Tissue	Expression measurements (age)	TAIR accession number
Seeds, stage 8, w/o siliques; walking-stick to early curled cotyledons embryos	8 wk	ME00319
Seeds, stage 9, w/o siliques; curled cotyledons to early green cotyledons embryos	8 wk	ME00319
Seeds, stage 10, w/o siliques; green cotyledons embryos	8 wk	ME00319
Vegetative rosette	7 days	ME00319
Vegetative rosette	14 days	ME00319
Vegetative rosette	21 days	ME00319
Leaf	15 days	ME00319
Flower	28 days	ME00319
Root	15 days	ME00319
Root, 1x MS agar	8 days	ME00319
Root, 1x MS agar, 1% sucrose	8 days	ME00319
Seedling, green parts, 1x MS agar	8 days	ME00319
Seedling, green parts, 1x MS agar, 1% sucrose	8 days	**ME00319**
Root, 1x MS agar	21 days	**ME00319**
Root, 1x MS agar, 1% sucrose	21 days	**ME00319**
Seedling, green parts, 1x MS agar, 1% sucrose	21 days	**ME00319**
Seedling, green parts, 1x MS agar	21 days	**ME00319**
Pollen	Mature pollen	**ME00319**

RESULTS

Preferential induction of duplicate genes by abiotic and biotic stresses

To investigate the expression evolution of duplicate genes in response to external factors (Table 2.1), we first studied how often these duplicate genes are induced by environmental stresses using microarray data analysis. The proportion of duplicate genes up-regulated under abiotic stress in roots or shoots is significantly higher than that of other genes in the genome (one-tailed t-test, $P \leq 0.01$) (Table 2.2). We obtained the same conclusion for duplicate genes in response to biotic stress induced by pathogen infections or pathogenic molecules (Table 2.2). The data suggest that duplicate genes are preferentially involved in stress responses.

Expression diversity in response to developmental changes

We next studied how duplicate genes respond to developmental processes. The differentially regulated genes were detected across 79 different tissues using one-way ANOVA. We found a higher frequency of gene duplicates than the other genes in the genome displaying differential expression in various developmental stages. Among five representative tissues (leaf, flower, root, seed, and pollen), the proportion of duplicate genes that were differentially expressed was significantly higher than that of the other genes in the genome. The data suggest that duplicate genes increase expression diversity during development, similar to the findings in *Drosophila* and yeast [103].

Table 2.2 Number of duplicate genes that were up-regulated in response to abiotic and biotic stresses

Stimulus	Observed number (percentage)	Expected number [a] (percentage)	P-value
Abiotic stress			
Cold, Root	343 (8.48%)	223 (5.4%)	1.13E-17
Cold, Shoot	397 (9.82%)	247 (6%)	1.60E-24
Genotoxic, Root	231 (5.71%)	177 (4.3%)	1.31E-05
Genotoxic, Shoot	330 (8.2%)	266 (6.4%)	1.33E-05
Osmotic stress, Root	405 %10.0%)	447 (10.9%)	0.071
Osmotic stress, Shoot	433 (10.7%)	342 (8.3%)	5.48E-08
Salt, Root	467 (11.5%)	334 (8.1%)	2.84E-15
Salt, Shoot	361 (8.9%)	278 (6.7%)	4.68E-08
UV, Root	341 (8.4%)	253 (6.1%)	2.78E-09
UV, Shoot	493 (12.2%)	335 (8.1%)	6.96E-21
Wound, Root	272 (6.7%)	182 (4.4%)	1.33E-12
Wound, Shoot	335 (8.3%)	245 (5.9%)	5.84E-10
Drought, Root	304 (7.5%)	205 (4.9%)	1.74E-13
Drought, Shoot	302 (7.5%)	243 (5.9%)	2.98E-05
Heat, Root	603 (14.9%)	525 (12.7%)	5.56E-05
Heat, Shoot	505 (12.5%)	493 (12.0%)	0.346283116

Oxidative stress, Root	244 (6.0%)	169 (4.1%)	8.92E-10
Oxidative stress, Shoot	283 (7.0%)	232 (5.6%)	0.00020
Sulfate deficient, Root	94 (2.3%)	74 (1.8%)	0.013
Biotic stress			
Pseudomonas syringae[b]	723 (17.9%)	539 (13.1%)	3.50E-19
NPP1[c]	626 (15.5%)	365 (8.9%)	6.98E-49
HRP Z[d]	593 (14.7%)	362 (8.8%)	4.80E-39
Flagellin[e]	518 (12.8%)	311 (7.5%)	3.52E-36
LPS[f]	378 (9.3%)	223(5.4%)	4.37E-28
Botrytis cinerea[g]	2126 (52.6%)	1789 (43.6%)	7.00E-31
Phytophthora infestans[h]	737 (18.2%)	551 (13.4%)	2.72E-19

[a] Expected number was calculated using the proportion of up-regulated genes from all annotated genes excluding gene duplicates using the t-test ($P \leq 0.01$). χ^2-test (d.f. =1) was used to test the difference between observed and expected numbers of gene duplicates. [b] *Pseudomonas syringae* pv. tomato DC3000, virulent pathogen infection; [c] NPP1, treatment of GST-Necrosis-inducing Phytophthora protein 1, a pathogen derived elicitor; [d] HRP Z, treatment of Hairpin Z, a proteinaceous elicitor of plant hypersensitive responses; [e] FLAGELLIN, treatment of FLAGELLIN, Flg22, *P. syringae*-derived peptide elicitor of plant defense response; [f] LPS, treatment of LPS, a pathogen derived elicitor constitutively present in the pathogen cell wall capable of inducing a plant host defense response; [g] treatment of pathogenic fungus, *Botrytis cinerea*, and [h] treatment of fungus-like pathogen, *Phytophthora infestans.*

Faster expression divergence in response to environmental factors than to developmental processes

We now consider the relative contributions of environmental and developmental factors to expression divergence between duplicate genes. To test this, we analyzed the Pearson's correlation coefficient of expression between gene duplicates in the developmental (R_{dev}) or environmental (R_{env}) process using the same number of expression datasets: 63 in different developmental stages and 63 treatment and time-course combinations in roots and shoots, respectively. The distributions of expression correlation coefficients were compared using the Wilcoxon rank-sum test [115]. As expected, correlation coefficients of expression profiles between randomly chosen genes showed a normal distribution with mean zero, and there was no significant difference in expression variation among random gene pairs in all three conditions (data not shown). Interestingly, the expression divergence of duplicate genes under environmental stress is significantly greater than that under developmental process (Figure 2.1A, $P \leq 2.2 \times 10^{-16}$). Furthermore, we analyzed the correlation coefficient difference (D) of each gene duplicate in developmental and environmental processes ($D_i = R_{dev,i} - R_{env,i}$, for the *i*th duplicate gene pair). The cumulative probability difference for the gene duplicates and random gene pairs between environmental and developmental processes was significantly different by either the Kolmogorov-Smirnov (KS) test [116] or the Wilcoxon rank sum test (Figure 2.1B). Taken together, the data indicate that expression divergence between

gene duplicates occurs faster in response to the environmental stresses than to the developmental changes.

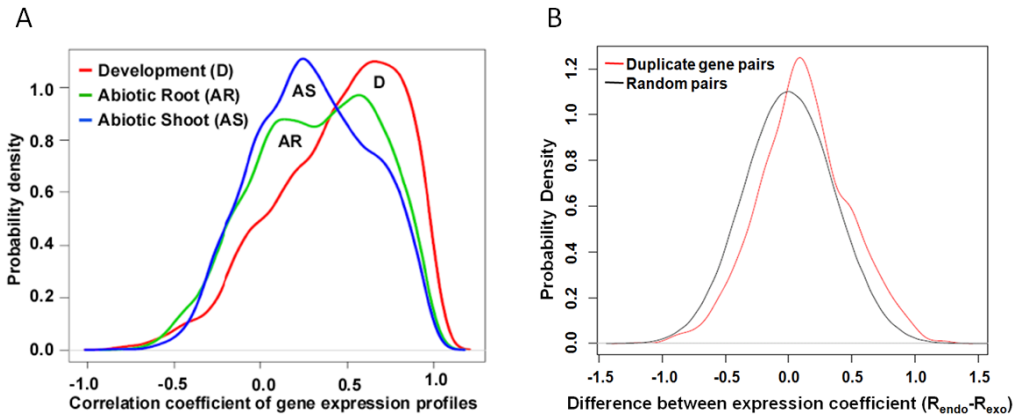

Figure 2.1 Distributions of expression correlations between gene duplicates
A. Distributions of expression correlations between gene duplicates in environmental factors and in developmental processes. The probability density of expression correlation coefficient is plotted against correlation coefficients. The distribution of expression correlations between duplicate genes in various developmental stages (red) is significantly shifted to the right of those subjected to abiotic treatments in roots (blue) and in shoots (green) (Wilcoxon rank sum test, $P \leq 2.2x10^{-16}$). The right-shift indicates less expression divergence. **B.** Probability density plot of difference ($d = R_{endo} - R_{env}$) between the duplicate genes and the randomly paired genes. Based on the Wilcoxon rank sum test, the distribution for the duplicate genes is shifted to the right of that for the randomly paired genes. For example, in one gene duplicate, R_{dev} is significantly larger than $R_{env(shoots)}$ ($d = 0.1460$) and $R_{env(roots)}$ ($d = 0.1282$) ($P \leq 2.2 \times 10^{-16}$).

To test if external factors are more effective in promoting expression divergence than other biological processes, we classified recent WGD duplicate genes into GO Slim biological processes [114] and analyzed expression correlation coefficients of gene duplicates in each category (Figure 2.2A). The levels of expression divergence between gene duplicates were highest in extracellular transport, signal transduction, stress

response and transcription and lowest in the cellular and developmental processes such as energy pathway, protein metabolism, intracellular transport, DNA/RNA metabolism, and cell organization and biogenesis.

To infer the biological processes responsive to external conditions, duplicate genes in the "transport" category were divided into extracellular and intracellular sub-groups. Indeed, duplicate genes in the "extracellular transport" showed the highest level of expression divergence, whereas those in the "intracellular transport" displayed a low level of expression divergence (Figure 2.1B). This supports the notion that gene expression divergence occurs at a faster rate in response to external than to internal factors. Note that biological processes in "response to external stresses" and "extracellular transport" would be directly affected by external conditions. So, we compared gene expression divergence in two groups: "other processes" and "response to external factors", which include external stresses (abiotic, biotic and other) and extracellular transport. The expression divergence is significantly faster in response to the external factors than to the other processes (Wilcoxon rank-sum test, $P \leq 1.29 \times 10^{-7}$) (Figure 2.2B). Note that this analysis may underestimate the difference because some potential external factors related to signal transduction and transcription were included in the "other processes".

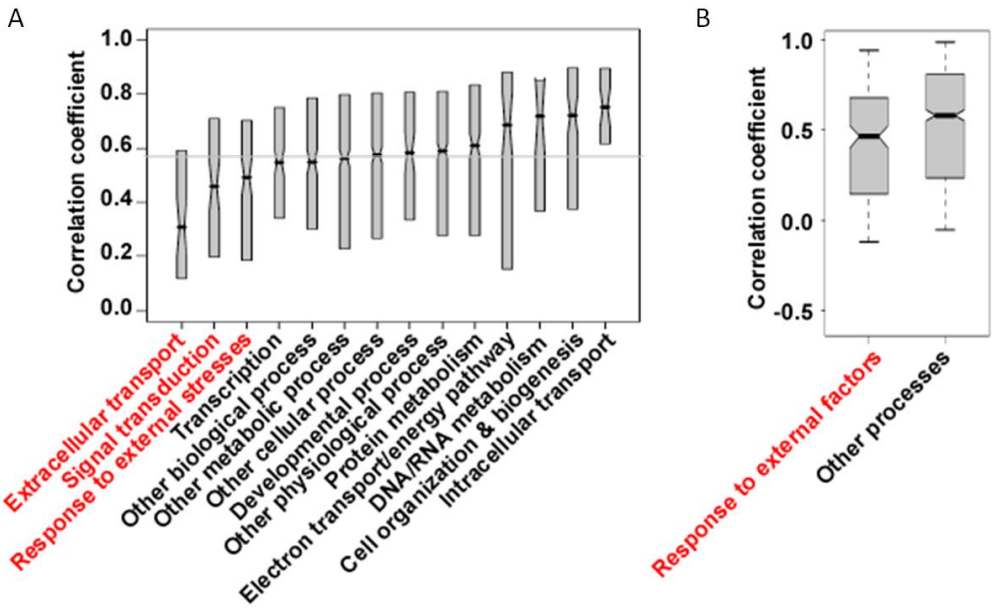

Figure 2.2 External factors accelerate expression divergence between duplicate genes

A. Distribution of expression correlation between duplicate genes in biological process classifications (except for biological functions unknown). Boxplots of expression correlations indicate distributions of 75th percentiles, median (red dots), and 25th percentiles. The levels of gene expression divergence were orderly arranged based on the median values that increased from the external factors (left) to the internal factors (right). The horizontal line shows median correlation coefficient using a total of 2022 gene duplicates. **B.** Gene duplicates that respond to external factors are significantly more diverged than gene duplicates in other processes. Gene duplicates in the extracellular transport and in response to external stresses were grouped into one category (response to external factors), and all other gene duplicates were considered to be in the "other processes". Dotted vertical lines indicate the range of gene expression divergence.

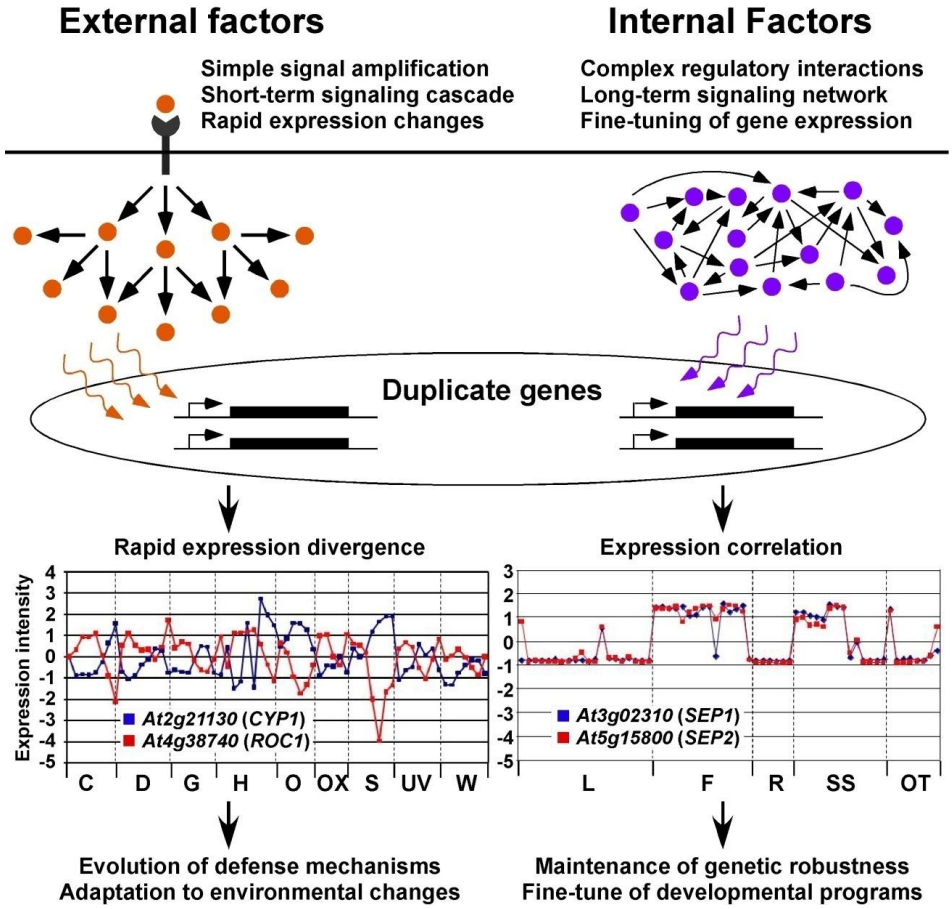

Figure 2.3 Different strategies for the evolution of duplicate genes in the external and internal processes

Expression intensities were standardized into the z score (z_{ij}) of gene i in condition j as

$$z_{ij} = \frac{(x_{ij} - \overline{x_i})}{\sqrt{s_{ii}}}$$

where x_{ij} is an expression value before standardization; $\overline{x_i}$ is a mean expression value of gene i; and s_{ii} is a standard deviation of gene i expression across the conditions (1…j). In response to external factors, gene duplicates (*At2g21130* and *At4g38720*) underwent fast expression divergence under various stress responses, leading to the evolution of an adaptive mechanism to adverse growth environments. In contrast, duplicate genes (*SEP1* and *SEP2*) encoding proteins involved in floral organ identity are both differentially regulated during developmental processes, but their expression patterns are highly correlated. C: cold; D: drought; G: genotoxic; H: heat; O: osmosis; OX: oxidative stress; S: salt; UV: UV-B; W: wounding; L: leaf; F: flower; R: root; SS: seed and seedling; and OT: others.

