Evolution of Duplicate Gene Sequences, Expression Patterns, and Functions in the Brassicaceae and Other Rosids

by

Shao-Lun Liu

M.Sc., The National Changhua University of Education, 2004

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2011

Abstract

Duplicated genes are considered as raw materials for evolutionary innovations. They are common in eukaryotic genomes, particularly in plants due to the high incidence of whole genome duplication. Thus, understanding the factors that contribute to the retention of duplicated genes is a fundamental topic in evolutionary biology. I tackle this topic by examining how reciprocal expression (RE) among different organ and tissue types, as well as protein subcellular relocalization (PSR), contributes to the retention of duplicated genes. From analyses of microarray data across 83 different organ/cell types and developmental stages in *Arabidopsis thaliana*, I determined that more than 30% of duplicate pairs showed RE patterns (chapter 2). Reconstructing their ancestral expression pattern, more RE cases resulted from gain of a new expression pattern (neofunctionalization) than from partitioning of ancestral expression patterns (subfunctionalization), with pollen being a common location for expression gain (chapter 2). During the analysis on RE, I found a dramatic example of neofunctionalization for a pair of protein kinase genes, *SSP* and *BSK1*, in the Brassicaceae (chapter 3). *BSK1* and *SSP* have opposite expression patterns in pollen compared with all other parts of the plant. I determined that *BSK1* retains the ancestral expression pattern and function and that the ancestral function of *SSP* was lost by deletions in the kinase domain. I revealed that *SSP* changed its function from a component of the brassinosteroid signaling pathway to being a paternal regulator of embryogenesis. I also found that two reciprocally expressed duplicated gene pairs, a peroxidase gene pair and a *CDPK* gene pair, in Brassicaceae showed PSR and evidence for neofunctionalization (chapter 2). To better understand how PSR can contribute to the retention of duplicated genes, I focused on a particular example for a pair of the chloroplast-origin ribosomal protein S13 (*rps13*) genes in rosids (chapter 4). One encodes chloroplast-imported RPS13 (nucp

rps13), while the other encodes mitochondria-imported RPS13 (numit *rps13*). I provided evidence that numit *rps13* genes have experienced adaptive and convergent evolution. My thesis provides important insights into the evolutionary importance of RE and PSR on the retention of duplicated genes in plants.

Preface

Chapter 2 has been submitted for publication. **Liu, S.-L.**, Baute, G.L., Adams, K.L. (2011) Reciprocal, organ and cell-type-specific expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. I conducted most of the analyses and wrote most of the manuscript. GLB developed the Perl scripts for the bioinformatics analyses. KLA conceived the study, helped with the experimental design, and edited the manuscript.

Chapter 3 has been published. **Liu, S.-L.**, Adams, K.L. (2010) Dramatic change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the Brassicaceae. *Molecular Biology and Evolution* 27: 2817–2828. I conducted all the analyses and lab experiements, and I wrote most of the manuscript. KLA conceived the study, helped with the experimental design, and edited the manuscript.

Chapter 4 has been published. **Liu, S.-L.**, Adams, K.L. (2008) Molecular adaptation and expression evolution following duplication of genes for organellar ribosomal protein S13 in rosids. *BMC Evolutionary Biology* 8: 25 (16 pages). I conducted all the analyses and lab experiments, and I wrote most of the manuscript. KLA conceived the study, helped with the experimental design, and wrote the manuscript.

Similar information is listed in the footnotes on the first pages of these chapters.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Dr. Keith Adams, for the opportunity to pursue one of the most compelling evolutionary questions, “why duplicated genes are important over plant evolution”, and advance my knowledge of evolutionary genomics and phylogenetics in plants. I also would like to thank other members of my committee, Drs. Quentin Cronk, Sean Graham, and Loren Rieseberg, for helpful discussions over the course of my PhD study. I thank my friends, Drs. Michael Barker, William Cheung, Nolan Kane, and Renchao Zhou, for enjoyable discussions of various scientific ideas and helping me learn the programming knowledge with the use of R and Perl, as well as my lab members, Gregory Baute, James Robertson, and other people in Keith Adams’ lab, for the input of their ideas and comments on my PhD researchs. I also thank Huei-Jiun Su and Liang-Chi Wang for collecting papaya pollen and assaying gene expression in papaya pollen, Yu-Ti Cheng and Fang Xu for helping with the *Agrobacterium*-mediated transformation assay, and Xuguang Liu for helping with the use of a confocal laser scanning microscope. My PhD research was funded by a grant from the Natural Science and Engineering Research Council of Canada to Keith Adams. I was partly supported by graduate fellowships from the University of British Columbia. This dissertation is dedicated to my family and friends in Taiwan for their ongoing support.

1 Introduction

Gene duplication is considered to be one of the most important types of genetic variation (reviewed in Dermuth and Hahn 2009; Hasting et al. 2009). Duplicated genes can be formed at a faster rate than other types of mutation (reviewed in Hasting et al. 2009). For instance, the frequency of changes in gene copy number can be two or four orders of magnitude more than for point mutations (reviewed in Lupski 2007). Most new genes evolved from pre-existing genes when considering that the majority of genes in a genome belong to a gene family (Ohno 1970). After formation, duplicated genes can diverge to acquire new expression patterns and functions. In this introduction, I will provide an overview of the following: what kinds of molecular mechanisms can create duplicated genes, why gene duplication is so important over evolution, what are the evolutionary fates of duplicated genes, and the goals of my dissertation that will advance our knowledge of the evolutionary importance of gene duplication over plant evolution.

1.1 Origin and Types of Gene Duplications

Duplicated genes can be created by various molecular mechanisms: 1) a whole genome duplication event (i.e., polyploidization) in which the entire set of chromosomes was duplicated by the fusion of unreduced gametes or endoreplication; 2) a segmental duplication event in which a region of a chromosome was duplicated by chromosome rearrangement or recombination; and 3) a tandem duplication event in which genes were amplified by unequal crossing over during meiosis between misaligned chromosomes.

Among these mechanisms, duplicated genes were largely originated from tandem duplication

and whole genome duplication in various eukaryotes. Likewise, both tandem duplication and whole genome duplication greatly contribute to the increase of genome complexity in plants. It is estimated that about 15% of duplicated genes are derived from tandemly duplicated genes within angiosperm genomes (Rizzon et al. 2006). In terms of ancient whole genome duplication events, several studies have shown that more than 90% of angiosperms have undergone at least one round of ancient whole genome duplication (i.e., paleopolyploidization) (e.g., Blanc and Wolfe 2004a; Sterck et al. 2005; Cui et al., 2006; Barker et al. 2008; Barker et al. 2009; Schmutz et al. 2010; Shi et al. 2010; Tang et al. 2010) and a recent study indicates that a polyploidy event occurred in the common ancestor of all angiosperms (Jiao et al. 2011).

1.2 Evolutionary Importance of Duplicated Genes

Since gene duplication played an important role in the creation of additional genes in a genome, it has been considered as an important molecular mechanism to provide the raw genetic sources for the morphological and physiological innovation over organismal evolution (e.g., Ohno 1970; Zhang 2003). The innovation of these traits (i.e., gain of a new function) would help organisms to adapt to adverse environments or occupy new niches (e.g., Monson 2003; Benderoth et al. 2006; Swingley et al. 2008). In addition, gene duplication, particularly whole genome duplication, can serve as an important molecular mechanism for speciation that potentially contributed to species diversification during eukaryotic evolution (reviewed in Van de Peer et al. 2009). Below I will elaborate on phenotypic consequences of gene duplication, as well as its contribution to speciation.

1.2.1 Phenotypic Consequences of Gene Duplication

To illustrate the phenotypic consequence of gene duplication, I herein provide two examples that shed light on how gene duplication contributes to changes in flowering time in plants. These two examples help illustrate why gene duplication has been so important during the evolution of plants.

Two recently published studies showed the effect of gene duplication on delay in flowering time. Interestingly, both examples involved the duplication of the Flowering Locus T (*FT*) gene, which belongs to the phosphatidylethanolamione-binding (*PEPB*) protein family. The first example is from a study of domesticated sunflower - *Helianthus annus* (Blackman et al. 2010). In their study, they identified four tandemly arrayed *FT* genes in both wild and domesticated sunflowers and named them *HaFT1* to *HaFT4*. Based on functional assays, *HaFT3* lost its expression and became a pseudogene. Both *HaFT2* and *HaFT4* retain the ancestral expression patterns and functions, where they show transcription and translation in leaves, and the proteins then are transferred to the shoot apex through the stem, similar to the orthologus *FT* gene in *Arabidopsis*. In contrast, *HaFT1* lost its ancestral expression pattern in leaves and acquired a new expression pattern in shoot apexes. In addition, *HaFT1* in domesticated sunflower populations possesses a frameshift mutation (TG→C) in the third exon that results in a new domain at the C-terminal end of the protein. The framshift mutation in the third exon in the *HaFT1* of domesticated sunflower is not seen in wild sunflower populations (Blackman et al. 2010). From transgenic assays, the authors demonstrated that the delay in flowering time in the domesticated sunflower results from the dominant-negative interference of the domesticated allele of *HaFT1* with an activating paralog, *HaFT4*.

A second example was recently reported by Pin et al. (2010) in cultivated, biennial beets (*Beta vulgaris* ssp. *vulgaris*). Cultivated, biennial beets are unable to have reproductive shoots unless the plants experience a period of cold treatment over winter, referred to as vernalization. Compared with orthologous genes in *Arabidopsis*, there are two copies of *FT* genes in cultivated beets. These two *FT* genes are located on different chromosomes, but the type of duplication event that created them is not known. The authors found that these two *FT* genes (named *BvFT1* and *BvFT2*) are highly associated with the regulation of flowering time in cultivated beets. Both copies are predominantly expressed in leaves. Although this expression pattern is similar to their ancestral expression pattern, expression of these two copies diverges in a temporal manner. *BvFT1* is expressed in the leaves of plants that are not competent to flower (i.e., plants grown in a short-day photoperiod), whereas *BvFT2* is expressed in the leaves of plants that are competent to flower (i.e., plants grown in a long-day photoperiod) (Pin et al. 2010). In addition to the cues of photoperiods, vernalization can suppress the expression of *BvFT1* (Pin et al. 2010). Without vernalization, *BvFT1* can still show expression in a long-day photoperiod. Antagonistic functions were found in this duplicated pair where *BvFT1* suppresses the flowering and *BvFT2* promotes flowering, indicating that the function of *BvFT2* is more like the ancestral state (as mentioned above in the domesticated sunflower example) and *BvFT1* acquired a new function after gene duplication (Pin et al. 2010). From sequence similarity analysis, *BvFT1* differs from *BvFT2* by three amino acid changes in the fourth exon in an external loop of the *PEPB* protein. These three amino acid mutations in *BvFT1* result in a functional shift from flowering promoter to flowering repressor (Pin et al. 2010). Thus, the suppression effect of flowering will be relieved after vernalization because the expression of *BvFT1* was suppressed and expression of *BvFT2* was induced (Pin et al. 2010).

1.2.2 Speciation

In addition to the innovation of morphological traits, gene duplication can serve as an important molecular mechanism for speciation in plants (Werth and Windham 1991; Lynch and Force 2000a; Taylor et al. 2001). After gene duplication, the functionally redundant duplicated genes can be reciprocally lost in two isolated populations. For example, copy one is retained by population1 and copy two is retained by population 2. Such a reciprocal gene loss (RGL) pattern will result in lowered hybrid fitness in some F2 individuals between two populations (see a nice illustration in figure 1 of Sémon and Wolfe 2007), and multiple RGLs could result in reproductive isolation between two populations (e.g., Scannell et al. 2006). The frequency of speciation by RGL can be higher when duplicated genes are located on different chromosomes or are separated by a long distance within a chromosome such that they can be easily segregated in F2 hybrids between two isolated populations. Based on this hypothesis, gene duplication is expected to promote speciation.