There is experimental support for the above conclusion. For example, *SEP1* (formerly *AGL2*) and *SEP2* (formerly *AGL4*) are gene duplicates expressed at the flower developmental stage [117], and their expression patterns are correlated throughout plant development (Figure 2.3, right lower panel). The two genes have a redundant function in the floral organ identity, and single-gene knockout shows no developmental defect [118]. In contrast, cyclophillin gene duplicates, At2g21130 (*CYP1*) and At4g38740 (*ROC1*), are induced by abiotic and biotic stresses (Figure 2.3, left lower panel) [119, 120]. Their expression levels are highly variable among various external stimuli, suggesting that the gene duplicate is involved in different regulatory networks. Although some extreme examples exist in large datasets, the experimental data collectively support the above notion.

DISCUSSION

Environmental stresses are often associated with a short-term cascade and/or simple signal amplification, leading to rapid changes in gene expression [121]. Therefore, external conditions may promote organisms to acquire an adaptive mechanism, as predicted by B. McClintock [122], through diversification of duplicate genes [98, 121] after WGD [101, 105-107]. Many plants respond to environmental stresses (e.g., drought and salt) by inducing the expression of stress-related genes and/or gene products [112, 113]. On the other hand, developmental programs affect gene expression via long-term, multi-stage, and complex molecular interactions, corresponding to a relatively slow rate of expression divergence between duplicate genes.

We propose a model (Figure 2.3) for different evolutionary fates of duplicate genes in response to external and internal processes. In external processes, duplicate genes diverge in expression relatively rapidly in response to abiotic and biotic stresses, which may facilitate subfunctionalization [123], neofunctionalization [124] and the evolution of an adaptive mechanism to environmental changes [122]. In internal processes (e.g., development), duplicate genes tend to be co-expressed. In development, a relatively slow rate of expression divergence between the duplicates may provide selective advantage via dosage-dependent gene regulation that enables organisms to fine-tune complex regulatory networks. Therefore, during evolution duplicate genes may promote an adaptive mechanism against environmental changes or provide genetic robustness and dosage-dependent regulation during organismal development.

The proposed model is also supported by other studies in yeasts. Stress-responsive WGD duplicate genes have more different expression profile than other WGD duplicate genes [23].

Gene expression evolution in external factors may be associated with protein sequence evolution. Changes in gene expression patterns may differentiate interacting protein partners. To optimize affinity with new interacting partners, duplicate genes may undergo rapid evolution of protein sequences associated with evolution of expression patterns. In human, genes involved in immune responses are under positive selection, and their protein sequences diverge faster than that of other genes [125].

3. GENE EXPRESSION DIVERSITY AMONG POLYPLOIDS BY DUPLICATE GENES

This chapter is being revised for publication after initial review by:

Misook Ha, Eun-deok Kim and Z. Jeffrey Chen

BACKGROUND AND RATIONALE

Polyploidy or whole genome duplication (WGD) is formed by duplication of a single genome (autopolyploidy) or combination of two or more distinct genomes (allopolyploidy) [126]. During evolution, duplicate genomes undergo massive gene loss and sequence changes, deletions, insertions, translocations, and/or other chromosomal rearrangements through a process known as diploidization [127]. Recent genome sequencing of many organisms has revealed that WGD occurred in most organisms during evolution. After WGD, some duplicate genes may undergo rapid loss [100, 128], while others are retained and provide dosage-dependent selective advantage, neofunctionalization, and/or subfunctionalization [11, 98, 123, 129, 130]. WGD affects genomic instability as well as gene expression patterns, leading to changes in growth, development, and reproduction [131]. At the expression level, change of gene expression is a major mechanism that polyploids establish novel traits and better fitness [16, 121]. Consequently, the merger of two distinct genomes in a new allopolyploid may generate novel expression variation (e.g., hybrid vigor) and increase fitness [98, 126, 132-134]. However, the mode and tempo of expression differentiation of duplicate genes in allopolyploids have been poorly understood. Within species, expression divergence between duplicate genes tends to expand regulatory networks and contribute to species-specific morphological diversity including variation in cell cycle control and organ development [19]. Therefore, one hypothesis is that retained duplicate genes derived

from ancient polyploidy events in the progenitor species increase expression diversity in allopolyploids or new polyploids, leading to adaptation and speciation.

To test the hypothesis, we analyzed gene expression microarray data in a natural allotetraploid *A. suecica* and two resynthesized allopolyploids derived from *Arabidopsis thaliana* and *A. arenosa*. The natural and resynthesized allotetraploids morphologically resemble the extant natural allotetraploid *A. suecica* that contains *A. thaliana* and *A. arenosa*-like genomes [97, 135]. The new allopolyploid lineages are suitable for testing the above hypothesis because *A. thaliana* and *A. arenosa* diverged recently, only ~6 million years ago (Mya) [95], after a WGD about 20 Mya [105, 136, 137]. The two species share >90% of nucleotide sequence identity in coding regions, and >90% of the ~26,000 70-mer *A. thaliana* oligos cross-hybridize with *A. arenosa* genes [138] (data not shown).

A. thaliana is inbreeding, whereas *A. arenosa* is outcrossing, and the difference in mating systems may promote gene expression divergence between species for adaptation [139]. Indeed, >15% of genes are expressed differently between the two species [97], which is reminiscent of the expression divergence observed between Drosophila species [140]. Over 68% of the genes that are expressed nonadditively (differently from the mid-parent value) in the resynthesized allotetraploids are also differentially expressed between parents, indicating a biological mediation for transcriptome divergence between the two species. To test the hypothesis that duplicate genes change expression regulation among different polyploid species, we first compared expression divergences of single-copy and duplicate genes between *A. thaliana* and *A. arenosa*. We then examined how duplicate

genes affect expression change in new round of polyploidy detecting expression variations of homoeologous single-copy and duplicate genes between each resynthesized allopolyploid and their progenitors. Finally, we examined the expression diversity of single-copy and duplicate genes in *A. suecica*, a natural allotetraploid species derived from extant progenitors *A. thaliana* and *A. arenosa*.

MATERIALS AND METHODS

Plant materials.

All plants were grown in vermiculite mixed with 30% soil in a growth chamber with growth conditions of 22°/18° (day/night) and 16 hr of illumination per day. The accessions included *A. thaliana* diploid ecotype Landsberg erecta (L*er*), tetraploid A. arenosa (Arabidopsis Biological Resource Center, accession no. 3901, 2n = 4x = 32), and natural *A. suecica* (9502) (2n = 4x = 26). Autotetraploid *A. thaliana* (2n = 4x = 20) was obtained through colchicine treatment of L*er* (accession no. CS3900). Rosette leaves prior to bolting were collected for the analysis of DNA, RNA, and gene expression variation.

DNA microarray experiments and data analysis.

To reduce the effects of gene copy number and genomic differences on expression, we performed DNA microarrays using comparative genomic hybridization (CGH) between *A. thaliana* autotetraploid (At4, accession no. CS3900) and *A. arenosa* (Aa, accession no.CS3901). Genomic DNA was isolated and sheared using a sonicator.

Probe labeling, slide hybridization, and washing were performed as previously described [97]. Raw data were collected using Genepix Pro4.1 after the slides were scanned using Genepix 4000B. The data were processed using a lowess function to remove nonlinear components and analyzed using a linear model [138]. Genomic hybridization intensities were analyzed [97], and the genes with statistically significantly different hybridization signals were excluded for the study. The genomic microarray data were deposited in Gene Expression Omnibus (GEO): accession no. GSE9512.

Analysis of gene expression data.

Genome-wide gene expression microarray data were obtained from three sets of comparisons between (1) *A. thaliana* and *A. arenosa*, (2) Allo733 and an artificial mix of two parents, and (3) Allo738 and an artificial mix of two parents [97]. Another set of microarray data was obtained using a comparison between mRNA from *A. suecica* and an artificial mRNA mix of two parents. The microarray data were deposited in GEO (accession no. GSE13468). Microarray data from two biological replications and two dye-swap experiments (8 hybridizations each) were analyzed using a linear model, and the results were adjusted for multiple comparisons [97]. The differentially expressed genes that were statistically significant under both common and per-gene variances were selected for the study.

Identification of expression of duplicate and single-copy genes.

Entire cDNA and protein sequences of *A. thaliana* were downloaded from TAIR database (ftp://ftp.arabidopsis.org/home/tair/Genes/). All-against-all protein sequence

alignment was performed using BLAST program. A locus was considered to be a single-copy gene if a protein sequence did not align with any other proteins using BLAST search (E <= 0.01). In this study, we used well-characterized gene duplicates that arose from the most recent WGD event ~20-40 Mya [105, 136, 137]. The WGD duplicate gene set was further processed to remove ambiguous loci as previously published [129]. Newly annotated pseudogenes and the genes with no detectable expression signals were excluded. To avoid the possibility of cross-hybridization among paralogous genes in microarrays, a microarray probe is selected only if its 70mer oligonucleotide probe does not match any other cDNA sequences with ≥ 70% identity and did not have 17 contiguous bases identical to any other cDNAs in *A. thaliana*.

Identification of paralogs.

The same definition of paralogs as Gu et al. [141] was used. Two genes were defined as paralogous if their protein sequences were matched using all-against-all BLAST with following criteria: (1) E value ≤ 10^{-10}; (2) sequence identity is greater than 30%; (3) The length of the alignable region between two protein sequences is greater than 50% of the longer sequence. With these criteria, close paralogous genes for duplicate genes were identified, and the number of paralogs of each duplicate gene was counted.

Assignment of gene ontology.

The Gene Ontology for *A. thaliana* was downloaded from The Arabidopsis Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/OLD/ATH_GO_GOSLIM.20080419.txt) released on 19 April 2008. GOSlim

was used to classify 13 biological process categories. Among the 2694 gene duplicates and 1347 single-copy genes, 2380 (88%) and 1205 (89%) genes were assigned using GOSlim biological process classifications.

Logistic regression model.

A simple logistic regression model was used to test the association of gene expression variation with the number of paralogs. The log odds ratio is related to the categories of gene duplicate numbers by linear model.

$$Ln(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x$$

where, p(x) represents the proportion of differentially expressed genes, and x is the category of gene duplicate numbers (x = 0, number of paralogs = 0; x = 1, number of paralogs = 1; x = 2, number of paralogs = 2 to 9; and x =3, number of paralog ≥ 10). The regression coefficient, parameter β_1, measures the degree of association between the tendency of the differential expression and the number of paralogs.

DNA methylation in the 5' upstream regions.

Genome-wide DNA methylation data were obtained from the published work [142]. Genes methylated in the 5' upstream region are defined as presence of two or more adjacent methylated probes (immunoprecipitated DNA /Input ≥ 1.28) within a 1-kbp 5' upstream region. A total of 965 genes were detected to be methylated in the 5' upstream regions. Fifty-eight out of 2,694 duplicate genes and 54 out of 1,347 single-copy genes, respectively, are methylated in their 5' upstream regions.

RESULTS

Sequence conservation between *A. thaliana* and *A. arenosa*.

We selected the duplicate genes from the most recent WGD event that occurred 20-40 Mya because they are accurately detected and generally conserved [105, 136, 137]. It is unknown, however, how orthologous genes have been conserved between *A. thaliana* and *A. arenosa* after they split ~6 Mya [95]. In a pilot project, we sequenced one *A. arenosa* BAC in the vicinity of the *FLC* locus. The BAC had a ~110-kb insert and consisted of 32 genes including *FLC* (At5g10140). The gene orders and genomic organization were completely colinear between *A. thaliana*- and *A. arenosa*-derived sequences (Figure 3.1), suggesting that these two regions are highly conserved. Within the ~110-kb regions, the nucleotide sequence identities between *A. thaliana* and *A. arenosa* were 94.6% in exons, 79.3% in introns, 82.8% in untranslated regions (UTRs), and 42.6% in aligned intergenic regions. The high level of sequence identity (~95%) in coding regions agrees with the previous data based on randomly sequenced cDNA fragments and clones [21, 55, 138]. Together, the available data suggest that coding sequences between *A. thaliana* and *A. arenosa* are highly conserved, and the oligo-gene microarrays designed from *A. thaliana* genes can be used to study *A. arenosa* genes [97].

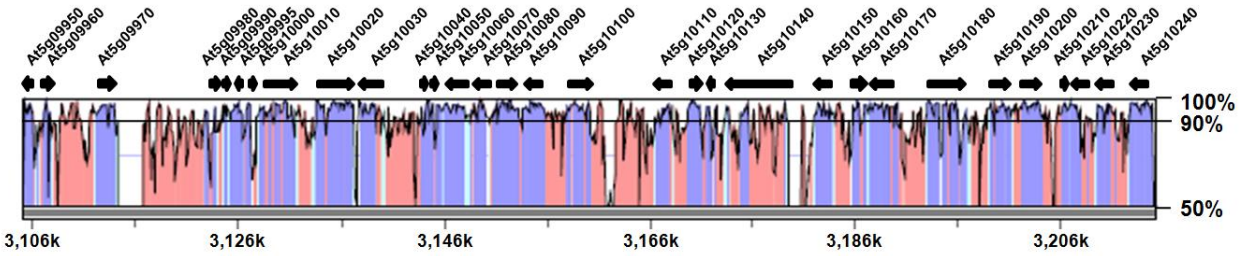

Figure 3.1 Sequence comparison between A. thaliana and A. arenosa in the vicinity of FLC on chromosome 5
BAC sequence from *A. arenosa* was aligned to *A. thaliana* genomic sequence using genome VISTA. Purple, blue, and pink represent protein coding exons, introns, and inter-genic regions, respectively. The percentage of nucleotide sequence identity between *A. thaliana* and *A. arenosa* was shown in the right. The orientation and names of loci are shown at the top, and the genomic coordinates are shown in the bottom.

To test the expression evolution of single-copy and duplicate genes, we applied the following criteria for the duplicate genes: (1) duplicate genes are present in both *A. thaliana* and *A. arenosa* prior to speciation; (2) orthologous genes equally hybridize with microarray probes; and (3) paralogous genes do not cross-hybridize. To satisfy these requirements, we selected the duplicate genes with the same hybridization intensities in comparative genomic hybridization (CGH) between *A. thaliana* and *A. arenosa*. Sequence similarity among oligonucleotide probes and target sequences is a major determinant of cross-hybridization in spotted oligonucleotide microarrays [143-145]. To minimize cross-hybridization, we selected duplicate and single-copy genes based on 70-mer sequences that had ≤70% of sequence identity with any other cDNAs and did not have 17 contiguous bases complementary to any other cDNAs. With these selection criteria cross-hybridization should be negligibly small [144, 145].

A single-copy gene was defined by its protein sequence that did not match any other paralogous proteins using BLASTp ($E \leq 0.01$) [141]. Only the single-copy genes with the same hybridization intensities in CGH between *A. thaliana* and *A. arenosa* were included in the study. Consequently, 1347 single-copy and 2694 WGD duplicate genes that have unique probes in the spotted oligo-gene microarrays were used for further analysis.

Expression divergence of duplicate genes between species.

We tested if more duplicate genes are differentially expressed than single-copy genes between *A. thaliana* and *A. arenosa* that split ~6 Mya [95]. Microarray data from four dye-swap experiments were analyzed using a linear model, and the results were adjusted for multiple comparisons. A gene is differentially expressed if its expression level is significantly different using both tests of common and per-gene variances [97, 138]. A total of 3,923 out of 26,090 (~15%) genes were expressed differently between the two species. By comparing these genes with 2,694 duplicate genes and 1,347 single-copy genes, we found that the proportion of duplicate genes (~18%, 478/2,694) that expressed differently between the two species was significantly higher than that of single-copy genes (~13%, 175/1,347) (Pearson's Chi-square test with Yates' continuity correction, $\chi^2 = 13.8$ and $P = 0.0002$, Figure 3.2A). The data suggest that expression divergence between duplicate genes plays a role in establishing different expression patterns in closely related species.

Expression divergence between single-copy and duplicate genes in allopolyploids.

To test how duplicate genes from the WGD event contribute to expression variation in new allopolyploids, we examined expression levels of homoeologous single-copy and duplicate genes in two newly resynthesized allotetraploids. The differentially expressed genes in the allotetraploids were detected by comparing mRNA levels in an allotetraploid with an equal mixture of RNAs from the two parents (mid-parent value, MPV) [97]. If gene expression is additive, the expression level of a gene in an allotetraploid should be equal to the sum of two parental loci (null hypothesis: 1 + 1 = 2). Nonadditive expression suggests repression (<2) or activation (>2) of a gene in the allotetraploid compared to MPV. This method may underestimate the number of nonadditively expressed genes because we could not detect a situation in which repression of one allele was compensated by the activation of another [97, 126]. A total of 1362 (~5.2%) and 1469 (~5.6%) genes are expressed nonadditively in two independent allotetraploid lineages, Allo733 and Allo738 [97].

By comparing nonadditively expressed genes with homoeologous single-copy and duplicate genes, we found that the proportion of homoeologous duplicate genes that were nonadditively expressed in both allotetraploids was significantly higher than that of homoeologous single-copy genes (Figure 3.2B and C, Pearson's Chi-square test with Yates' continuity correction, $P = 5x10^{-5}$). These data suggest rapid expression divergence between duplicate genes after immediate allopolyploidization. Note that the proportion of duplicate genes displaying expression changes may be underestimated because some nonadditively expressed duplicate genes were excluded (see Methods).

Resynthesized allopolyploids are suitable materials for the study because the exact progenitors are known. Moreover, homoeologous genomes in the resynthesized allotetraploids are relatively stable after selfing for 6 generations [97, 146]. Therefore, differences in microarray hybridization intensities are mainly due to gene expression changes rather than sequence differences.

To test if gene expression changes in resynthesized allotetraploids also occur in "old" allotetraploids, we examined the role of duplicate genes in expression diversity in a natural allotetraploid. *A. suecica* was formed by interspecific hybridization between extant *A. thaliana* and *A. arenosa* species from 12,000 to 300,000 years ago [135, 147]. Relative to two extant progenitors (MPV), 1,855 (~7%) of the genes are nonadditively expressed in *A. suecica*, which is consistent with the number of nonadditively expressed genes found in two resynthesized allotetraploids. Furthermore, we found that homoeologous duplicate genes were significantly enriched in nonadditively expressed genes in *A. suecica* compared to homoeologous single-copy genes (Figure 3.2D, Pearson's Chi-square test with Yates' continuity correction, $P < 0.04$). Although only one natural allotetraploid species was examined, the data suggest that similar to resynthesized allotetraploids, evolution of gene expression in natural allopolyploids is partly caused by the expression divergence between duplicate genes.

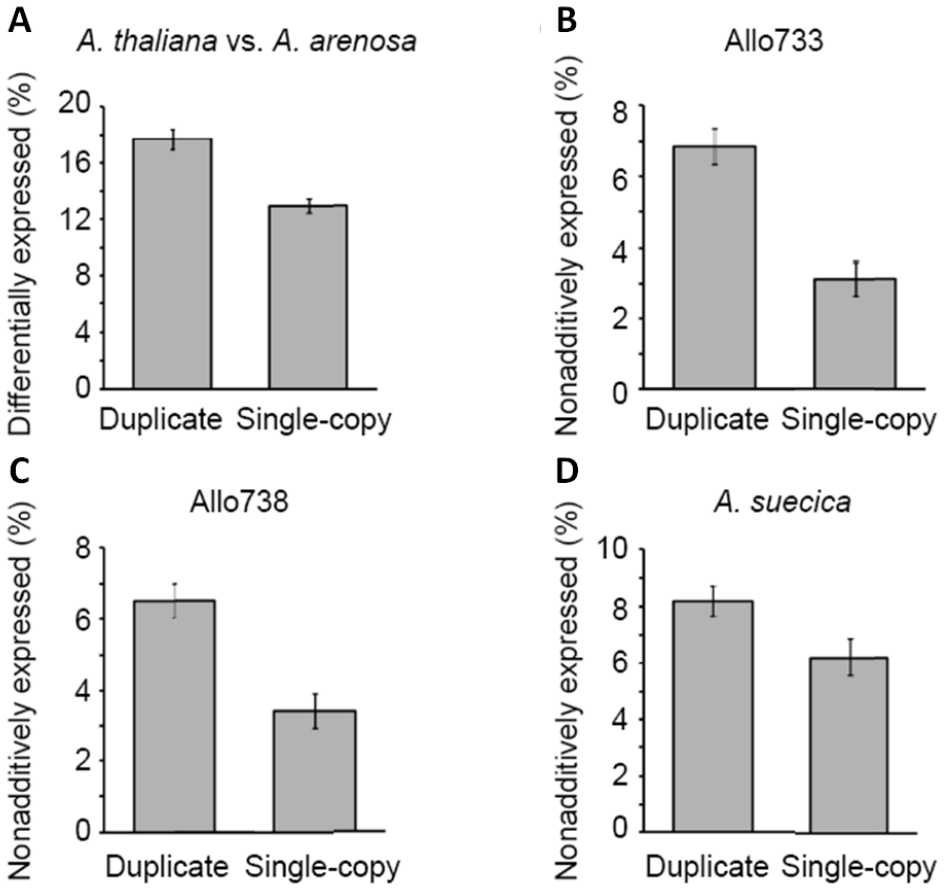

Figure 3.2 Expression change of duplicate genes between progenitors and in allopolyploids

Proportions of single-copy and duplicate genes are shown to be differentially expressed between *A. thaliana* (At4) and *A. arenosa* (Aa) (A) and nonadditively expressed in allotetraploids Allo733 (B), Allo738 (C) and *A. suecica* (D) relative to the mid-parent value (MPV). Pearson's Chi-square tests with Yates' continuity correction are as follows: (A) $\chi^2 = 13.8$, degrees of freedom, d.f. = 1, P = 0.0002; (B) $\chi^2 = 23.5$, d.f. = 1, P = 1.3×10^{-6}; and (C) $\chi^2 = 16.5$, d.f. = 1, P = 4.8×10^{-5}. (D) Fisher's Exact test, P = 0.049. Unless noted otherwise, the standard errors were estimated using 10,000 replications of bootstrapping.

Expression divergence between duplicate genes involved in external processes.

Duplicate genes after WGD are often differentially expressed in various developmental stages and environmental conditions, and external factors accelerate expression divergence between duplicate genes [129, 148], providing molecular bases for dosage-dependent selection and adaptive evolution [129, 130]. To understand how the expression of duplicate genes changes in response to external and internal signals, we compared distributions of single-copy and duplicate genes that were differentially or nonadditively expressed in various Gene Ontology Slim (GOSlim) biological processes. Compared to single-copy genes, duplicate genes were enriched in all functional categories except for nucleotide metabolism, which has a small number of single-copy genes. A small proportion of transport and transcription related duplicate genes showing expression changes may suggest that these genes are dosage-insensitive [130]. In the category of "response to stress and external stimuli" more duplicate genes than single-copy genes were differentially expressed between *A. thaliana* and *A. arenosa* (Figure 3.3A). Moreover, homoeologous duplicate genes in the "response to stress and external stimuli" showed a higher proportion of nonadditive expression than other homoeologous duplicate genes (Figure 3.3B, C and D, Pearson's Chi-square test with Yates' continuity correction, $P \approx 0$). We further tested whether changes in gene expression co-evolve with promoter sequences and regulatory elements as shown in yeast [149]. Indeed, compared to those without TATA box, the TATA-containing duplicate genes tend to be differentially expressed between two species and nonadditively in two independent allotetraploid lineages. These data suggest that following speciation and

allopolyploidization duplicate genes related to environmental cues and stress pathways undergo rapid changes in gene expression probably via divergence in regulatory elements. This may explain why gene duplicates are preserved in polyploids because their gene products are maintained by adaptive selection.

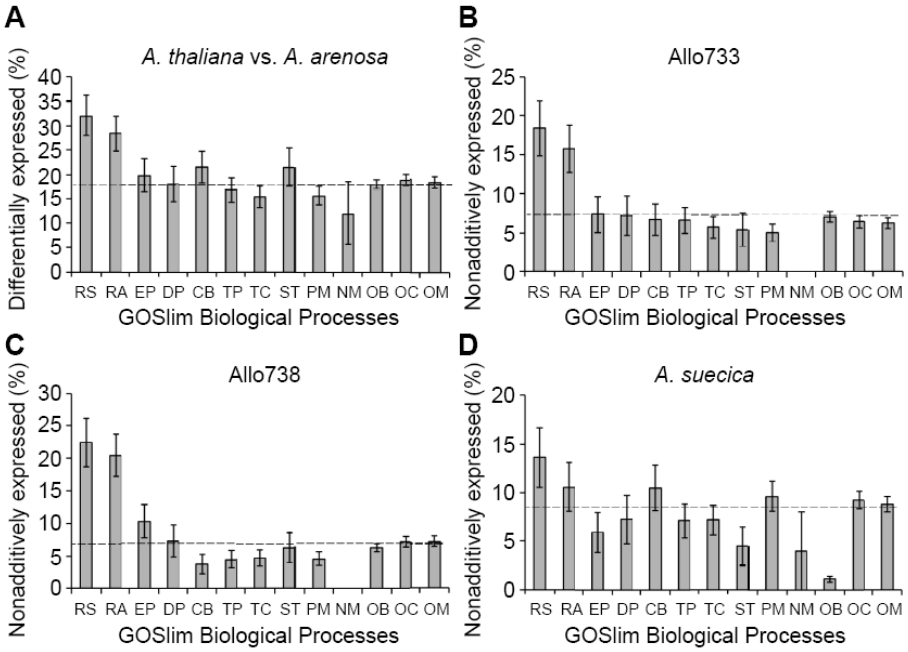

Figure 3.3 Differential expression of duplicate genes in GOSlim biological process classifications in allotetraploids and their progenitors

(A). Duplicate genes in external biological processes are differentially expressed between *A. thaliana* and *A. arenosa* (RS, $\chi^2 = 15.3$, d.f. = 1, $P = 9.4 \times 10^{-5}$ and RA, $\chi^2 = 10.0$, d.f. = 1, P = 0.002). (B). Duplicate genes in "response to stress and external stimuli" are differentially expressed in the allotetraploid Allo733 relative to the mid-parent value (RS, $\chi^2 = 21.6$, d.f. = 1, $P = 3.4 \times 10^{-4}$ and RA, $\chi^2 = 15.5$, d.f. = 1, $P = 8.0 \times 10^{-5}$). (C). Duplicate genes in "response to stress and external stimuli" are differentially expressed in the allotetraploid Allo738 relative to the mid-parent value (RS, $\chi^2 = 42.5$, d.f. = 1, $P = 7.2 \times 10^{-11}$ and RA, $\chi^2 = 39.0$, d.f. = 1, $P = 4.3 \times 10^{-10}$). (D). Duplicate genes in "response to stress" are differentially expressed in the nautral allotetraploid *A.suecica* relative to the MPV (randomization test, P = 0.02). RS: response to stress; RA: response to abiotic or biotic stimulus; EP: energy pathways; DP: developmental processes; CB: cell

organization and biogenesis; TP: transport; TC: transcription; ST: signal transduction; PM: protein metabolism; NM: nucleotide metabolism; OB: other biological processes; OC: other cellular processes; and OM: other metabolic processes. GoSlim biological processes are classified according to the TAIR release of 17 March 2007.

Rapid expression divergence in genes with multiple paralogs.

Retained gene duplicates after a series of WGD events expand gene families [121], probably via a "balanced gene drive" mechanism [150]. Functional compensation by duplicate genes may decrease constraints on gene dosage and increase variability of duplicate biological modules, which may facilitate balancing selection. To test this hypothesis, we examined the relationship between the number of paralogous genes and gene expression variation between species and in allotetraploids. We divided all single-copy and duplicate genes tested into four categories relative to the number of close paralogs (0 = single-copy, 1, 2-9, and ≥10). Among the genes that displayed differential expression patterns between species and nonadditive expression in the allotetraploids, those with 2-9 and ≥10 paralogs were significantly overrepresented, whereas single-copy genes were underrepresented (Figure 3.4, Pearson's Chi-square test with Yates' continuity correction, degree of freedom = 3, At4 vs. Aa, χ^2 = 80.0 and P = 0; Allo733, χ^2 = 69.6 and P = 5.3×10^{-15}; Allo738, χ^2 = 69.5 and P = 5.4×10^{-15}; *A. suecoca*, χ^2 = 125.4114 and P = 2.2×10^{-16}).

.

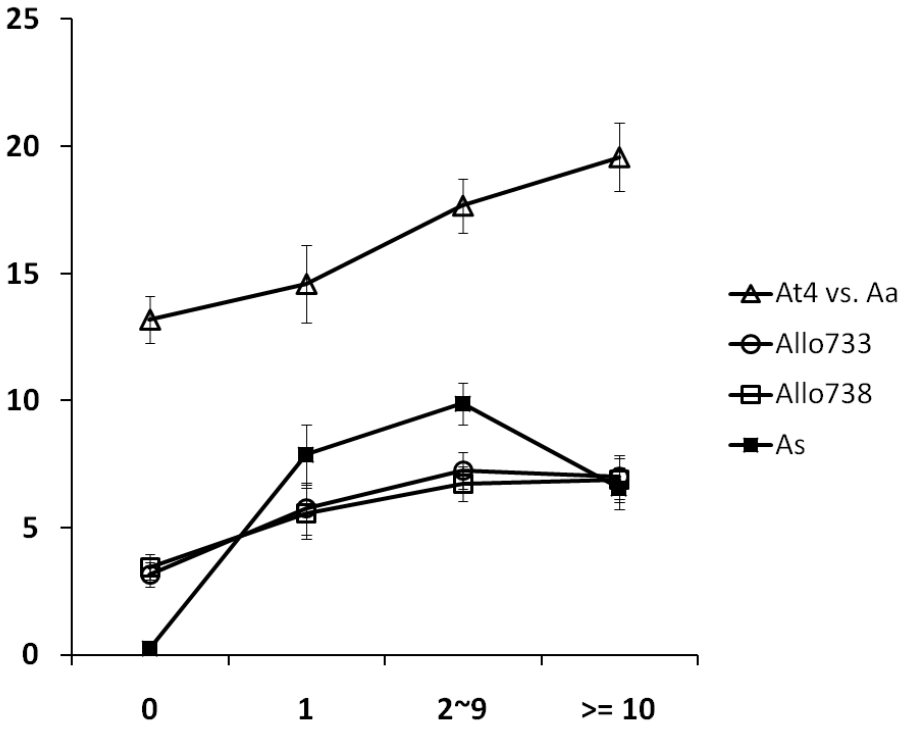

Figure 3.4 Differentially expressed genes with the number of paralogs
Differentially expressed genes with the number of paralogs between *A. thaliana* and *A. arenosa* (open triangle) and nonadditively expressed genes in the allotetraploids Allo733 (open circle) and Allo738 (open square). The proportion of differentially expressed genes significantly increases as the number of paralogs increases. The P values of logistic regression were 1.2×10^{-5} (*A. thaliana* and *A. arenosa*), 4.7×10^{-6} (Allo733), 6.6×10^{-5} (Allo738) and 6.9×10^{-16} (*A.suecica*).

Using a simple logistic regression model, we analyzed the association of differential gene expression with the number of paralogs. The regression coefficients for the number of paralogs with the proportion of differentially expressed genes between species and in two allotetraploids are significantly greater than 0 (At4 vs. Aa, $\beta_1 = 0.2$, $P = 1.2 \times 10^{-5}$; Allo733, $\beta_1 = 0.3$, $P = 4.7 \times 10^{-6}$; and Allo738, $\beta_1 = 0.2$, $P = 6.6 \times 10^{-5}$; As, $\beta_1 = 0.5$, $P = 6.9 \times 10^{-16}$). The data suggest that the proportion of differentially expressed genes

between species and in allopolyploids increases as the number of paralogs increases (Figure 3.4). Although the overall expression distribution in *A. suecica* is higher than in two resynthesized allotetraploids, a low frequency of nonadditively expressed single-copy genes in *A. suecica* suggest functional fixation of homoeologous single-copy genes in natural allopolyploids. Alternatively, over time sequence divergence among homoeologous duplicate genes (≥10 copies) may account for decreased frequency of nonadditively expressed genes in *A. suecica*

Gene expression diversity expands as the number of paralogs increases, which supports the models of dosage-dependent positive selection and balanced molecular drive [150, 151] and may explain the high level of duplicate-gene preservation in polyploid species and gene expression variation associated with hybrid vigor in allopolyploids [126, 152].

Promoter regions of duplicate genes are less methylated.

The gene expression changes between closely related species may be caused by *cis*-regulatory elements as well as *trans*-acting and species-specific factors [153]. Upstream sequence divergence between duplicate genes may alter binding affinities to RNA polymerases, transcription factors, and epigenetic modifiers such as DNA methylation and histone methylation that are essential for transcriptional regulation. Although DNA methylation occurs in coding and noncoding regions [154], DNA methylation in the promoter regions is generally associated with transcriptional silencing [142, 154]. To examine a role of DNA methylation in expression divergence between

single-copy and duplicate genes, we investigated DNA methylation patterns in the upstream regions of single-copy and duplicate genes. Using genome-wide DNA methylation data [142], we found that many genes are highly methylated in the upstream regions. A gene is considered to be methylated in the upstream region if two or more adjacent probes are significantly methylated within a 1-kbp upstream region. We found that in *A. thaliana* the proportion of duplicate genes with DNA methylation in the upstream regions is significantly lower than that of single-copy genes (Figure 3.5, Pearson's Chi-square test with Yates' continuity correction, $P < 0.001$). Avoidance of DNA methylation in the promoter regions of duplicate genes implies that duplicate genes have a higher potential of regulation than sing-copy genes through interactions with transcription and *trans*-acting factors in the interspecific hybrids and allopolyploids, which may promote expression divergence between duplicate genes in closely related species and interspecfic hybrids and allopolyploids.

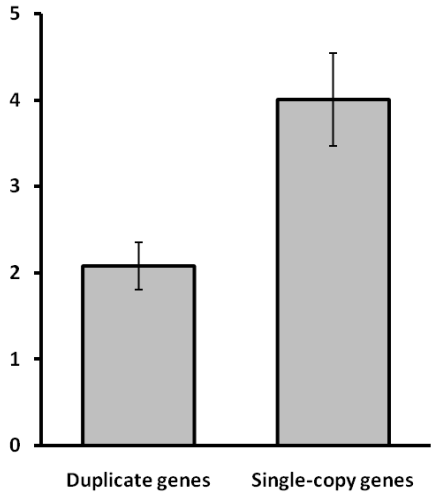

Figure 3.5 DNA methylation in promoter of duplicate and single-copy genes
Duplicate genes are void of DNA methylation in the upstream regions than single-copy genes (Pearson's Chi-square test with Yates' continuity correction, $P < 0.001$).