Recently, there are three nice examples supporting the speciation by RGL in plants. The first example is a pair of histidinol-phosphate transaminase genes (AT1G71920 and AT5G10330) in *Arabidopsis thaliana* (Bikard et al. 2009). This duplicated gene pair arose from a dispersed duplication event where AT1G7920 is located on chromosome 1 and AT5G10330 is located on chromosome 5. The knock-out phenotype of AT5G10330 in *Arabidopsis thaliana* ecotype Columbia has been shown to be embryo-lethal (Tzafrir et al. 2004), suggesting that the function of these genes is important for embryo development. Functional analysis showed that AT1G71920 lost its expression in ecotype Columbia-0 (Col), whereas AT5G10330 had a deletion in its protein sequence that disrupts the function in ecotype Cape Verde Island Cvi-0

(Cvi) (Bikard et al. 2009). Due to genetic redundancy, the loss of each copy is compensated by the other copy in both ecotypes of *Arabidopsis thaliana*. As expected, the authors found a hybrid incompatibility when they crossed these two ecotypes.

The second example, reported by Yamagata et al. (2010), is a pair of nuclear-encoded mitochondrial ribosomal protein L27 (*mtRPL27*) genes in both domesticated and wild rice. The function of *mtRPL27* gene is essential for the late developmental stage of pollen (Yamagata et al. 2010). This duplicated gene pair is derived from a segmental duplication event in the common ancestor of the AA genome of rice and the genes are located on chromosome 4 and chromosome 8. (Although the authors showed that there was a subsequent tandem duplication of *mtRPL27* on chromosome 8 of domesticated rice, I introduce these two duplicated genes as a whole unit for simplifying the story.) Analysis of these two duplicated genes in domesticated rice (*Oryza sativa* ssp. *japonica* cultivar variety T65) and wild rice (*O. glumaepatula*) revealed that both copies still remain in domesticated rice, whereas only one copy was found on chromosome 4 of wild rice and the other copy was missing due to a deletion event. Furthermore, the authors demonstrated that the loss of function for the copy on chromosome 4 of domesticated rice is due to the failure of expression. Therefore, hybrid incompatibility was found when these two species were crossed.

The last example is reciprocal gene loss of a pair of small plant-specific proteins, *DOPPELGANGER1* (*DPL1*) and *DOPPELGANGER2* (*DPL2*), between two subspecies, *Oryza sativa* sub. *japonica* cultivar Nipponbare and *O. sativa* sub. *indica* cultivar Kasalath, in rice (Mizuta et al. 2010). The disruption of *DPL* genes would result in the failure of pollen germination. This duplicated gene pair was derived from a recent small duplication event, occurring after the speciation event between *Oryza* and *Brachypodium*. Based on the map-based cloning of these two duplicated *DPL* genes, *DPL1* is located on chromosome 1 and *DPL2* is

located on chromosome 6 in these two rice subspecies. From the gene structure and expression analysis, *DPL1* of *O. sativa* sub. *indica* had a 518 bp insertion by a transposable element in the coding region and showed no evidence of expression, whereas *DPL2* of *Oryza sativa* sub. *japonica* remained expressed but had a nucleotide substitution from A to G in the second intron that led to the loss of splicing and an intron-retained transcript with a premature stop codon, resulting in the loss of its function. Due to such a reciprocal gene loss between these two subspecies, the hybrid incompatibility was observed when crossing these two subspecies (Mizuta et al. 2010). From these studies, it is clear that postzygotic reproductive barriers by RGL can happen between two different populations within a species, two different subspecies within a species, or two different species. In addition, the origin of a gene duplication that contributes to speciation by RGL can be either small-scale duplication or large-scale duplication.

How frequent does gene duplication contribute to plant speciation? It is less known how frequent small-scale gene duplications such as tandem duplication contributes to speciation. In constrast, the speciation process by whole genome duplication can be readily accessed by investigating the evolution of chromosome number along a phylogeny (e.g., Otto and Whitton 2000; Wood et al. 2009). Wood et al. (2009) recently estimated the frequency of speciation events by polyploidization (i.e., whole genome duplication) based on the cytological and phylogenetic data in ferns and angiosperms. Based on their results, 15% of the speciation events in angiosperms and 31% in ferns were found to be associated with polyploidization, indicative of a high frequency of polyploid speciation in angiosperms and ferns.

1.3 Evolutionary Fates of Duplicated Genes

Although I introduced the evolutionary fate of duplicated genes in the context of gain of new functions in previous sections, there are actually several possible fates of a newly duplicated gene. Most frequently one member of a newly formed gene pair is lost (Walsh 1995). This is because the rate of deleterious mutations is much higher than beneficial mutations (Kimura 1983), and if the duplicated genes act redundantly one copy is free to accumulate deleterious mutations and lose its function. The majority of duplicated genes therefore will be expected to revert to single copy status. Computational simulation indeed supports such a view that the rate of gene loss after duplication is at least an order of magnitude higher than gene divergence (e.g., Ohta 1987; Walsh 1995). From empirical observations, the half-life of duplicated genes may only be about 3-7 Myr in eukaryotes (Lynch and Conery 2000), suggesting that gene loss after gene duplication is the most common evolutionary fate for duplicated genes over a short period of evolutionary time. Although the reversion to single copy after gene duplication is expected, there are still numerous duplicated genes retained in most eukaryotic genomes, particularly in plants. Several different models that attempt to explain why and how duplicates are retained have been proposed in the literature. Here I will introduce some of the most popular models: 1) neofunctionalization; 2) subfunctionalization; 3) subneofunctionalization; and 4) gene dosage balance. More thorough discussion of possible evolutionary routes to duplicate gene retention can be found in some recent reviews, e.g., Sémon and Wolfe (2007), Conant and Wolfe (2008), Hahn (2009), and Innan and Kondrashov (2010).