DISCUSSION

All flowering plants genome that have been sequenced underwent three or more rounds of polyploidy (WGD) over the last ~300 million years (Myr) [155]. It has been shown that duplicate genes retained in one WGD tend to be retained again after subsequent genome duplication [121]. Retained duplicate genes after WGD are enriched with genes involved in transcriptional regulation and in response to environmental changes [129]. In this study we first compared expression divergence between WGD duplicate genes between closely related species of a shared ancestor. Second, we investigated expression divergence between duplicate genes after additional new and old events of allopolyploidization. Consistent with previous findings in Drosophila, yeast, and mouse [141, 156], our data suggest that duplicate genes increase expression diversity

within and between Arabidopsis species. Moreover, among nonadditively expressed genes in the allotetraploids, more homoeologous duplicate genes than homoeologous single-copy genes of progenitor species are expressed nonadditively in response to the new event of allopolyploidization. Nonadditive expression in new allopolyploids may result in nonadditive expression and/or novel phenotypic variation [126]. Expression modulation of nonadditively expressed duplicate genes may lead to coexistence of the two distinct genomes within a single cell, induce novel phenotypes (e.g., heterosis), and facilitate interspecies hybrid speciation. For example, nonadditive expression of *A. thaliana* and *A. arenosa FLC* loci determines flowering time variation [21] and reproductive isolation.

Another conrtirbution of gene duplicates in expression evolution is that duplicate genes rapidly diverge in expression in response to changes in environmental conditions following new polyploidization event. Duplicate genes may expand regulatory networks of gene expression, conferring adaptive evolution. High rates of single-nucleotide polymorphisms (SNPs) have been associated with gene families in response to environmental conditions among different strains of *A. thaliana* [157]. Presence of sequence variation in regulatory regions may induce expression divergence between duplicate genes in external biological processes. The 5’ upstream regions of duplicate genes are less methylated, which may facilitate expression divergence between duplicate genes through interactions with transcription factors and *cis*- and *trans*-acting proteins. The level of gene expression diversity increases as the gene duplicate number increases, suggesting that duplicate genes subsequently retained after many duplication events tend

to diverge expression among polyploids. This is an alternative explanation to purifying selection against reduction of dosage-sensitive gene duplicates after WGD [130]. Genome duplication may be considered a mutagen, and imbalance in gene-copy numbers would be deleterious and selected against [151]. Duplicate genes may reduce constraints of dosage-dependent regulation and increase expression potentials. Indeed, duplicate genes are enriched in the external processes and diverge rapidly in response to environmental stresses [129]. Our experimental data reinforce the roles for duplicate genes in dosage-dependent regulation and adaptive evolution.

In conclusion, the large number of homoeologous duplicate genes that are preserved in polyploid species provides a beneficial effect of genomic obesity on morphological and adaptive evolution among species. WGD gene duplicates increase expression diversity as well as promotes regulatory sequence divergence [141, 149]. Expression diversity conferred by duplicate genes derived from previous WGD expands gene expression regulatory networks and facilitates organismal adaptation to its environment after new round of polyploidization [129]. Polyploidization preserves dosage-responsive genes and eliminates dose-sensitive genes [130, 148], which may be subsequently maintained by stabilizing selection. The preservation of duplicate genes complements the diploidization process following WGD that leads to the reduction of duplicate genes and genome size by genomic rearrangement and gene loss [100, 128]. The empirical and experimental data suggest that both preservation and reduction of duplicate genes are actively operative during eukaryotic genome evolution and polyploid speciation.

4. SMALL RNAS IN POLYPLOIDS

This chapter is in preparation for a manuscript by:

Misook Ha, Jie Lu, Lu Tian, Vanitharani Ramachandran, Xuemei Chen, Xiu-Jie Wang, and Z. Jeffrey Chen

BACKGROUND AND RATIONALE

Both allopolyploids and autopolyploids are prevalent in nature, suggesting an evolutionary advantage of having multiple sets of genetic material for adaptation and development. Moreover, heterozygosity and intergenomic interactions in allopolyploids may lead to phenotypic variation and growth vigor.

Arabidopsis suecica is a natural allotetraploid derived from extant *A. thaliana* and *A. arenosa* species ~12,000 to 300,000 years ago [147]. *A. thalian*a and *A. arenosa* diverged ~6 million years ago (Mya) [95], similar to the divergent time between human and chimpanzee [158]. Resynthesized allotetraploids were produced by pollinating tetraploid *A. thaliana* with *A. arenosa* [54, 97]. They are genetically stable, resemble A. suecica (Figure 4.1) and display morphological vigor, making them a suitable system for studying mechanisms for changes in gene expression and growth vigor [1].

In closely related species, transcriptome divergence occurs at 15-45% of the genes in Arabidopsis [97] or Drosophila [140]. Among 3,900 genes that are differentially expressed between *A. thaliana* and *A. arenosa*, ~68% are nonadditively expressed in the allotetraploids [97]. Significantly, many microRNA (miRNA) targets including transcription factor genes were nonadditively expressed in the two allotetraploids, suggesting a role of miRNAs in interspecies variation of gene expression and development [28].

MicroRNAs (miRNAs) and small interfering RNAs (siRNAs) are produced in diverse species and control gene expression and epigenetic regulation [81, 93, 159, 160].

Although physiological and developmental roles of miRNAs and siRNAs have been extensively studied in plants and animals, expression diversity and evolution of miRNAs and siRNAs in closely related species are poorly understood. Therefore, I performed comprehensive analyses of miRNA expression and siRNA distribution in two closely related species (Arabidopsis thaliana and A. arenosa), a natural allotetraploid (A. suecica), and two resynthesized allotetraploid lines (F1 and F7) derived from A. thaliana and A. arenosa.

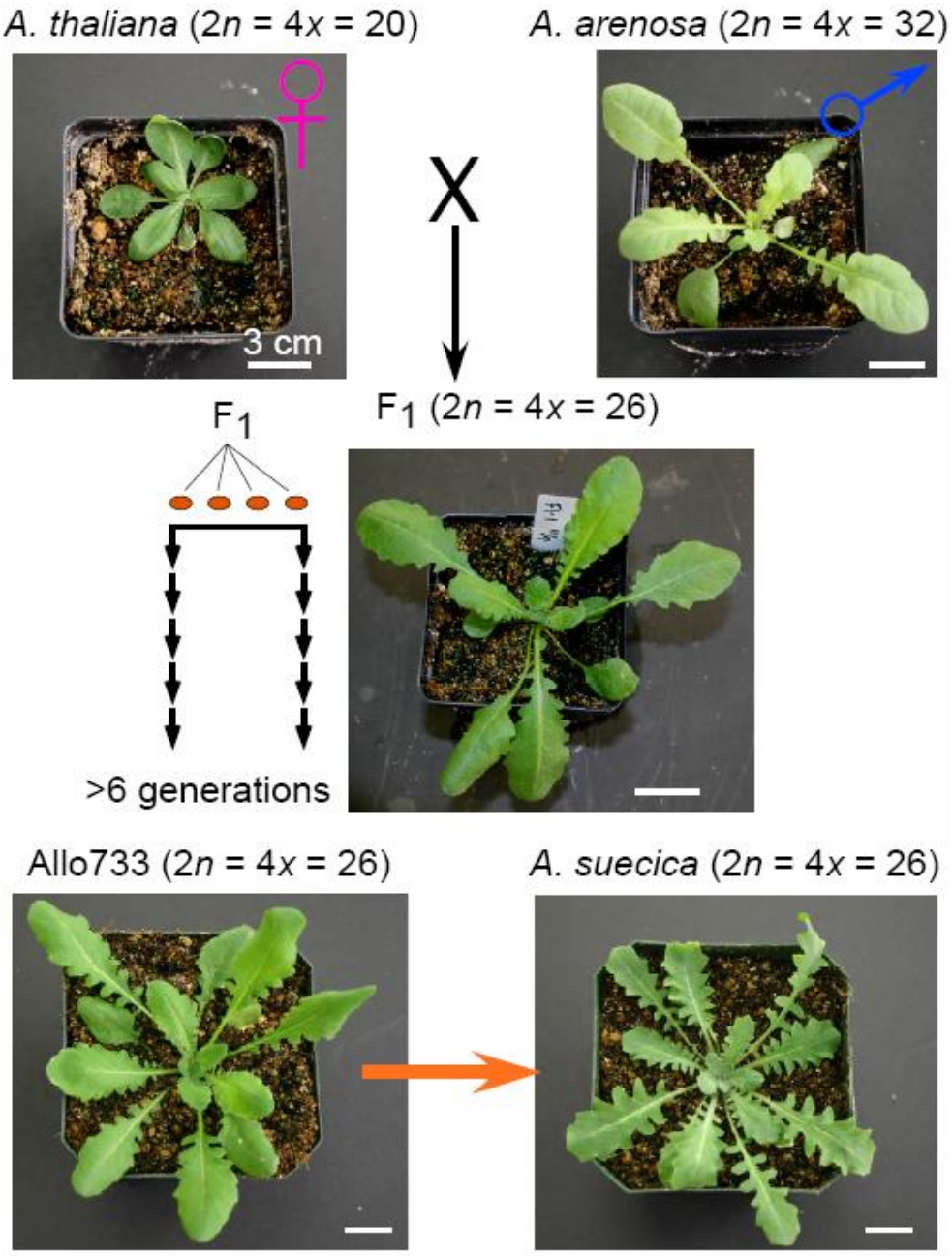

Figure 4.1 Plant materials used for sequencing small RNAs
Nascent allotetraploids (F_1) were produced by hybridization between *A. thaliana* and *A. arenosa*. Multiple F_1 allotetraploids were selfed for more than six generations to reduce the level of heterozygosity. A stable allotetraploid line (Allo733 in the 7th generation) and the natural allotetraploid *A. suecica* are shown.

MATERIALS AND METHODS

Plant materials.

Plant materials including natural *A. suecica* (#9502), resynthesized allotetraploids and their progenitor were produced as previously described [97], expect for the resynthesized allotetraploids that were one generation older than those used previously.

Small RNA library construction and sequencing.

To determine small RNA profiles in Arabidopsis allotetraploids and their progenitors, we made 10 small RNA libraries from rosette leaves (L) and flower buds (F) in five lines, *A. thaliana*, *A. arenosa*, Allo(F_1), Allo733(F_7), and *A. suecica* (Figure 4.1). Small RNAs of 15-100-nt in size were purified in a 15% polyacrylamide gel and ligated to the 5' and 3' RNA adaptors, respectively. The resulting RNA bands ranged from 55- to 140-nt. The ligated RNAs from each sample were reverse- transcribed, and the first-stranded cDNAs in each sample were amplified using the primer pair that contains two specific nucleotides as a "barcode". Four "barcoded" samples were pooled, and a small aliquot of pooled DNA was cloned and sequenced to determine the quality and representation of cloned products. After the quality control was done, the pooled DNA was subjected to high-throughput pyrosequencing that yields a total of ~1.5-million reads in seven runs.

The raw sequences were processed by identifying the barcodes, base call quality, and removing adaptor sequences. Each of the 10 libraries had a unique 4-nucleotide tag

within the 5'-end adaptor sequences. After trimming adaptor sequences at both 5' and 3' ends, pyrosequenced cDNAs were aligned to full genomic sequences of *A. thaliana*. To characterize small RNA populations in closely related species and exclude contaminant sequences, we used the sequences matching perfectly *A. thaliana* Columbia or Ler genomic sequences. Sequences matching other cellular RNAs including pre-tRNAs, rRNAs, snoRNAs, and snRNAs were regarded as degradation products of other cellular RNAs and excluded for analysis. Also excluded were sequences matching genome sequences of mitochondria and chloroplasts. Remaining sequences 20-25-nt in length that were identical to the *A. thaliana* genome were considered to be unique small RNAs (sRNAs) (Table 4.1). A genomic region matching several sRNAs were clustered into an sRNA locus if they were overlapping or located within 250-bp. The annotation (version TAIR8, April 2008) of genes, repeat elements, pseudogenes associated with transposable elements, and pseudogenes not associated with transposable elements were down loaded from The Arabidopsis Information Resource (TAIR) (ftp://ftp.arabidopsis.org/home/tair/Genes/ TAIR8_genome_release).

Small RNA generating genes.

To make sure the presence of small RNAs around genes, we considered only genes with at least 10 small RNA sequencing reads in the 1kb 5' upstream region and 3' downstream region. A total of 1437 elements including 473 genes generated siRNAs in 5' upstream regions, and additional 610 elements including 180 genes are generating siRNAs in their 3' downstream regions.

MicroRNA microarray experiments and data analysis.

We analyzed miRNA abundance in *A. thaliana*, *A. arenosa*, resynthesized allotetraploid (Allo733F_7), and *A. suecica* using miRNA microarrays (ComiMatrix http://www.combimatrix.com/products_microrna.htm). Small RNAs were enriched from total RNA using miRVana kit (Ambion, Austin, TX). Custom designed chips, each containing four identical arrays spotted with anti-sense DNA oligo nucleotides corresponding to miRNAs (version 9.1, http://microrna.sanger.ac.uk), ta-siRNAs and selected endogenous siRNAs (ASRP), were obtained. After prehybridization, the chip was hybridized to small RNAs labeled with Cy5 (Mirus Bio corporation) overnight at 37°C. Posthybridization washes were done sequentially once each with 6xSSPET and 3xSSPET and twice with 0.5xSSPET. Finally hybridized chip was scanned using a Genepix 4000B (Molecular devices) and data was extracted using a software from CombiMatrix. A total of 9 hybridization intensities in three biological replicates per miRNA were obtained and log values of hybridization intensities were used for analysis. If miRNA hybridization intensities are not significantly higher than background signal (probe with 2 nucleotide mutation), the miRNA is considered to be not expressed. miRNAs from other species (rice, poplar) showed below background signal and were excluded from analysis.

For expressed miRNAs and ta-siRNAs, a statistical analysis was conducted to examine different level of miRNA accumulation in 4 different species. We used a linear model to exclude technical variation from biological variation.

For each microRNA feature, the linear model for the intensity of miRNA g, replicate h, species i,

$$\text{Log}\,(Y_{ghi}) = \mu + G_i + R_j + S_k + (G*S)_{ik} + (G*R)_{ij} + (G*R*S)_{ijk} + \varepsilon_{ijk}$$

where g = 1… 84; h = 1, 2, 3; i = 1, 2, 3, 4.

G, R, S are main sources of variation from gene (G), replicate (R), species (S).

To test difference of miRNA accumulation between two species (e.g. *A. thaliana* and *A. arenosa*), the null hypothesis (H_o) is $S_k + (G*S)_{ik} = S_i' + (G*S)_g'_I'$ was tested for individual genes. Differential accumulation between *A. thaliana* and *A. arenosa* was tested using t-test for 84 distinct mature miRNA sequences. Non-additive accumulation was tested using linear contrast. For each mature miRNA sequence, miRNA accumulation level in an allopolyploid was compared with mid-parent values using F-test for linear contrast.

$$H_0: l = (S+G*S)_{allo,g} - \{((S + G*S)_{At,g} + (S + G*S)_{Aa,g})/2\} = 0.$$

Test static: F = SSC /MSerror, $df_1 = 1$, $df_2 = 27 - 3 = 24$

The type I error rate of 84 tests were adjusted using the false discovery rate (FDR) (Benjamini and Hotchberg). The significance level alpha = 0.05 was chosen for these investigations [129, 161].

Gene expression and DNA methylation data.

Affymetrix expression data for gene expression level in leaf (slide name: ATGE_91_A, ATGE_91_B, ATGE_91_C) and flower (slide name: ATGE_92_A, ATGE_92_B, ATGE_92_C) were downloaded from the AtGenExpress expression atlas

at TAIR http://www.arabidopsis.org/info/expression/ATGenExpress.jsp. Expression value from each microarray experiment was normalized using the GC-RMA method and averaged three replicates [111, 129]. Gene expression data comparing Allo733 and mid-parent value were obtained from the previous study [97].

Small RNA blot analysis and miRNA target validation.

Total RNA was isolated from leaves and flower buds using TRIZOL (Invitrogen) according to manufacturer's instructions. Twenty micrograms of total RNA was separated on a 15% polyacrylamide-urea gel and blotted on Hybond-N+ membranes (Amersham). The probes were made by end-labeling 21- to 24-mer DNA oligonucleotides that corresponded to the antisense strand of microRNAs using T4 polynucleotide kinase. RNA blot analysis was performed using a previously published protocol [162]. RNA ligase-mediated (RLM)-5' RACE analysis of miRNA target genes was performed using the GeneRacer Kit according to the manufacturer's instructions (Invitrogen). The 5' ends of pri-miRNA cDNA were amplified with the GeneRacer primers. PCR products were gel-purified and cloned into pGEM T-easy vector, and individual inserts (5-20) were sequenced to estimate the transcript frequency.

RESULTS

Dynamic changes in small RNA profiles among closely related species

MiRNAs are produced from genetic loci independent of the targets and serve as negative regulators of gene expression by targeting RNA degradation or translational repression [81, 163], while siRNAs are generated from endogenous loci and repeat sequences or from exogenous agents such as viruses and mediate RNA degradation and RNA-directed DNA methylation and chromatin remodeling [159, 160]. To test roles of RNA-mediated pathways in response to polyploidization, we generated ~1.5-million small RNA (sRNA) sequences by pyrosequencing in ten libraries: five from leaves and five from flower buds of *A. suecica*, resynthesized allotetraploids (F_1 and F_7), and their parents *A. thaliana* L*er* and *A. arenosa*. To characterize small RNA population comparable between species and not to include contaminant sequences, only sequences matching perfectly to *A. thaliana* Columbia or Ler genome sequences were used. After removal of adaptor sequences and sequences identical to known cellular RNAs (mRNAs, snoRNAs, pre-tRNAs, rRNAs, and snRNAs) and chloroplast and mitochondrial genomes, 467,589 unique sRNA sequences of 20-25-nt in length representing 28,834 distinct loci in ten libraries were further analyzed (Table 4.1). Many sRNAs originating from *A. arenosa* and *A. suecica*, not perfectly matching to A. thaliana genome were excluded from further analyses because the complete *A. arenosa* genome sequence is unknown.