1.3.1 Neofunctionalization

The concept, gain of new function (i.e., neofunctionalization) after gene duplication, was developed by Ohno (1970) in his book "Evolution by gene duplication". The neofunctionalization model anticipates the gain of a new function by one gene in a duplicated pair. In this model, the duplicated genes are expected to have redundant functions so that one copy is free to accumulate mutations without affecting the gene's ancestral role. This copy may be free to have a higher rate of mutation accumulation in comparison to its duplicated partner, referred to as relaxation of purifying selection, or accumulate beneficial mutations to adaptively acquire a new function, referred to as neofunctionalization by an adaptive process. As discussed below with a few examples, this process may happen by completely neutral or nearly neutral processes (Dykhuizen–Hartl effect), or by an adaptive process (positive Darwinian selection).

1.3.1.1 Dykhuizen–Hartl Effect (Neutral Process)

In this scenario, a new function arises as a by-product of the accumulation of mutations in a *cis*-regulatory region or protein coding region because one copy of a duplicated gene pair is free to accumulate mutations by relaxation of purifying selection. The fixation of these selectively neutral mutations could be caused by stochastic fluctuation in gene frequency (i.e., genetic drift), which is a common process in eukaryotes with small population sizes. Although these mutations that result in gain of a new function were fixed into the population by non-adaptive mechanisms, this new function might be merely non-adaptive, referred to as the Dykhuizen-Hartl effect (Dykhuizen and Hartl 1980; Zhang et al. 1998). Thus, there is no need for positive Darwinian selection to fix the acquired new function in the population. One possible example is a pair of

transcription factor II A γ genes, *TFIIAγ1* and *TFIIAγ5*, in rice (Sun and Ge 2010). *TFIIAγ* is a small subunit of transcription factor II A that is required by RNA polymerase II for transcription. *TFIIAγ1* and *TFIIAγ5* originated from an ancient whole genome duplication event, which occurred about 70 Mya. From protein sequence rate analysis, *TFIIAγ1* evolved two times faster than *TFIIAγ5*. But, there is no evidence of positive selection for any codons in *TFIIAγ1* based on codon-based positive selection analysis using the evolutionary software PAML, suggesting that the accelerated evolution in *TFIIAγ1* might be due to relaxation of purifying selection. From gene expression assays, *TFIIAγ1* is expressed at a low level under normal growing conditions and showed up to 400 times up-regulation in response to biotic and abiotic stresses, whereas *TFIIAγ5* is expressed in all organs but showed no response to biotic or abiotic stresses, suggesting that *TFIIAγ1* might have evolved to acquire an uncharacterized new function under biotic or abiotic stresses. Alternatively, *TFIIAγ1* could have lost or be losing function but still expressed. Based on the evidence, the authors argued that the Dykhuizen–Hartl effect better explains the gain of new function of *TFIIAγ1* due to the accelerated evolution and the lack of positive Darwinian selection (Sun and Ge 2010).

1.3.1.2 Positive Darwinian Selection (Adaptive Process)

The second scenario is neofunctionalization by an adaptive process. Darwinian positive selection could play an important role in promoting the fixation of beneficial mutations that lead to neofunctionalization following duplication. Under this model, once one copy has gained a new function, deleterious mutations to this gene would carry a cost and the gene has a greater chance of being retained. One example for the neofunctionalization by positive selection is two functionally redundant cytochrom P450 genes (*CYP98A8* and *CYP98A9*) formed by retroposition

in the Brassicaceae (Matsuno et al. 2009). Their ancestral gene, *CYP98A3*, is involved in the formation of lignin monomers. *CYP98A3* is expressed in many organ types but not in pollen, whereas *CYP98A8* and *CYP98A9* are highly expressed in pollen. Their new function is involved in a novel phenolic pathway for pollen development. Codon-based methods showed that several codons in these two genes have undergone repeated amino acid changes, suggesting that positive Darwinian selection has contributed to the acquisition of new function for *CYP98A8* and *CYP98A9* in pollen (Matsuno et al. 2009).

1.3.2 Subfunctionalization

Although the neofunctionalization model has been popular since it was proposed, another compelling model, referred to as subfunctionalization, was developed to serve as alternative hypothesis for the retention of duplicated genes. Subfunctionalization explains the retention of duplicated genes by partitioning of multiple ancestral functions, or partitioning of ancestral expression patterns. This model has been popularized by Hughes (1994) and Force et al. (1999). After gene duplication, each copy retains part of its ancestral function or partial expression pattern compared with the ancestral state. Regulatory subfunctionalization and protein subfunctionalization will result in two different outcomes. With regulatory subfunctionalization, duplicated genes can share the same protein function, but they exhibit the spatial or temporal partitioning in expression. With protein subfunctionalization, the duplicated genes have distinct protein functions. It is certainly imaginable that the retention of some duplicated genes might be driven by both regulatory and protein subfunctionalization. Based on population genetic simulation, subfunctionalization can be neutral in the early stage of duplicated genes by accumulating deleterious mutations in both copies without interrupting the total function of their

ancestral state (Lynch and Force 2000b). Thus, it has been argued that subfunctionalization is likely more important than neofunctionalization for the retention of duplicated genes, particularly in organisms with small effective population sizes, which experience more genetic drift (Lynch and Force 2000b). As I will present in more detail later, it does not mean that positive selection cannot act on subfunctionalized gene duplicates for their long-term preservation. Therefore, based on whether positive selection is involved or not, subfunctionalization can be further classified into two different categories: 1) duplication, degeneration, and complementation (DDC) model and 2) escape from adaptive conflict (EAC) model.