Table 4.1 List of small RNA sequencing statistics in allopolyploids and their progenitors.

Leaves	**AtL**	**AaL**	**Allo(F_1)L**	**Allo(F_7)L**	**AsL**	**Total**
Total reads	69319	137214	106049	82034	79432	474048
Filtered	7450	10899	4224	1593	5624	29790 (6%)
Unique sRNAs	61869	126315	101825	80441	73808	444258 (94%)
Perfect matches	42489 (69%)	44600 (35%)	42126 (41%)	22462 (28%)	21450 (29%)	173127 (39%)
1-2 mismatches	13346 (22%)	21243 (17%)	15702 (15%)	14722 (18%)	17082 (23%)	82095 (18%)
miRNAs	8930 (14%)	34856 (28%)	27439 (27%)	3619 (4.5%)	6881 (9.3%)	72795 (16%)
tasiRNAs	960 (1.6%)	3 (~0%)	36 (~0%)	41 (~0%)	176 (0.2%)	1216 (0.3%)
Flower Buds						
	AtF	**AaF**	**Allo(F_1)F**	**Allo(F_7)F**	**AsL**	**AsF**
Total reads	69399	73980	81549	76799	87216	388943
Filtered	1901	1598	1821	6817	3064	15201 (4%)
Unique sRNAs	61498	72382	79728	69982	84152	367742 (96%)
Perfect matches	41674 (68%)	7842 (11%)	24040 (30%)	21500 (31%)	28048 (33%)	123104 (33%)
1-2 mismatches	14711 (23%)	10781 (15%)	13582 (17%)	15111 (22%)	14672 (17%)	67857 (18%)
miRNAs	5426 (9%)	3176 (4.3%)	6699 (8.4%)	6587 (9.4%)	10833 (13%)	32721 (9%)
tasiRNAs	170 (0.3%)	4 (~0%)	54 (~0%)	282 (0.4%)	67 (~0%)	577 (0.2%)

In *A. thaliana*, 24-nt sRNAs were most diverse, accounting for 58% of the total distinct sRNA species, while 23-nt, 22-nt, and 21-nt sRNAs represent 18%, 10%, and 9% of total distinct sRNA species, respectively (Figure 4.2A). MiRNAs and ta-siRNAs are usually 20-21-nt in size, whereas siRNAs are generally 23-24-nt in size and derived from repetitive sequences, transposons, intergenic regions, and some genic regions (usually from both strands with a relatively equal frequency) [164]. In *A. thaliana*, an overall sequencing frequency of miRNAs (27%), ta-siRNAs (4%), and siRNAs (68%) was similar to the published data [164-166], indicating a good representation of sRNAs in these small RNA libraries. Although the number of distinct miRNAs was relatively small, the sequence frequency of 21-nt RNAs (mainly miRNAs) was high (~30%), next to that of 24-nt RNAs (~39%).

To test if tissue-specific siRNAs are generated, we compared siRNAs in *A. thaliana* leaves and flowers. Most siRNAs (~96%) were present in either leaves or flowers, and only 4% were found in both tissues (Figure 4.2B), suggesting divergent siRNA populations in two developmental stages [164, 166]. The low overlapping percentage (~4%) of siRNAs in different tissues may correlate with diverse siRNA-generating loci or same genomic regions that produce different siRNAs. To discern these possibilities, we clustered siRNAs that overlapped or in adjacent genomic regions (≤250bp) into a siRNA locus. Among 11,329 siRNA loci (>= 10 reads), ~70% (8,170/11,762) generated siRNAs in both leaves and flowers, and only ~30% (3,592) were found to produce siRNAs in either leaves or flowers (Figure 4.2C), suggesting that common siRNA loci produce diverse siRNA populations during leaf and flower

development. This does not preclude the possibility that some tissue-specific siRNAs are generated from different siRNA loci.

Interestingly, 52-57% siRNA loci of *A. thaliana* origin were found in resynthesized allotetraploids (F_1 and F_7) and *A. suecica* (Figure 4.2D). As those allopolyploids contain both genomes derived from *A.thaliana* and *A.arenosa*, this result suggests stable inheritance of parental siRNAs in allopolyploids. Except for a small amount of siRNAs in coding regions, most siRNAs originated from repeats, transposons, and pseudogenes (Figure 4.2E). Are siRNAs reactivated in repetitive DNA and transposons as a response to "genomic shock" [122] in interspecific hybrids and allopolyploids? The siRNA distribution in *A. thaliana* was consistent with genome-wide density of repeats and transposons (Figure 4.2E). Moreover, siRNA distribution trends were similar in *A. thaliana*, resynthesized allotetraploids, and *A. suecica*, whereas siRNA densities in F_1, F_7, and natural allopolyploids were half of that in *A. thaliana*, consistent with the notion that *A. thaliana* siRNAs are stably inherited in the allopolyploids. A few abundant and conserved siRNA loci of *A. arenosa* and *A. thaliana* origins examined were also maintained in the allotetraploids, but this does not rule out the possibility that some species-specific siRNAs may accumulate differently in allopolyploids. Two siRNA peaks in *A. suecica* and *A. thaliana* were found near chromosome 2 and 4 centromeres, and another siRNA peak in *A. thaliana* existed near the chromosome 5 centeromere (Figure 4.2E). High siRNA densities near the centromeres coincided with dense methylation of centromeric repeats and transposons [32, 39], which may lead to siRNA accumulation

and DNA hypermethylation of *A. thaliana* homoeologous centromeres in *A. suecica* [167].

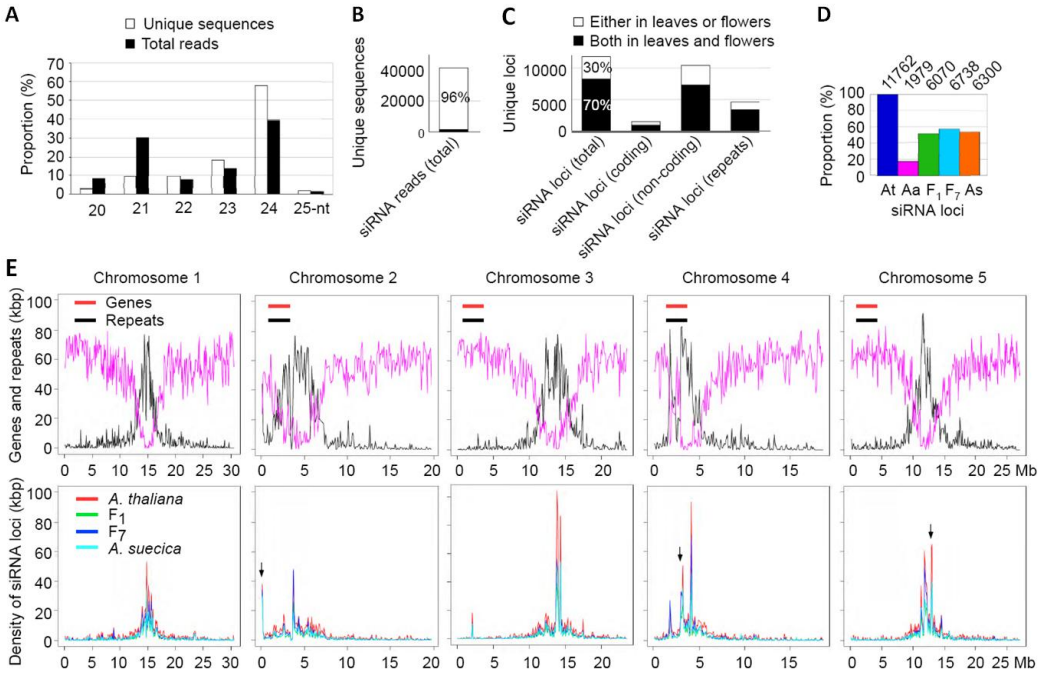

Figure 4.2 sRNAs in allopolyploids and their progenitors
a. Distribution of small RNAs by size in the percentages of unique sequences (unfilled) and total reads (black) in *A. thaliana*. **b**. Total siRNA reads in both leaves and flowers (black) and in either one tissue (unfilled) in *A. thaliana*. **c**. Distribution of siRNA loci in total, coding regions, non-coding sequences, and repeats including transposable elements. **d**. Percentage of *A. thaliana* siRNAs loci in *A. thaliana* (blue), *A. arenosa* (red), F_1 (green), F_7 (cyan), and *A. suecica* (orange). The small number of siRNA loci detected in *A. arenosa* is due to lack of its genomic sequences**. e**. Distribution of genes (red) and repeats including transposable elements (black) in *A. thaliana* in 100kb sliding windows. **f:** Density of siRNAs along five chromosomes of *A. thaliana*. Red, green, blue, and cyan indicate siRNAs in *A. thaliana*, F_1, F_7, and *A. suecica*, respectively. Arrows indicate the peaks that were differently from those in repeats (**e**).

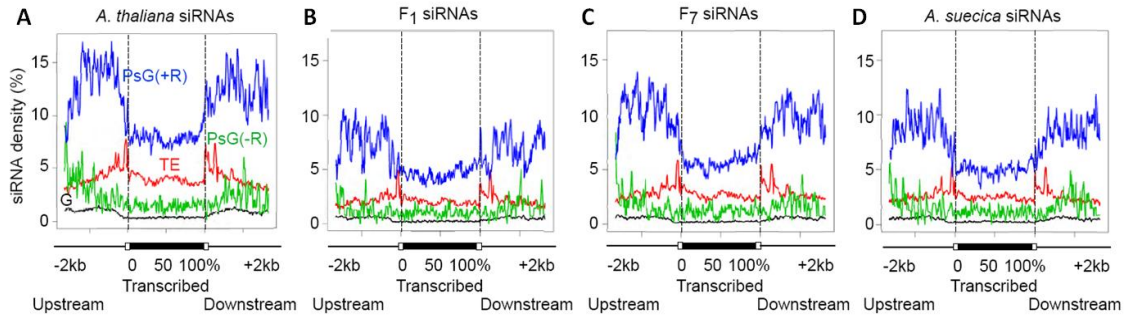

Figure 4.3 Densities of sRNAs around pseudogenes, transposable elements and transcribed genes

A. Distribution of A. thaliana siRNAs in transcribed, upstream, and downstream regions of pseudogenes with repeats (PsG+R, blue) and without repeats (PsG-R, green), transposable elements (TE, red), and all genes (G). **B-D.** The same plots as in a except that allotetraploid F1, F7, and A. suecica siRNAs of A. thaliana origin, respectively, were used. e. Distribution of siRNAs in transcribed, 5' and 3' regions in A. thaliana (black), F1 (green), F7 (blue), and A. suecica (red).

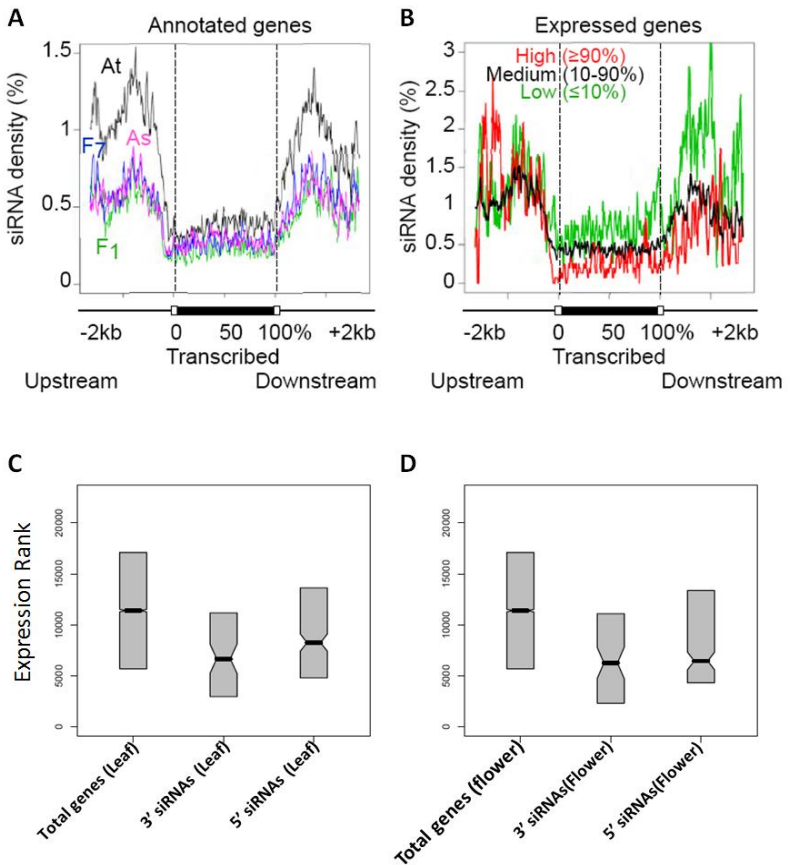

Figure 4.4 Density of siRNAs in transcribed regions and upstream and downstream sequences in allotetraploids and their progenitors

A. Distribution of siRNAs in transcribed, 5' and 3' regions in *A. thaliana* (black), F_1 (green), F_7 (blue), and *A. suecica* (red). **B**. Distribution of *A. thaliana* siRNAs in 5' and 3' regions of the genes that are expressed at high (≥90%, red), medium (10-90%, black), and low (≤10%, green) levels. **C**. Distribution of expression level of siRNA generating genes in *A. thaliana* leaves. Total genes (Leaf), all expressed genes in leaves; 3' siRNAs (Leaf), expressed genes with siRNAs in the 3' end; 5' siRNAs (Leaf), expressed genes with siRNAs in the 5' end. D. Distribution of expression level of siRNA generating genes in *A. thaliana* flowers. Total genes (flower), all expressed genes in flowers; 3' siRNAs (Flower), expressed genes with siRNAs in the 3' end; 5' siRNA (Flower), expressed genes with siRNAs in the 5' end.

Compared to the transcribed genes (G), siRNA accumulated at high levels in the upstream and downstream regions of pseudogenes associated with transposable elements (PsG + R), followed by transposons (TE) (Figure 4.3A-D). However, pseudogenes without repeats (PsG - R) less likely generated siRNAs. This implies that repeat regions are predisposed to siRNA biogenesis. The siRNA distribution trends in these regions were similar in *A. thaliana*, *A. suecica*, and resynthesized allotetraploids (F_1 and F_7), again suggesting stable inheritance of siRNAs during allopolyploid formation. A slightly lower density of siRNAs in the F_1 than F_7 allotetraploid and *A. suecica* (Figure.4.3B-D) indicates that a few generations is required to establish siRNA-mediated chromatin modifications in allopolyploids [168].

The siRNA distribution trends were also similar in the 5' and 3' ends of the transcribed regions in *A. thaliana* and resynthesized and natural allotetraploids (Figure 4.4A). The siRNAs predominated in the 5' and 3' ends with peaks near 1000-bp upstream or downstream of the transcribed regions. Again, siRNA densities in the allotetraploids fell in the middle of *A. thaliana* siRNA plots, and siRNA densities in F_1 allotetraploids were slightly lower than those in F_7 and *A. suecica*.

Do the siRNAs located in the 5' and 3' ends correlate with gene expression? We classified gene expression levels as high (≥90%), medium (10-90%) and low (≤10%) using publically available gene expression microarrays in Arabidopsis leaves [94]. siRNAs accumulated at high levels in the 3' ends of the poorly expressed genes, whereas siRNA levels were low in the highly expressed genes (Figure 4.4B), suggesting a potential role of siRNAs in gene repression via siRNA regulation in the 3' ends. No

correlation between siRNAs and gene expression was found in a previous study [169] probably because different statistical tests and/or samples were used.

To test if the presence or absence of siRNAs affects gene expression level, we identified 679 protein-coding genes that generated siRNAs in genic regions in *A. thaliana* as well as in synthetic allopolyploids and *A. suecica*. The total gene expression values were obtained from published data [94] and normalized to yield similar distributions in both leaves and flowers (Figure 4.3C, D). The genes containing siRNAs in the 5' and 3' ends (within 1,000-bp) were expressed at significantly lower levels in leaves ($P = 2x10^{-7}$ and $3x10^{-7}$, Wilcoxon rank-sum test) and flowers ($P = 9x10^{-12}$ and $1x10^{-7}$ for 5' and 3', respectively) than all genes expressed in the corresponding tissues. Moreover, reduction of gene expression level was more significant in sRNAs generated in 3' downstream region. The data suggest that siRNAs around coding region are associated with low abundance of transcripts in the corresponding tissue. sRNAs generated in 3' down-stream region may be more effective in down-regulating the gene expression as shown in FLC [170].

The majority of siRNA-generating genes (377/679) matched transposable elements or contained transposable elements within introns, and the siRNA-generating genes (~30%) are preferentially methylated (Pearson's Chi-squared test, $\chi^2 = 92$, d.f. = 2, $P = 2.2e^{-16}$). This suggests that siRNAs originate from nearby transposons within the genes, leading to silencing or activation of the genes through siRNA-mediated DNA methylation [93, 160].

The nonadditively expressed genes in the allotetraploids were not associated with siRNAs, which is consistent with few transposons and repetitive elements that are reactivated in the allotetraploids [97]. Only 14 (2.1%) of 1,914 nonadditively expressed genes matched siRNAs in two allotetraploids, which is a significantly low frequency compared with that of additively expressed genes (χ^2 = 32.9074, d.f. = 1, P = $9.7x10^{-9}$). This is probably because transposons are under-represented in the microarrays. Alternatively, rapid changes of siRNAs in the early stages (F_1) of allotetraploids may be sequestered in later generations (>F_6) [97].

Although the distributions of sRNAs are highly conserved in allopolyploids, we found significant decrease of siRNA abundance in allopolyploids. And most of lost siRNAs in F1 were regenerated in F7 or conserved in A.suecica. We detected total 6,012 transposable elements generating sRNAs in at least one species (number of reads >= 10). Majority of them (5123/6000) conserved sRNA generation in A.thaliana and allopolyploids (Figure 4.5A). F1 allopolyploid showed the biggest change of small RNAs around transposable elements. In F1 allopolyploid, 329 transposons lost sRNA generation and only 12 transposons newly generate sRNAs. Majority of transposons that lost sRNA generation regenerate sRNAs in F7 (311/329) or conserved sRNA generation in *A. suecica* (259/329) making sRNA profiles of F7 and A.suecica more similar to A. thaliana than F1 allopolyploid.

To infer sRNA abundance in allopolyploids and progenitor species, we examined sequencing read frequency perfectly matching the *A. thaliana* genome (Figure 4.5B). In *A. arenosa*, sRNAs matching the *A. thaliana* genome are mostly derived from miRNA

loci. This shows high sequence conservation of miRNAs among species. Notably, F1 allopolyploid had a significantly reduced amount of siRNAs from repeat elements making miRNA proportion high (Chi-squared test, Chi-squared = 15683.42, df = 5, P ≈ 0). In F7 allopolyploid, the composition of sRNAs matching repeat elements was higher than F1 but still less than *A. thaliana* (Chi-squared test, P ≈ 0). Abrupt reduction of siRNAs in transposable elements of F_1 allopolyploid may be associated with activation of transposable elements in response to genomic shock as McClintock suggested. Reduced amounts of repeat region matching sRNAs in A.suecica may be due to sequence change in those genomic regions as well as loss of sRNAs.

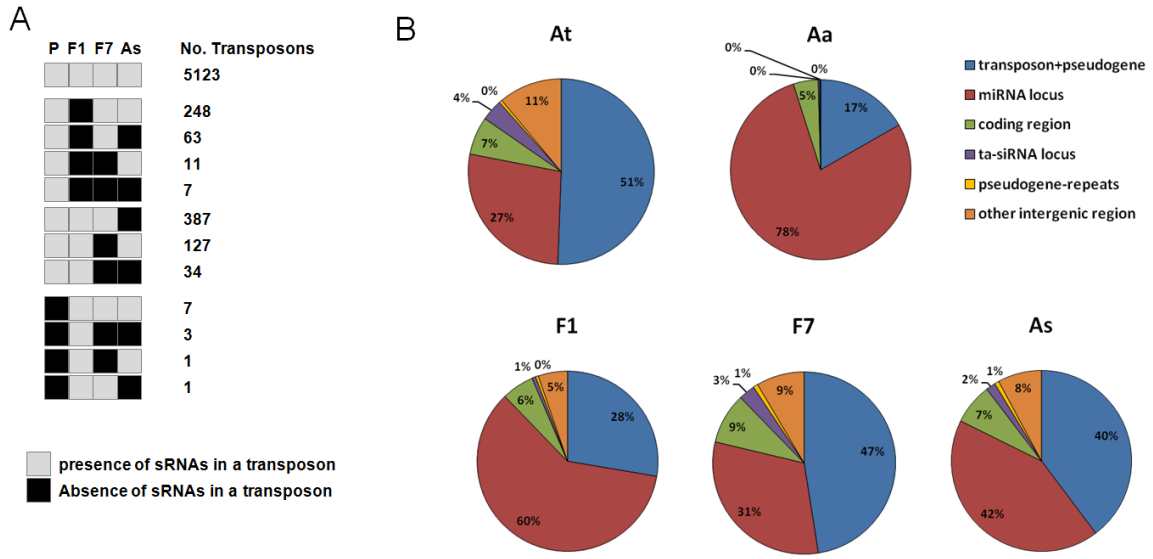

Figure 4.5 Small RNA composition in allopolyploids and their progenitors
A. Preservation, loss and gain of small RNA generation around transposable elements among allopolyploids and their progenitors. Light grey represents presence of small RNAs around a transposon and black represents absence of small RNAs around a transposable elements. **B.** Sequence abundance of A.thaliana genome derived small RNAs in allopolyploids and the progenitors. Pie charts show proportion of sequences perfectly matching repeat elements (blue), miRNA loci (red), coding regions (green), ta-

siRNA loci (violet), pseudogenes not associated with repeat elements (yellow) and other intergenic regions (orange).

Sequence conservation and expression divergence among miRNAs in closely related species.

MicroRNA loci were identified by mapping known miRNAs and hairpin-derived sRNAs [81]. Unlike siRNAs, unique miRNAs were commonly identified in allotetraploids and their progenitors (Figure 4.6), suggesting conservation of miRNA sequences and their roles in gene expression and development. The majority of miRNA loci (153) sequenced in *A. thaliana* were also present in *A. arenosa* (107), allotetraploid F_1 (114) and F_7 (127), and *A. suecica* (121). Except for miR163, the same miRNAs had identical mature sequences in *A. thaliana*, *A. arenosa*, and allotetraploids (Figure 4.6). A few miRNAs undetected in allotetraploids and *A. arenosa* may be species-specific or of low abundance. Similarly, the majority of ta-siRNAs were present in the allotetraploids and their progenitors, and a few ta-siRNAs were not found in *A. arenosa* and the allotetraploids.

Despite sequence conservation, miRNA expression levels were highly variable in *A. thaliana*, *A. arenosa*, *A. suecica*, and the allotetraploids (Figure 4.7). Among 85 distinct miRNAs and 23 ta-siRNAs from 6 loci spotted on the microarrays, 69 miRNAs and 17 ta-siRNAs were expressed above the detection level and confirmed by the sequencing data in *A. arenosa*, *A. suecica*, and resynthesized allotetraploids. In leaves, 35 (~51%) distinct miRNAs and 8 ta-siRNA from 2 ta-siRNA loci showed differential expression between *A. thaliana* and *A. arenosa* or between an allotetraploid (Allo733F_7

or *A. suecica*) and MPV, while in flowers 33 (~40%) miRNAs were differentially expressed in these comparisons. Interestingly, miRNA accumulation levels were nonadditive, which is reminiscent of nonadditive expression of many protein-coding genes in the resynthesized allotetraploids [97]. For example, 12 miRNAs were expressed at higher levels in *A. thaliana* than in *A. arenosa*, and many were down-regulated in F_7 allotetraploids. Likewise, 18 miRNAs were expressed at higher levels in *A. arenosa* than in *A. thaliana*, and the majority of these miRNAs remained highly expressed in the allotetraploids. The data suggest an expression dominance of *A. arenosa* miRNAs over *A. thaliana* in the resynthesized allotetraploids, a direction consistent with repression of *A. thaliana* rRNA genes [171] and many protein-coding genes in the allopolyploids [97].

Many miRNAs showing expression variation in allotetraploids and their progenitors are expressed differently in leaves and flowers, suggesting a general role for miRNAs in tissue-specific expression and development. miRNA abundance in flower buds varied dramatically between the two species with different flower morphologies. *A. thaliana* is inbreeding and has small and white flowers, while *A. arenosa* is outcrossing and has large and pink flowers. The data suggest a role of miRNA regulation in flower morphology between two closely related species. Differential miRNA accumulation between resynthesized and natural allotetraploids may suggest an evolutionary role of miRNAs in growth and development in response to physiological changes and environmental cues.

Among the miRNAs and ta-siRNAs examined, the expression levels estimated from microarrays and sequencing data closely matched those of small RNA blot analysis.

The miRNA frequencies detected by sequencing within one species represented a significant correlation with values of microarray. However, it is not reasonable to compare sequencing frequency among A.thaliana and A.arenosa to infer difference of miRNA accumulation level. Because, we cannot normalize small RNA sequencing frequency in A.arenosa without correctly filtering out contaminant sequences and without knowing genome sequences. The 24-nt miR163, a recently evolved miRNA [172], was abundant in *A. thaliana* but undetectable in *A. arenosa* leaves. In *A. arenosa* flower buds, a 23-nt RNA was detected but at a level 30-fold lower than in *A. thaliana*, suggesting sequence divergence and expression diversity of this young miRNA. In leaves miR159 and 403 were expressed at higher levels in *A. arenosa* than in *A. thaliana* and allotetraploids. Ta-siR255 accumulated at higher levels in *A. arenosa*, F_3 and natural allotetraploids than other lines. Interestingly, ta-siR255 and miR159 expression levels altered in two independent F_1 lines and also changed during selfing (F_3) and in *A. suecica*. The data suggest that intergenomic interactions, genetic segregation, and evolutionary force play roles in miRNA expression in interspecific hybrids and allopolyploids. Although expression variation of miRNAs and ta-siRNA in flowers was generally low, miR156 and miR171 accumulated at higher levels in the flowers than in the leaves, indicating a role for these miRNAs in flower development. Except for miR156 that showed a positive correlation between ploidy and expression levels, miRNA abundance was generally inversely correlated with ploidy levels in *A. thaliana* isogenic diploids (At2) and autotetraploids (At4), suggesting a dosage dependent mechanism for miRNA accumulation in polyploids [1].

miRNAs and target regulation in closely related species and allopolyploids

Do changes in miRNA abundance affect nonadditive expression of target genes in allopolyploids? Among 10 miRNA target genes that are nonadditively expressed in both allopolyploids relative to the progenitors, log-fold changes in miRNA and target expression levels showed a significant negative correlation ($r = -0.764$ and $P = 0.01$, $N = 10$) (Figure 4.8A). Based on the correlation analysis, nonadditive accumulation of miRNAs explains ~58% of non-additive expression of the target genes in allopolyploids. This suggests that differential regulation of miRNAs plays a role in interspecies variation of protein-coding gene expression. Interestingly, the majority (9/10) of nonadditively expressed miRNA targets were down-regulated in allotetraploids, which represents a trend that 909 out of 1187 nonadditively expressed genes are down-regulated [97] ($P = 0.5326$).

To test miRNA function in allotetraploids, we analyzed cleavage sites and frequency of the targets that are nonadditively expressed and negatively correlated with miRNAs. Target binding sites are much conserved in *A. thaliana* and *A. arenosa*, consistent with high conservation of miRNA target sequences in plants [173]. Among six miRNA targets analyzed, the cleavage sites were identical in two related species (Figure 4.8B). These show that many miRNA and their target sequences are highly conserved among close species.

Despite high sequence conservation of miRNAs and miRNA targets, miRNAs tend to change accumulation level between progenitors and in allopolyploids. Moreover, miRNA targets change expression level in negative correlation with change of miRNA

accumulation. Therefore high conservation and change of accumulation level of miRNAs can be one mechanisms causing expression variation of protein coding genes among allopolyploids and their progenitors.

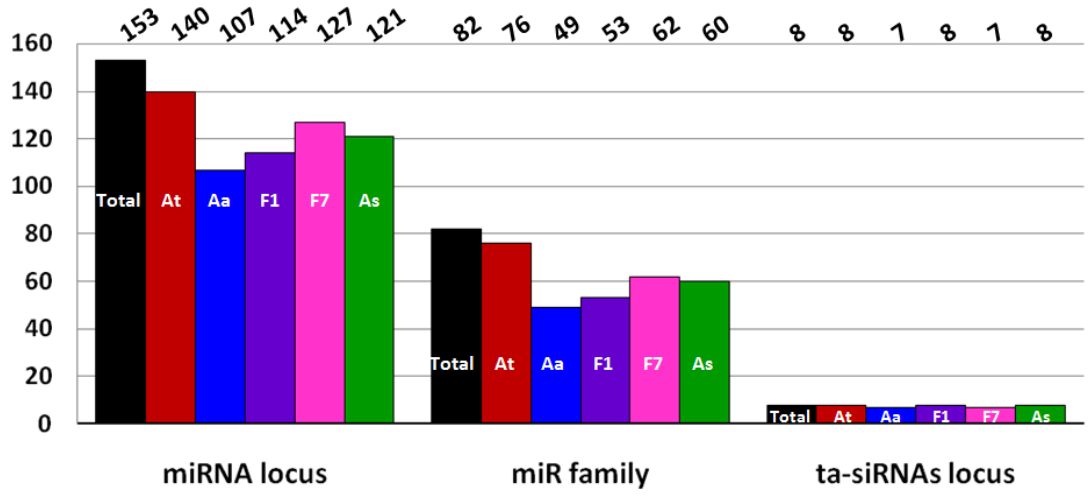

Figure 4.6 Conservation of miRNA and ta-siRNA loci among allopolyploids and their progenitors
Number of distinct miRNAs and ta-siRNA loci identified in total 10 libraries (black), *A. thaliana* (red), *A. arenosa* (blue), F_1, (violet), F_7 (pink), and *A. suecica* (green).

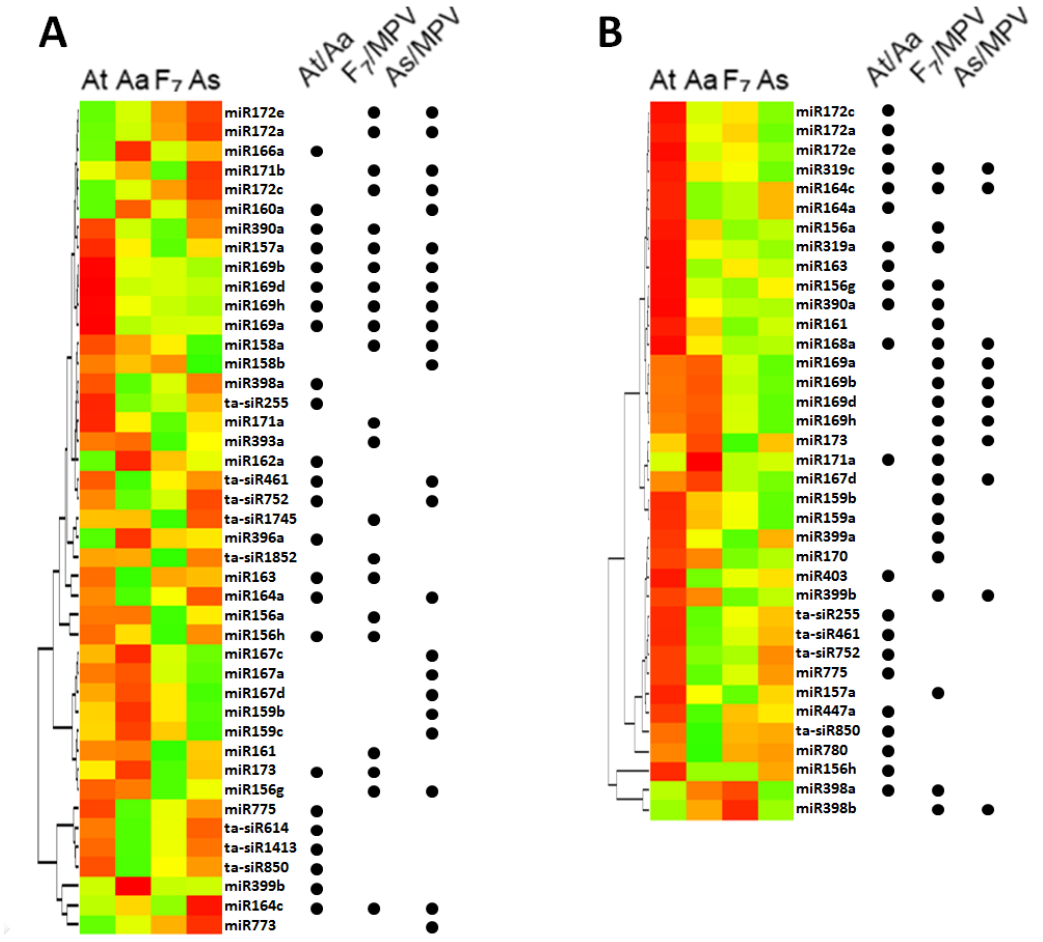

Figure 4.7 Change of miRNA expression levels among allopolyploids and the progenitors

A. Hierarchical cluster analysis of miRNA expression variation in leaves in *A. thaliana* (At), *A. arenosa* (Aa), resynthesized allotetraploid (F_7), and *A. suecica* (As). Expression intensities were standardized into the z score (z_{ij}) for each miRNA i in species j as

$$z_{ij} = \frac{(x_{ij} - \overline{x_i})}{\sqrt{s_{ii}}}$$

where x_{ij} is an expression value prior to standardization; $\overline{x_i}$ is a mean expression value of gene i; and s_{ii} is a standard deviation of miRNA i expression across the species (1…4).. Black, open, and grey circles in the heat maps indicate the targets that were expressed differently between *A. thaliana* and *A. arenosa*, mid-parent value (MPV) and F_7, and MPV and *A. suecica*, respectively. **B**. Hierarchical cluster analysis of miRNA expression variation in flower buds. The colors, symbols, and lines were the same as in **a**.