1.3.2.1 Duplication, Degenerative, and Complementation (DDC) Model (Neutral Process)

When the DDC model was proposed, Force et al. (1999) particularly focused on regulatory aspects. However, the DDC model should be also applicable to protein function. In this model, both copies will accumulate deleterious mutations in the *cis*-regulatory regions by neutral processes or genetic drift, and then each copy will only retain partial ancestral expression pattern (i.e., subfunction) over evolution and complement each other to cover the full spectrum of ancestral expression pattern. It has been shown that the probability of DDC subfunctionalization is greater than that of regulatory neofunctionalization, especially in organisms with small effective population sizes because the process largely depends on genetic drift (Force et al. 1999). A good example of the DDC model in plants was illustrated in Force et al. (1999). A pair of MADS-box transcription factor genes (*ZAG1* and *ZMM2*) in maize, which originated from a polyploidization event, showed a reciprocal expression pattern where *ZAG1* is highly expressed

throughout carpel development but weakly expressed in stamens, and *ZMM2 is* highly expressed in stamens (Mena et al., 1996). In comparison to their putative ancestral expression pattern from a single orthologous gene in Arabidopsis (*AGAMOUS*) and Antirrhinum (*PLENA*) that are highly expressed in both carpels and stamen (Yanofsky et al. 1990; Coen and Meyerowitz 1991; Bradley et al 1993), it is plausible to infer that *ZAG1* and *ZMM2* have been subfunctionalized (i.e. DDC) over evolution. A second example of the DDC model in plants is the organ-specific reciprocal gene silencing pattern for a pair of alcohol dehydrogenase genes (*AdhA*) in polyploidy cotton (Adams et al. 2003). In this example, one copy is only expressed in petals and stamens and the other copy is only expressed in carpels, indicative of regulatory subfunctionalization across different organ types.

1.3.2.2 Escape from Adaptive Conflict (EAC) Model (Adaptive Process)

Under the EAC model, a new function arises in the pre-duplication, single copy, ancestral gene, but that new function somewhat reduces the performance of the original function (i.e., improving either function comes at a cost to the other). The gain of a new function is therefore a hurdle for the improvement of the original function and vice versa. It is proposed that this dilemma situation, adaptive conflict, can be resolved by gene duplication (e.g., Hittinger and Carrol 2007). After gene duplication, one copy can be free to improve the new function and the other copy can improve the original function. Since both subfunctionalization by EAC and neofunctionalization by adaptive process show evidence of positive Darwinian selection, two major criteria are necessary to distinguish between them. First, both copies show adaptive changes in the EAC model whereas only one copy shows adaptive change in the neofunctionalization model. Second, the ancestral function will be improved in the EAC model but not in the neofunctionalization

model. Des Marais and Rausher (2008) showed a nice example of the EAC model in tandemly duplicated copies of the anthocyanin biosynthesis pathway gene dihydroflavonol-4-reductase (*DFR*) in morning glories (*Ipomea*). The major function of *DFR* is in the regulation of flower color by regulating the anthocyanin biosynthesis pathway. After the first gene duplication event, one copy (*DFR*-A/C clade) underwent repeated positive Darwinian selection, inferred based on codon-based models of sequence evolution while the copy (*DFR*-B clade) showed an improvement of its ancestral function by increasing its enzyme activity on dihydrokaempferol (*DHK*), dihydroquercetin (*DHQ*) and dihydromyricetin (*DHM*). In addition, the copy in the *DFR*-A/C clade was shown to have lost its ancestral function, suggesting that a new function has been acquired after gene duplication although this new function remains uncharacterized. This example nicely showed two criteria that fit the EAC model.

1.3.3 Subneofunctionalization

He and Zhang (2005) proposed a new model for the fates of duplicated genes, which they referred to as subneofunctionalization, based on results from protein-protein interaction patterns of duplicated genes in yeast and gene expression patterns of duplicated genes in human. In this model, some duplicated genes evolved to have mixture properties of both subfunctionalization and neofunctionalization. In other words, duplicated genes can be first subfunctionalized and then one copy gains a new function. Concurrently, Rastogi and Liberles (2005) also proposed that subfunctionalization serves as a transition state to neofunctionalization for duplicated genes based on *in silico* simulations. Later, combining *in silico* modeling and yeast microarray data, MacCarthy and Bergman (2007) found that younger duplicated genes tend to have higher proportions of subfunctionalization than older ones, which consist of more cases of

neofunctionalization. Based on these studies, subfunctionalization is an important force for the retention of duplicated genes in the early stage after gene duplication, but it is substituted by neofunctionalization later on. However, Sémon and Wolfe (2008) found that slowly evolving, duplicated genes in frogs tend to be subfunctionalized and argued that subfunctionalization can contribute to the retention of duplicated genes longer than evolutionary biologists previously thought. Apparently, subfunctionalization is important not only for the retention of young duplicated genes but also for the long term retention of slow-evolving, duplicated genes.

1.3.4 Gene Dosage Balance

The gene dosage balance hypothesis states that genes whose products interact with more gene products, referred to as "connected" genes, are more dosage-sensitive after a small-scale gene duplication than after a large-scale gene duplication (reviewed in Freeling and Thomas 2006; Birchler and Veitia 2007). How does gene dosage balance influence the retention of duplicated genes over evolution? After gene duplication, gene dosage balance is commonly used to explain the situation that either gene loss or gene gain is beneficial to maintain the proper amount of protein dosage in genomes (e.g., Birchler et al. 2001; Veitia 2002; Papp et al. 2003; Veitia 2003). Based on this model, duplicated genes that have a large pleotropic effect (or more interactions with other proteins) are expected to be retained more often after a large-scale duplication event (e.g., whole genome duplication and segmental duplication) than a small-scale duplication event (e.g., tandem duplication and retroposition) (reviewed in Freeling 2009). For example, genes whose products function as transcription factors and in signal transduction are expected to be retained after whole genome duplication such that the entire genetic network still has the proper amount of upstream regulator in regulating the doubled amount of product in the downstream

pathway. In contrast, genes involved in the terminal genetic network have a higher chance to be retained after small-scale duplication (e.g., tandem duplication) because the retention of these genes will not cause any imbalance effect for the entire genetic network. As expected, genes involved in transcriptional factors and signal transduction are overrepresented in duplicated genes from whole genome duplication while genes involved in abiotic or biotic stresses are overrepresented in those from tandem duplication in *Arabidopsis*, rice, and kiwifruit (Blanc and Wolfe 2004b; Seoighe and Gehring 2004; Maere et al. 2005; Tian et al. 2005; Shi et al. 2010). These observations support the gene dosage balance hypothesis and shed light on the importance of gene dosage balance on the retention of duplicated genes.