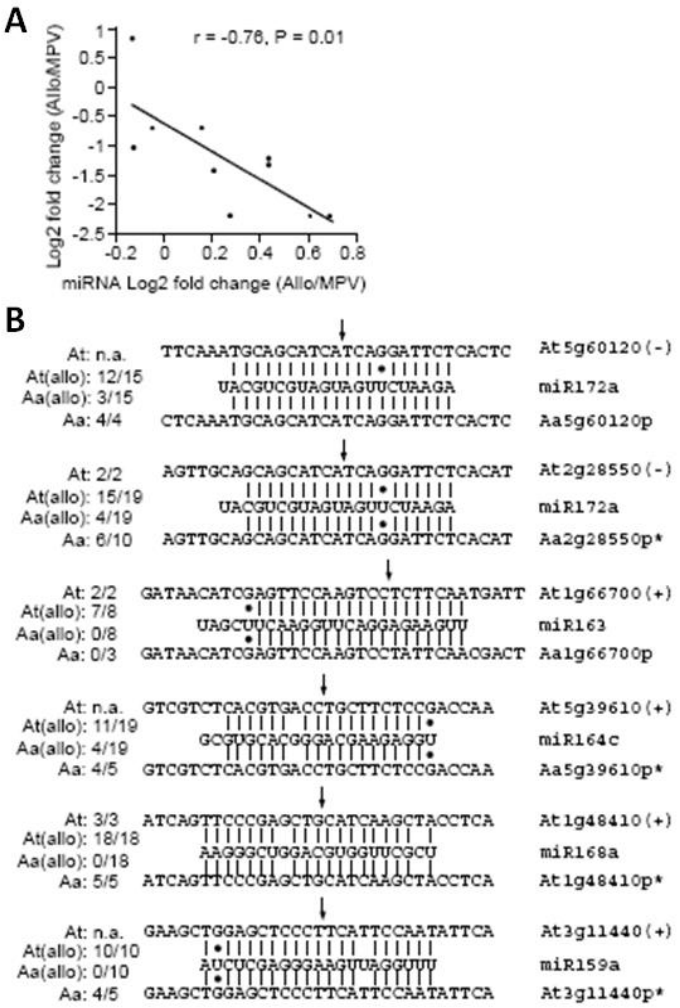

Figure 4.8 Negative correlations between miRNA and their target expression changes in allopolyploids and validation of miRNA targets in allopolyploids
A. Inverse expression correlation between miRNAs and their targets that are nonadditively expressed in allotetraploid Allo733. **B.** miRNA preference of A. thaliana or A. arenosa targets in the seventh generation of an allotetraploid (Allo733F_7). The same locus identity for A. thaliana was used for A. arenosa, except that a "p" was added at the end of A. areonsa locus to indicate "putative". The nominators indicate the number of sequences that matched corresponding loci in A. thaliana (At), A. arenosa (As), A. thaliana homoeologs in allotetraploid [At(allo)], and A. arenosa homoeologs in allotetraploid [Aa(allo)], respectively. n.a.: not analyzed in this study. Arrows indicate cleavage sites, while asterisks indicate A. arenosa targets that perfectly matched A. thaliana ones in the region shown. Dots indicate wobble base pairs between U and G.

DISCUSSION

A role for miRNAs and siRNAs in gene expression diversity and genome stability

The data suggest roles of siRNAs and miRNAs in genomic stability and gene expression diversity in closely related species and allopolyploids. Repeat-associated endogenous siRNAs diverged rapidly among closely related species [160]. These siRNAs are associated with the genes that are constitutively repressed but do not play a role in gene expression changes within and between species. However, siRNAs are directly related to genomic stability and centromere function [160]. Active biogenesis of siRNAs from transposons and repeats is essential for RNA-mediated DNA methylation and chromatin modifications, a vicious cycle for the establishment and stable maintenance of heterochromatin and centromeres. Low siRNA accumulation levels in F_1 but not in F_7 and natural allotetraploids may suggest an association with the high levels of centromeric disjunction and infertility during early stages of allotetraploid formation [146]. It is conceivable that loss of siRNAs in F_1 allotetraploids leads to genomic instability and lethality, whereas in genetically stable allotetraploids siRNA production and heterochromatin formation are well maintained.

MiRNAs, on the contrary, are conserved in sequence but expressed at different levels between the closely related species and in allotetraploids. In fish, several conserved miRNAs such as miR-454a, miR-145, and miR-205 displayed spatial expression differences between two closely related species, medaka and zebrafish [90]. The spatial

and temporal regulation of conserved miRNAs may play an important role in shaping developmental and physiological changes during animal evolution [91]. In Arabidopsis, many miRNAs and their targets are expressed differently between closely related species. This nonadditive accumulation of miRNAs is consistent with nonadditive expression of many protein-coding genes in resynthesized allotetraploids [97]. The miRNAs that are highly expressed in *A. thaliana* is repressed in the resynthesized allotetraploids, a direction consistent with silencing of *A. thaliana* rRNA genes [174] and many protein-coding genes in the allotetraploids [97]. The repression of nonadditively expressed targets of *A. thaliana* origin is associated with preference of miRNAs for *A. thaliana* targets over *A. arenosa* targets. This suggests a miRNA-dependent mechanism for repression of *A. thaliana* genes in interspecific hybrids and allopolyploids. Finally, some miRNAs that were repressed in new allotetraploids were expressed at high levels in *A. suecica*. The differential accumulation of miRNAs between resynthesized and natural allotetraploids may suggest a role of miRNAs in allopolyploid evolution.

Collectively, our data suggests a model that explains nonadditive expression of target genes and phenotypic variation in interspecies hybrids and allopolyploids. MicroRNA loci in different species inherited from the ancestor may diverge in sequence and expression patterns (e.g., tissue-specificity), gain new expression patterns, or undergo gene loss, as a consequence of genetic and epigenetic changes during evolution [1]. Over time, the expression differences are fixed such that the regulatory networks are finely tuned in each species. Combination of two miRNA loci with differential expression patterns will perturb the regulatory balance of miRNAs and their targets in interspecific

hybrids and new allopolyploids. As a result, the accumulation levels of miRNAs is nonadditive [1, 97], leading to nonadditive expression of some targets in the interspecific hybrids and allotetraploids. Although the cause is unknown, preference of miRNAs for degrading *A. thaliana* or *A. arenosa* targets leads to nonadditive expression of targets in allopolyploids [97]. The repression of *A. thaliana* homoeologous loci [97] and accumulation of *A. thaliana*-centromeric siRNAs associated with changes in DNA methylation [167] may be similar to the repression of transposons through maternal transmission of endogenous siRNAs in *Drosophila virilis* [175]. Interspecific hybrids and allotetraploids can only be produced using *A. thaliana* as the maternal parent [54, 168]. Many miRNA targets encode transcription factors or proteins that are important to growth and development in animals and plants [69, 81, 83, 176]. For example, miR164 is responsible for cell patterning and organ boundaries [177, 178]; TAS3 mediates transition from juvenile to adult development [88]; and miR168 and miR403 are predicted to be part of feedback regulation in miRNA biogenesis [164, 179]. Differential accumulation of these miRNAs may lead to physiological and developmental differences in closely related species and interspecific hybrids and allopolyploids in plants and animals.

REFERENCES

1. Chen, Z.J., *Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids.* Annual Review of Plant Biology, 2007. **58**(1): p. 377-406.
2. Bowers, J.E., et al., *Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.* Nature, 2003. **422**: p. 433-438.
3. Blanc, G., K. Hokamp, and K.H. Wolfe, *A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome.* Genome Res., 2003. **13**: p. 137-144.
4. Kellis, M., B.W. Birren, and E.S. Lander, *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.* Nature, 2004. **428**(6983): p. 617-624.
5. *The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.* Nature, 2007. **449**(7161): p. 463-467.
6. De Bodt, S., S. Maere, and Y. Van de Peer, *Genome duplication and the origin of angiosperms.* Trends Ecol. Evol., 2005. **20**: p. 591-597.
7. Jaillon, O., *Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.* Nature, 2004. **431**: p. 946-957.
8. Aury, J.M., *Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia.* Nature, 2006. **444**: p. 171-178.
9. Maere, S., *Modeling gene and genome duplications in eukaryotes.* Proc. Natl Acad. Sci. USA, 2005. **102**: p. 5454-5459.
10. Scannell, D.R., et al., *Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts.* Nature, 2006. **440**: p. 341-345.
11. Gu, Z., et al., *Role of duplicate genes in genetic robustness against null mutations.* Nature, 2003. **421**: p. 63-66.
12. Aravin, A.A., et al., *Developmentally regulated piRNA clusters implicate MILI in transposon control.* Science, 2007. **316**: p. 744-747.
13. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.
14. Lynch, M. and A. Force, *The Probability of Duplicate Gene Preservation by Subfunctionalization.* Genetics, 2000. **154**(1): p. 459-473.
15. Ohno, S., *Evolution by Gene Duplication*. 1970.
16. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes.* Mol Biol Evol, 2003. **20**(9): p. 1377-419.
17. Britten, R.J., *Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(21): p. 13633-13635.

18. The Chimpanzee, S.a.A.C., *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.
19. Cohn, M.J., et al., *Hox9 genes and vertebrate limb specification.* Nature, 1997. **387**(6628): p. 97-101.
20. Belting, H.-G., C.S. Shashikant, and F.H. Ruddle, *Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(5): p. 2355-2360.
21. Wang, J., et al., *Nonadditive Regulation of FRI and FLC Loci Mediates Flowering-Time Variation in Arabidopsis Allopolyploids.* Genetics, 2006. **173**(2): p. 965-74.
22. Gu, X., Z. Zhang, and W. Huang, *Rapid evolution of expression and regulatory divergences after yeast gene duplication.* Proceedings of the National Academy of Sciences, 2005. **102**(3): p. 707-712.
23. Wapinski, I., et al., *Natural history and evolutionary principles of gene duplication in fungi.* Nature, 2007. **449**(7158): p. 54-61.
24. Teichmann, S.A. and M.M. Babu, *Gene regulatory network growth by duplication.* Nat Genet, 2004. **36**(5): p. 492-6.
25. Gu, Z., et al., *Duplicate genes increase gene expression diversity within and between species.* Nat Genet, 2004. **36**(6): p. 577-579.
26. Bomblies, K. and J.F. Doebley, *Pleiotropic Effects of the Duplicate Maize FLORICAULA/LEAFY Genes zfl1 and zfl2 on Traits Under Selection During Maize Domestication.* Genetics, 2006. **172**(1): p. 519-531.
27. Adams, K.L., et al., *Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(8): p. 4649-4654.
28. Ha, M., et al., *Interspecies regulation of microRNAs and their targets.* Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 2008. **In Press, Corrected Proof**.
29. Gidoni, D., W.S. Dynan, and R. Tjian, *Multiple specific contacts between a mammalian transcription factor and its cognate promoters.* Nature, 1984. **312**(5993): p. 409-413.
30. Zhang, Z., J. Gu, and X. Gu, *How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution?* Trends in Genetics, 2004. **20**(9): p. 403-407.
31. Wang, D., et al., *Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors.* Genome Res., 2007. **17**(8): p. 1161-1169.
32. Zilberman, D., et al., *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.* Nat Genet, 2007. **39**(1): p. 61-69.
33. Cubas, P., C. Vincent, and E. Coen, *An epigenetic mutation responsible for natural variation in floral symmetry.* Nature, 1999. **401**(6749): p. 157-161.

34. Finnegan, E.J., W.J. Peacock, and E.S. Dennis, *Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(16): p. 8449-8454.
35. Jullien, P.E., et al., *Maintenance of DNA Methylation during the Arabidopsis Life Cycle Is Essential for Parental Imprinting.* Plant Cell, 2006. **18**(6): p. 1360-1372.
36. Kinoshita, T., et al., *One-Way Control of FWA Imprinting in Arabidopsis Endosperm by DNA Methylation.* Science, 2004. **303**(5657): p. 521-523.
37. Cao, X. and S.E. Jacobsen, *Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing.* Current Biology, 2002. **12**(13): p. 1138-1144.
38. Cao, X., et al., *Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation.* Current Biology, 2003. **13**(24): p. 2212-2217.
39. Cokus, S.J., et al., *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008.
40. Zhang, X., et al., *Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis.* Cell, 2006. **126**(6): p. 1189-1201.
41. Gehring, M., *DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation.* Cell, 2006. **124**: p. 495-506.
42. Jullien, P.E., et al., *Polycomb group complexes self-regulate imprinting of the Polycomb group gene MEDEA in Arabidopsis.* Curr. Biol., 2006. **16**: p. 486-492.
43. Kohler, C. and U. Grossniklaus, *Epigenetic inheritance of expression states in plant development: the role of Polycomb group proteins.* Curr. Opin. Cell Biol., 2002. **14**: p. 773-779.
44. Bernstein, B.E., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**: p. 315-326.
45. Lee, T.I., *Control of developmental regulators by Polycomb in human embryonic stem cells.* Cell, 2006. **125**: p. 301-313.
46. Negre, N., *Chromosomal distribution of PcG proteins during Drosophila development.* PLoS Biol., 2006. **4**: p. e170.
47. Turck, F., et al., *Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27.* PLoS Genetics, 2007. **3**(6): p. e86.
48. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-45.
49. Turner, B.M., *Cellular Memory and the Histone Code.* Cell, 2002. **111**(3): p. 285-291.
50. Cao, X. and S.E. Jacobsen, *Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes.* Proc. Natl Acad. Sci. USA, 2002. **99**(suppl. 4): p. 16491-16498.
51. Costa, S. and P. Shaw, *Chromatin organization and cell fate switch respond to positional information in Arabidopsis.* Nature, 2006. **439**: p. 493-496.

52. Zhou, C., et al., *HISTONE DEACETYLASE19 Is Involved in Jasmonic Acid and Ethylene Signaling of Pathogen Response in Arabidopsis.* Plant Cell, 2005. **17**(4): p. 1196-1204.
53. Chen, Z.J. and C.S. Pikaard, *Transcriptional analysis of nucleolar dominance in polyploid plants: Biased expression/silencing of progenitor rRNA genes is developmentally regulated in Brassica.* Proceedings of the National Academy of Sciences, 1997. **94**(7): p. 3442-3447.
54. Comai, L., et al., *Phenotypic Instability and Rapid Gene Silencing in Newly Formed Arabidopsis Allotetraploids.* Plant Cell, 2000. **12**(9): p. 1551-1568.
55. Lee, H.S. and Z.J. Chen, *Protein-coding genes are epigenetically regulated in Arabidopsis polyploids.* Proc Natl Acad Sci U S A, 2001. **98**(12): p. 6753-6758.
56. Vaughn, M.W., et al., *Epigenetic Natural Variation in Arabidopsis thaliana.* PLoS Biology, 2007. **5**(7): p. e174.
57. Zhang, X., et al., *Global Analysis of Genetic, Epigenetic and Transcriptional Polymorphisms in Arabidopsis thaliana Using Whole Genome Tiling Arrays.* PLoS Genetics, 2008. **4**(3): p. e1000032.
58. Mette, M.F., et al., *Transcriptional silencing and promoter methylation triggered by double-stranded RNA.* Embo Journal, 2000. **19**(19): p. 5194-5201.
59. Herr, A.J., et al., *RNA polymerase IV directs silencing of endogenous DNA.* Science, 2005. **308**: p. 118-120.
60. Kanno, T., *Atypical RNA polymerase subunits required for RNA-directed DNA methylation.* Nature Genet., 2005. **37**: p. 761-765.
61. Onodera, Y., *Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation.* Cell, 2005. **120**: p. 613-622.
62. Pontier, D., *Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis.* Genes Dev., 2005. **19**: p. 2030-2040.
63. Boutet, S., et al., *Arabidopsis HEN1: A Genetic Link between Endogenous miRNA Controlling Development and siRNA Controlling Transgene Silencing and Virus Resistance.* Current Biology, 2003. **13**(10): p. 843-848.
64. Yu, J., et al., *The Genomes of Oryza sativa: A History of Duplications.* PLoS Biol, 2005. **3**(2): p. e38.
65. Baumberger, N. and D.C. Baulcombe, *Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(33): p. 11928-11933.
66. Morel, J.-B., et al., *Fertile Hypomorphic ARGONAUTE (ago1) Mutants Impaired in Post-Transcriptional Gene Silencing and Virus Resistance.* Plant Cell, 2002. **14**(3): p. 629-639.
67. Chen, X., *A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development.* Science, 2004. **303**(5666): p. 2022-2025.
68. Doench, J.G. and P.A. Sharp, *Specificity of microRNA target selection in translational repression.* Genes Dev., 2004. **18**(5): p. 504-511.

69. Llave, C., et al., *Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA.* Science, 2002. **297**: p. 2053-2056.
70. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II.* EMBO J, 2004. **23**(20): p. 4051-60.
71. Lee, Y., et al., *The nuclear RNase III Drosha initiates microRNA processing.* Nature, 2003. **425**(6956): p. 415-9.
72. Park, W., et al., *CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana.* Curr Biol, 2002. **12**(17): p. 1484-95.
73. Yu, B., et al., *Methylation as a crucial step in plant microRNA biogenesis.* Science, 2005. **307**(5711): p. 932-5.
74. Lund, E., et al., *Nuclear export of microRNA precursors.* Science, 2004. **303**(5654): p. 95-8.
75. Han, M.H., et al., *The Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation.* Proc Natl Acad Sci U S A, 2004. **101**(4): p. 1093-8.
76. Vazquez, F., et al., *The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing.* Curr Biol, 2004. **14**(4): p. 346-51.
77. Schwarz, D.S., et al., *Asymmetry in the assembly of the RNAi enzyme complex.* Cell, 2003. **115**(2): p. 199-208.
78. Bao, N., K.-W. Lye, and M.K. Barton, *MicroRNA Binding Sites in Arabidopsis Class III HD-ZIP mRNAs Are Required for Methylation of the Template Chromosome.* Developmental Cell, 2004. **7**(5): p. 653-662.
79. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-854.
80. Lim, L.P., et al., *Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.* Nature, 2005. **433**(7027): p. 769-773.
81. Bartel, D., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**: p. 281-297.
82. Farh, K.K.-H., et al., *The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution.* Science, 2005. **310**(5755): p. 1817-1821.
83. Palatnik, J.F., et al., *Control of leaf morphogenesis by microRNAs.* Nature, 2003. **425**(6955): p. 257-263.
84. Guo, H.-S., et al., *MicroRNA Directs mRNA Cleavage of the Transcription Factor NAC1 to Downregulate Auxin Signals for Arabidopsis Lateral Root Development.* Plant Cell, 2005. **17**(5): p. 1376-1386.
85. Navarro, L., et al., *A Plant miRNA Contributes to Antibacterial Resistance by Repressing Auxin Signaling.* Science, 2006. **312**(5772): p. 436-439.
86. Song, C.-P., et al., *Role of an Arabidopsis AP2/EREBP-Type Transcriptional Repressor in Abscisic Acid and Drought Stress Responses.* Plant Cell, 2005. **17**(8): p. 2384-2396.

87. Borsani, O., et al., *Endogenous siRNAs Derived from a Pair of Natural cis-Antisense Transcripts Regulate Salt Tolerance in Arabidopsis.* Cell, 2005. **123**(7): p. 1279-1291.
88. Yoshikawa, M., et al., *A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis.* Genes Dev., 2005. **19**: p. 2164-2175.
89. Fahlgren, N., et al., *Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA Affects Developmental Timing and Patterning in Arabidopsis.* Current Biology, 2006. **16**(9): p. 939-944.
90. Ason, B., et al., *From the Cover: Differences in vertebrate microRNA expression.* Proceedings of the National Academy of Sciences, 2006. **103**(39): p. 14385-14389.
91. Niwa, R. and F.J. Slack, *The evolution of animal microRNA function.* Current Opinion in Genetics & Development, 2007. **17**(2): p. 145-150.
92. Axtell, M.J. and D.P. Bartel, *Antiquity of MicroRNAs and Their Targets in Land Plants.* Plant Cell, 2005. **17**(6): p. 1658-1673.
93. Chapman, E.J. and J.C. Carrington, *Specialization and evolution of endogenous small RNA pathways.* Nat Rev Genet, 2007. **8**(11): p. 884-896.
94. Schmid, M., et al., *A gene expression map of Arabidopsis thaliana development.* Nat Genet, 2005. **37**(5): p. 501-506.
95. Koch, M.A., B. Haubold, and T. Mitchell-Olds, *Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae).* Mol Biol Evol, 2000. **17**(10): p. 1483-98.
96. Lee, H.-S., et al., *Sensitivity of 70-mer oligonucleotides and cDNAs for microarray analysis of gene expression in Arabidopsis and its related species.* Plant Biotechnology Journal, 2004. **2**(1): p. 45-57.
97. Wang, J., et al., *Genomewide nonadditive gene regulation in Arabidopsis allotetraploids.* Genetics, 2006. **172**(1): p. 507-17.
98. Ohno, S., *Evolution by Gene Duplication*. 1970, New York: Springer-Verlag.
99. Wolfe, K.H. and W.-H. Li, *Molecular evolution meets the genomics revolution.* Nat Genet, 2003. **33**: p. 255-265.
100. Wolfe, K. and D.S. Shields, *Molecular evidence for an ancient duplication of the entire yeast genome.* Nature, 1997. **387**: p. 708-713.
101. Vision, T.J., D.G. Brown, and S.D. Tanksley, *The origins of genomic duplications in Arabidopsis.* Science, 2000. **290**(5499): p. 2114-7.
102. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**: p. 459-473.
103. Gu, Z., et al., *Duplicate genes increase gene expression diversity within and between species.* Nat Genet, 2004.
104. Masterson, J., *Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms.* Science, 1994. **264**: p. 421-424.
105. Blanc, G., K. Hokamp, and K.H. Wolfe, *A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome.* Genome Res, 2003. **13**(2): p. 137-44.

106. Bowers, J.E., et al., *Unravelling angiosperm genome evolution by phylogenetic analysis of chromsomal duplication events.* Nature, 2003. **422**: p. 433-438.
107. Blanc, G. and K.H. Wolfe, *Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution.* Plant Cell, 2004. **16**(7): p. 1679-91.
108. Schmid, M., et al., *A gene expression map of Arabidopsis thaliana development.* Nat Genet, 2005. **37**(5): p. 501-6.
109. Wu, Z., et al. 2004 [cited; A model based background adjustment for oligonucleotide expression assays (http://www.bepress.com/jhubiostat/paper1/)].
110. Ihaka, R. and R. Gentleman, *A language for data analysis and graphics.* J Comput Graphical Statist, 1996. **5**: p. 299-314.
111. Affymetrix. *Statistical Algorithms Description Document* 2002 [cited; Statistical Algorithms Description Document (http://www.affymetrix.com/support/technical/whitepapers.affx)].
112. Xiong, L., K.S. Schumaker, and J.K. Zhu, *Cell signaling during cold, drought, and salt stress.* Plant Cell, 2002. **14 Suppl**: p. S165-83.
113. Seki, M., et al., *Monitoring the Expression Pattern of 1300 Arabidopsis Genes under Drought and Cold Stresses by Using a Full-Length cDNA Microarray.* Plant Cell, 2001. **13**(1): p. 61-72.
114. Berardini, T.Z., et al., *Functional annotation of the Arabidopsis genome using controlled vocabularies.* Plant Physiol, 2004. **135**(2): p. 745-55.
115. Wilcoxon, F., *Individual comparisons by ranking methods.* Biometrics Bulletin, 1945. **1**: p. 80-83.
116. Massey, F.J.J., *The Kolmogorov-Smirnov test of goodness of fit.* Journal of the American Statistical Association, 1951. **46**(253): p. 68-78.
117. Jack, T., *Relearning our ABCs: new twists on an old model.* Trends Plant Sci, 2001. **6**(7): p. 310-6.
118. Pelaz, S., et al., *B and C floral organ identity functions require SEPALLATA MADS-box genes.* Nature, 2000. **405**(6783): p. 200-3.
119. Luan, S., W.S. Lane, and S.L. Schreiber, *pCyP B: a chloroplast-localized, heat shock-responsive cyclophilin from fava bean.* Plant Cell, 1994. **6**(6): p. 885-92.
120. Marivet, J., P. Frendo, and G. Burkard, *Effects of Abiotic Stresses on Cyclophilin Gene-Expression in Maize and Bean and Sequence-Analysis of Bean Cyclophilin cDNA.* Plant Science, 1992. **84**(2): p. 171-178.
121. Seoighe, C. and C. Gehring, *Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome.* Trends Genet, 2004. **20**(10): p. 461-4.
122. McClintock, B., *The significance of responses of the genome to challenge.* Science, 1984. **226**: p. 792-801.
123. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**(1): p. 459-73.
124. Lynch, M., et al., *The probability of preservation of a newly arisen gene duplicate.* Genetics, 2001. **159**(4): p. 1789-804.
125. Williamson, S.H., et al., *Localizing Recent Adaptive Evolution in the Human Genome.* PLoS Genetics, 2007. **3**(6): p. e90.

126. Chen, Z.J., *Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids.* Annu Rev Plant Biol, 2007. **58**: p. 377-406.
127. Wolfe, K.H., *Yesterday's polyploids and the mystery of diploidization.* Nat Rev Genet, 2001. **2**(5): p. 333-341.
128. Kellis, M., B.W. Birren, and E.S. Lander, *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.* Nature, 2004. **428**(6983): p. 617-24.
129. Ha, M., W.H. Li, and Z.J. Chen, *External factors accelerate expression divergence between duplicate genes.* Trends Genet, 2007. **23**(4): p. 162-6.
130. Thomas, B.C., B. Pedersen, and M. Freeling, *Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes.* Genome Res, 2006. **16**(7): p. 934-46.
131. Storchova, Z. and D. Pellman, *From polyploidy to aneuploidy, genome instability and cancer.* Nat Rev Mol Cell Biol, 2004. **5**(1): p. 45-54.
132. Wendel, J.F., *Genome evolution in polyploids.* Plant Mol Biol, 2000. **42**(1): p. 225-49.
133. Soltis, D.E., P.S. Soltis, and J.A. Tate, *Advances in the study of polyploidy since Plant Speciation.* New Phytologist, 2003. **161**: p. 173-191.
134. Leitch, A.R. and I.J. Leitch, *Genomic plasticity and the diversity of polyploid plants.* Science, 2008. **320**(5875): p. 481-3.
135. O'Kane, S., B. Schaal, and I. Al-Shehbaz, *The origins of Arabidopsis suecica (Brassicaceae), as indicated by nuclear rDNA sequences, and implications for rDNA evolution.* Systematic Botany, 1995. **21**: p. 559-566.
136. Simillion, C., et al., *The hidden duplication past of Arabidopsis thaliana.* Proc Natl Acad Sci U S A, 2002. **99**(21): p. 13627-32.
137. Bowers, J.E., et al., *Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.* Nature, 2003. **422**: p. 433-438.
138. Lee, H.S., et al., *Sensitivity of 70-mer oligonucleotides and cDNAs for microarray analysis of gene expression in Arabidopsis and its related species.* Plant Biotechnology Journal, 2004. **2**(1): p. 45-57.
139. Bustamante, C.D., et al., *The cost of inbreeding in Arabidopsis.* Nature, 2002. **416**(6880): p. 531-534.
140. Ranz, J.M., et al., *Sex-dependent gene expression and evolution of the Drosophila transcriptome.* Science, 2003. **300**(5626): p. 1742-5.
141. Gu, Z., et al., *Duplicate genes increase gene expression diversity within and between species.* Nat Genet, 2004. **36**(6): p. 577-9.
142. Zilberman, D., et al., *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.* Nat Genet, 2007. **39**(1): p. 61-9.
143. Wright, M.A. and G.M. Church, *An open-source oligomicroarray standard for human and mouse.* Nat Biotechnol, 2002. **20**(11): p. 1082-3.

144. Kane, M.D., et al., *Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.* Nucleic Acids Res, 2000. **28**(22): p. 4552-7.
145. Chou, C.C., et al., *Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression.* Nucleic Acids Res, 2004. **32**(12): p. e99.
146. Comai, L., A.P. Tyagi, and M.A. Lysak, *FISH analysis of meiosis in Arabidopsis allopolyploids.* Chromosome Res, 2003. **11**(3): p. 217-26.
147. Jakobsson, M., et al., *A unique recent origin of the allotetraploid species Arabidopsis suecica: Evidence from nuclear DNA markers.* Mol Biol Evol, 2006. **23**(6): p. 1217-31.
148. Casneuf, T., et al., *Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana.* Genome Biol, 2006. **7**(2): p. R13.
149. Tirosh, I., et al., *A genetic signature of interspecies variations in gene expression.* Nat Genet, 2006. **38**(7): p. 830-4.
150. Freeling, M. and B.C. Thomas, *Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity.* Genome Res, 2006. **16**(7): p. 805-14.
151. Birchler, J.A., et al., *Dosage balance in gene regulation: biological implications.* Trends in Genetics, 2005. **21**(4): p. 219-226.
152. Birchler, J.A., D.L. Auger, and N.C. Riddle, *In search of the molecular basis of heterosis.* Plant Cell, 2003. **15**(10): p. 2236-9.
153. Wittkopp, P.J., B.K. Haerum, and A.G. Clark, *Evolutionary changes in cis and trans gene regulation.* Nature, 2004. **430**(6995): p. 85-8.
154. Zhang, X., et al., *Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis.* Cell, 2006. **126**(6): p. 1189-201.
155. Tang, H., et al., *Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps.* Genome Res., 2008: p. gr.080978.108.
156. Liao, B.Y. and J. Zhang, *Mouse duplicate genes are as essential as singletons.* Trends Genet, 2007. **23**(8): p. 378-81.
157. Clark, R.M., et al., *Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana.* Science, 2007. **317**(5836): p. 338-42.
158. Patterson, N., et al., *Genetic evidence for complex speciation of humans and chimpanzees.* Nature, 2006. **441**(7097): p. 1103-1108.
159. Baulcombe, D., *RNA silencing in plants.* Nature, 2004. **431**(7006): p. 356-363.
160. Lippman, Z. and R. Martienssen, *The role of RNA interference in heterochromatic silencing.* Nature, 2004. **431**: p. 364-370.
161. Benjamini, Y.H., Yosef, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
162. Tian, L. and Z.J. Chen, *Blocking histone deacetylation in Arabidopsis induces pleiotropic effects on plant gene regulation and development.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(1): p. 200-205.

163. Xuemei, C., *microRNA biogenesis and function in plants.* FEBS letters, 2005. **579**(26): p. 5923-5931.
164. Rajagopalan, R., et al., *A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana.* Genes Dev., 2006. **20**(24): p. 3407-3425.
165. Kasschau, K.D., *Genome-wide profiling and analysis of Arabidopsis siRNAs.* PLoS Biol., 2007. **5**: p. e57.
166. Lu, C., et al., *Elucidation of the Small RNA Component of the Transcriptome.* Science, 2005. **309**(5740): p. 1567-1569.
167. Chen, M., et al., *RNAi of met1 Reduces DNA Methylation and Induces Genome-Specific Changes in Gene Expression and Centromeric Small RNA Accumulation in Arabidopsis Allopolyploids.* Genetics, 2008. **178**(4): p. 1845-1858.
168. Wang, J., et al., *Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids.* Genetics, 2004. **167**(4): p. 1961-1973.
169. Kasschau, K.D. and J.C. Carrington, *A Counterdefensive Strategy of Plant Viruses: Suppression of Posttranscriptional Gene Silencing.* Cell, 1998. **95**(4): p. 461-470.
170. Swiezewski, S., et al., *Small RNA-mediated chromatin silencing directed to the 3â€² region of the Arabidopsis gene encoding the developmental regulator, FLC.* Proceedings of the National Academy of Sciences, 2007. **104**(9): p. 3633-3638.
171. Pikaard, C.S., *The epigenetics of nucleolar dominance.* Trends in Genetics, 2000. **16**(11): p. 495-500.
172. Allen, E., *Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana.* Nature Genet., 2004. **36**: p. 1282-1290.
173. Rhoades, M.W., et al., *Prediction of Plant MicroRNA Targets.* Cell, 2002. **110**(4): p. 513-520.
174. Chen, Z.J., L. Comai, and C.S. Pikaard, *Gene dosage and stochastic effects determine the severity and direction of uniparental ribosomal RNA gene silencing (nucleolar dominance) in Arabidopsis allopolyploids.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(25): p. 14891-14896.
175. Blumenstiel, J.P. and D.L. Hartl, *Evidence for maternally transmitted small interfering RNA in the repression of transposition in Drosophila virilis.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(44): p. 15965-15970.
176. Reinhart, B.J., et al., *MicroRNAs in plants.* Genes Dev., 2002. **16**(13): p. 1616-1626.
177. Mallory, A.C., et al., *MicroRNA Regulation of NAC-Domain Targets Is Required for Proper Formation and Separation of Adjacent Embryonic, Vegetative, and Floral Organs.* Current Biology, 2004. **14**(12): p. 1035-1046.
178. Laufs, P., et al., *MicroRNA regulation of the CUC genes is required for boundary size control in Arabidopsis meristems.* Development, 2004. **131**(17): p. 4311-4322.

179. Vaucheret, H., A.C. Mallory, and D.P. Bartel, *AGO1 Homeostasis Entails Coexpression of MIR168 and AGO1 and Preferential Stabilization of miR168 by AGO1*. Molecular Cell, 2006. **22**(1): p. 129-136.

VITA

Misook Ha was born in 1977 in Cheonan, Korea. In 1996, she studied at Pohang University of Science and Technology (POSTECH), Korea and received a Bachelor of Science degree in 2000. Two years later, she graduated with a Master of Science degree at POSTECH. In February 2002, she started her first job as a bioinformatician at Sejong Bioinformatics Institute in Seoul, Korea. From 2003 to 2004, she worked as a Research Scientist at Korea Research Institute of Bioscience and Biotechnology (KRIBB) in Daejon, Korea. In August 2004, she arrived in the United States to pursue a Ph.D. degree in the Interdisciplinary Genetics program at Texas A&M University. In Spring 2005, she joined the Chen laboratory to study expression evolution of duplicate genes. After the laboratory moved to The University of Texas at Austin in late 2005, she transferred her graduate study from Texas A&M University to The University of Texas at Austin and continued her pursuit for a Ph.D. degree in the Cellular and Molecular Biology program (Bioinformatics and Computational Biology track) under the supervision of Dr. Z. Jeffrey Chen.

Permanent Address: 950-4, Wallha-li, Seo-myon, Yongi-gun, Chungcheong-namdo, 339-840, South Korea

This dissertation was typed by the author.