Since gene dosage balance plays an important role in the retention of duplicated genes, the shrinkage or expansion of gene families over evolution should be largely attributed to the degree of dosage sensitivity. Cannon et al. (2004) investigated the relative contribution of segmental and tandem gene duplication on the evolution of 50 different large gene families in *Arabidopsis*. As predicted by the gene dosage balance hypothesis, gene families that function as transcription factors and in signaling contain a high proportion of segmental gene duplicates, whereas gene families that function in pathogen defenses comprise of high proportion of tandem gene duplicates. Such an observation further supports the importance of gene dosage balance for the retention of duplicated genes, whereby more “connected” genes (i.e., upstream regulator of genetic network such as transcription factor and signaling) would be likely retained after large scale duplication and less “connected” genes (i.e., terminal nodes of genetic network such as genes in response to biotic or abiotic stresses) should be likely retained by small scale duplication.

1.3.5 Protein Subcellular Relocalization After Gene Duplication

A potential outcome of evolutionary fates for the retention of duplicated genes is protein subcellular relocalization (PSR). Similar to the concepts proposed to the functional divergence of duplicated genes via changes in expression patterns or protein sequences, PSR can result from partitioning of the ancestral subcellular localization if a gene product is localized to two or more compartments (sublocalization) or gain of a new subcellular localization (neolocalization). PSR is often associated with changes in the N-terminal signal peptide (reviewed in Byun-McKay and Geeta 2007). In plants, there are relatively few studies about the PSR after gene duplication. Nakamura et al. (2000) reported an example of PSR after gene duplication for a pair of 2-mercaptopyruvate sulfurtranderases, *AtMST1* (AT1G79230) and *AtMST2* (AT1G16460), which were derived from the most recent whole genome duplication event during the evolution of *Arabidopsis* lineage. The *AtMST1* is located in mitochondria and the *AtMST2* is located in the cytoplasm. A further examination showed that *AtMST1* has a N-terminal extension compared with its duplicated partner, *AtMST2*, and the different N-terminal region contributes to the difference in protein subcellular localization for these two duplicated genes. Two other examples for PSR after gene duplication are two pairs of nuclear-encoded organellar ribosomal protein genes in plants (Adams et al. 2002a). A duplicated *rps13* gene and a duplicated *rps15A* gene acquired new N-terminal peptides and gained a new subcellular localization to mitochondria in comparison to their ancestral localization, the chloroplast and cytoplasm, respectively. These two examples provide some insights into how a duplicated gene can change its subcellular localization.

1.4 Dissertation Goals

1.4.1 Goal 1: To Access the Frequency of Reciprocal Expression Patterns Between Duplicated Genes and to Understand the Relative Contribution of Regulatory Sub- and Neofunctionalization

Reciprocal expression means that only one duplicated gene is expressed in one (or more) organ, tissue, or developmental stage, while the other duplicated gene is expressed in another (or multiple) organ, tissue, or developmental stage. It is the most extreme kind of expression divergence pattern between duplicate genes and it likely leads to retention of both copies. Reciprocal expression can result from regulatory subfunctionalization or neofunctionalization. How common are reciprocal expression patterns of duplicated genes in plants? What is the relative importance of regulatory sub- or neofunctionalization as a fate of duplicated genes in plants? To answer these questions, I analysed expression patterns of duplicated genes using microarray data from *Arabidopsis thaliana*. To infer if the reciprocal expression patterns are a result of regulatory subfunctionalization or neofunctionalization, I performed a maximum likelihood analysis of expression patterns from gene families and reconstructed the ancestral expression state of extant duplicated genes. This approach allowed me to make inferences of subfunctionalization or neofunctionalization for some genes.

1.4.2 Goal 2: To Investigate the Divergence in Expression Patterns and Function of a Duplicated Kinase Gene Pair

How do duplicated genes gain a new function? From my large-scale study investigating the frequency of reciprocal expression patterns, I identified a pair of reciprocally expressed protein kinase genes, the *SHORT SUSPENSOR* (*SSP*) gene and the *Brassinosteroid Kinase 1* (*BSK1*) gene, that had not be previously recognized as duplicates. *SSP* is involved in paternal control of zygote elongation by its transcription in the sperm cells of pollen and then its translation in the zygote (Bayer et al. 2009), whereas *BSK1* is involved in brassinosteroid signal transduction (Tang et al. 2008). How did these two duplicated genes diverge in their expression patterns and functions? To answer this question, I studied the ancestral expression pattern using two different approaches, and I did various analyses of the gene sequences to provide insights on how the original function of *SSP* was lost and how its sequence evolved. This study allowed me to illustrate a dramatic example of neofunctionalization following gene duplication by complete changes in expression pattern and function.

1.4.3 Goal 3: To Investigate Cases of Protein Subcellular Relocalization After Gene Duplication

How does subcellular localization change (PSR) after gene duplication? Is there evidence for neolocalization and neofunctionalization in genes that show PSR? From my large-scale study of the frequency of reciprocal expression patterns in *Arabidopsis thaliana*, I found two pairs of genes that were candidates for encoding proteins showing neolocalization. I first investigated the

subcellular localization of a pair of class III peroxidase proteins using a green fluorescence protein assay. Together with a pair of calcium-dependent protein kinase proteins, I conducted sequence rate analysis to test the hypothesis that the genes whose products have a new subcellular localization show a strongly asymmetric rate of amino acid sequence evolution, suggestive of neofunctionalization.

In addition, I studied the evolution of the gene for mitochondrial ribosomal protein S13 in rosids, which arose by duplication of its chloroplast homolog followed by neolocalization. In addition an *rps13* gene exists in the mitochondrion of some rosids. I performed various analyses to further characterize the evolution of the nuclear gene in various rosids. My questions included:
How much faster is the gene evolving compared with its chloroplast homolog? What kinds of amino acid changes have taken place, where are those amino acids located in the tertiary structure, and how do they compare with the sequence of the mitochondrial copy of the gene? Has there been adaptive evolution (positive selection) in the gene sequence? After finding intact and expressed *rps13* genes in both the nucleus and mitochondrion of *Malus* (apple), I tested the hypothesis that there has been expression partitioning (reciprocal expression) of the two genes in different organ types and/or stress conditions to preserve both genes.

2 Reciprocal, Organ and Cell Type-specific Expression Patterns and Regulatory Neofunctionalization between Duplicated Genes in *Arabidopsis thaliana*[1]

2.1 Introduction

Whole genome (WG) duplication has been a recurrent phenomenon during eukaryotic evolution (reviewed in Jaillon et al. 2009). For example, at least two rounds of ancient WG duplication occurred early during vertebrate evolution, which might contribute to genome complexity and species diversification in vertebrates (Dehal and Boore 2005). All angiosperms have undergone at least one round of ancient WG duplication during their evolutionary history (e.g., Blanc and Wolfe 2004a; Sterck et al. 2005; Cui et al. 2006; Barker et al. 2008; Barker et al. 2009; Schmutz et al. 2010; Shi et al. 2010; Tang et al. 2010; Jiao et al. 2011). In addition, many plants have experienced an evolutionarily recent polyploidy event and are cytologically polyploid. Tandem duplication is another major source of duplicate genes, often caused by unequal crossing over. Both tandem and WG duplicates greatly contribute to the complexity of the transcriptomes in flowering plants.

After duplication, expression divergence of duplicated genes can be an important factor for their evolutionary retention. One particular type of expression divergence of duplicated genes leads to

1 Chapter 2 has been submitted for publication. **Liu, S.-L**., Baute, G.L., Adams, K.L. (2011) Reciprocal, organ and cell-type-specific expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*.

a reciprocal expression pattern where only one copy is expressed in some organ or tissue types, while only the other copy is expressed in others. Reciprocal expression often arises by regulatory neofunctionalization, where one copy gains a new expression pattern in some organ or tissue types and loses its ancestral expression in others, or subfunctionalization where ancestral expression patterns are divided between the duplicates in different organ types, cell types, or developmental stages (Force et al. 1999; see Figure 2.1 in this study for detailed illustrations). Several examples of reciprocally expressed duplicated genes, caused by regulatory neofunctionalization or subfunctionalization, have been reported in plants (e.g, Adams et al. 2003; Drea et al. 2006; Liu and Adams 2007; Bottley et al. 2006; Chaudhary et al. 2009; Buggs et al. 2010; Liu and Adams 2010), suggesting that reciprocal expression can be an important factor for functional diversification of duplicated genes. Most previous studies of the evolution of duplicate gene expression on a large scale in plants used correlation methods to show considerable expression divergence between duplicated gene pairs (Blanc and Wolfe 2004b; Haberer et al. 2004; Casneuf et al. 2006; Ha et al. 2007; Ganko et al. 2007; Throude et al. 2009) but those studies were not designed to detect reciprocal expression patterns. In contrast, Duarte et al. (2006) used a two-way analysis of variance approach to study gene pairs that are differentially expressed in a manner indicative of regulatory subfunctionalization and/or neofunctionalization, using 280 regulatory gene pairs and six organ types. Only a few cases of reciprocal expression were discovered in their study, probably due to the limited number of organ types and developmental stages examined.

Arabidopsis thaliana is an excellent system for studying expression evolution of duplicated genes in plants. Large amounts of microarray data are available from previous studies, including a large scale study of expression in 63 different organ and tissue types and developmental stages (Schmid et al. 2005), as well as a study of 20 different cell types and developmental stages of

roots (Birnbaum et al. 2003; Brady et al. 2007), among others. The most recent WG duplication during the evolutionary history of *Arabidopsis thaliana* likely occurred at or near the base of the Brassicaceae family, referred to as the alpha WG duplication (Bowers et al. 2003; Barker et al. 2009). About 2500 pairs of genes we estimated to have been retained from the alpha WG duplication (Blanc et al. 2003). In addition, about 4000 genes in *Arabidopsis thaliana* have been identified as tandem duplicates in clusters of various sizes (Haberer et al. 2004; Rizzon et al. 2006).

The goal of this study is to estimate the frequency of reciprocal expression patterns of WG duplicates and tandem duplicates in *Arabidopsis thaliana* across a broad range of developmental stages, organ types, and cell types, and evaluate the evolutionary importance of reciprocal expression on the retention of duplicated genes in plants. I investigated the frequency of reciprocal expression in gene duplicates derived from tandem duplication and the most recent whole genome duplication across 83 different organ types, developmental stages, and cell types by using Ath1 microarray data from *Arabidopsis thaliana* (Schmid et al. 2005; Birnbaum et al. 2003; Brady et al. 2007). I then assessed whether reciprocal expression patterns resulted from sub- or neofunctionalization by applying a maximum likelihood method of ancestral character-state reconstruction across gene families. I next performed protein sequence rate analysis to see how often reciprocally expressed gene pairs show significantly asymmetric sequence rate evolution. In addition, I studied two gene pairs in detail to further characterize cases of neofunctionalization.

2.2 Materials and Methods

2.2.1 Duplicated Gene Pair Selection

I started with a data set of 2584 pairs of duplicated genes (5168 genes) derived from the most recent whole genome (WG) duplication event identified by Blanc et al. (2003), and 1826 clusters of tandemly duplicated genes (4970 genes) identified in the current study. Identification of tandem duplicates followed the three criteria described in Zou et al. (2009): (1) they belong to the same gene family; (2) they are located within 100 kb each other; and (3) they are separated by 10 or fewer genes that do not belong to the same gene family. Only tandem duplicate clusters with two members were analyzed. I first identified and excluded WG duplicates and tandem duplicates that are not included on the Affymetrix ATH1 microarray chip, which contains 22,746 probes (> 80% of known *Arabidopsis* genes). To avoid cross-hybridization, only those genes with unique probes on the chip were selected (those that are designated with an '_at' extension and without a 's' or 'x' suffix) (e.g., Casneuf et al. 2006; Ganko et al. 2007). I also excluded gene pairs from the WG duplicate analysis where one or both members are both part of WG and tandem duplicate pairs, to minimize confounding effects on comparisons of expression patterns between the WG duplicates. First, I obtained *Arabidopsis* gene families from PLAZA 1.0 (Proost et al. 2009) and implemented a maximum likelihood analysis for every gene family by RAxML v7.0.0 (Stamatakis 2006). I then performed a bootstrap analysis (Felsenstein 1985) and retained WG duplicates that only pair with each other in the terminal branch of the tree without subsequent duplication events such as tandem duplication with at least 50% bootstrap support. For tandem duplicates, I kept those that only pair with each other in the terminal branch of the tree. Finally, I excluded genes that were annotated as probable pseudogenes by TAIR

(http://www.arabidopsis.org/). After these filtration steps, 1538 WG duplicated pairs and 466 tandem duplicated pairs were used for further analyses (available upon request).

2.2.2 Microarray Data Analysis and Detection of Reciprocal Expression

Filtered ATH1 microarray data from 63 different organ types and developmental stages (ADA, *Arabidopsis* Development Atlas; Schmid et al. 2005; Table 2.1) were obtained from the TAIR website (http://www.arabidopsis.org/). ATH1 microarray data from 20 different cell types and developmental stages in roots (ARA, *Arabidopsis* Root Atlas; Birnbaum et al. 2003; Brady et al. 2007; Table 2.2) were downloaded from the AREX website (http://www.arexdb.org/database.jsp). There are three biological replicates for each microarray data in the ADA dataset, and three to four biological replicates for each microarray data in the ARA dataset. Raw CEL files were processed and normalized using the MAS5.0 algorithm in Bioconductor (http://www.bioconductor.org/). Absence or presence of expression was statistically determined by using the "mas5calls" function in Bioconductor. The statistical test performed the Wilcoxon signed rank-based gene expression absence/presence detection algorithm and generated a detection call (i.e., a probability value) to determine if the expression signal was significantly greater than background noise. Genes with a probability value less than 0.05 were designated as presence of expression, whereas genes with a probability value equal to or greater than 0.05 were assigned as absence of expression. To better visualize the reciprocal expression patterns of gene duplicates across different developmental stages, organ types, and cell types, I also generated graphs that contain the expression profiles between duplicated genes (available upon request). Expression profile analysis and all statistical tests were implemented using the statistical package R.

After assigning the absence and presence of expression, reciprocal expression between duplicated genes was determined based on the following Boolean criteria: (1) Let y_{ij} be the expression value, where i = 1, 2 for gene copy 1 and gene copy2, and j = 1, 2, … for different developmental stages, organ types, or cell types; (2) then, $\min(y_{ij}) = 0$ and $\max(y_{ij}) > 0$; and (3) last, $\max(y_{1j} - y_{2j}) > 0$ and $\min(y_{1j} - y_{2j}) < 0$. The first definition acquires designations of absence or presence of expression described in the previous paragraph. The second and third definitions ensure that there are some conditions for both duplicated genes where expression is equal to zero or above zero (at least once).

2.2.3 Gene Ontology Analysis and Chi-square (x^2) Tests

Gene ontology (GO) annotations were obtained from the website TAIR. Any difference of enrichment of GO categories between any two datasets was compared by using chi-square (x^2) tests with 10,000 Monte Carlo simulations in the statistical package R. To correct for multiple testing, I implemented the 5% false discovery rate (FDR) adjustment algorithm described in Storey and Tibshirani (2003) using the function "p.adjust" in the software R. An FDR adjusted *P* value (or *Q* value) smaller than 0.05 was considered as a significant difference. I next compared the ratio of genes in each GO category between reciprocally expressed gene duplicates and all gene duplicates. At each developmental stage, organ type, or cell type, I also compared the ratio of genes in each GO category between neofunctionalized gene duplicates and all reciprocally expressed gene duplicates. To see if expression gain exceeds expression loss in certain organ types, developmental stages, or cell types, I compared the ratio of expression gain and expression loss at each developmental stage, organ type, and cell type by applying x^2 tests described

previously (Barker et al. 2008). This analysis allowed us to see if certain developmental stages, organ types, or cell types have any bias toward expression gain or expression loss in reciprocally expressed gene duplicates.

2.2.4 Inference of the Most Recent Common Ancestral (MRCA) Expression

Annotated protein sequences in *Arabidopsis thaliana* (TAIR, v8) were downloaded from the TAIR website (http://www.arabidopsis.org/). I obtained gene families from PLAZA 1.0 (Proost et al. 2009). For the MRCA analysis, I followed the analytical procedure described in Zou et al. (2009) and Liu and Adams (2010). Briefly, reconstruction of the MRCA expression between gene duplicates with reciprocal expression was conducted with a maximum likelihood algorithm using the program MultiState in the package BayesTraits v.1.0 (Pagel and Meade 2009). To take the uncertainty of phylogenetic tree topology into account, 100 bootstrapped trees deduced from maximum likelihood analyses by RAxML v7.0.0 were imported into BayesTraits and each tree was rooted at the midpoint using the program Reroot in the package Phylip v.3.68 (Felsenstein 2009). Prior to gene family phylogenetic analysis, protein sequences were aligned using the MUSCLE program (Edgar 2004) with default settings. Two evolutionary transition rates comprising forward (from presence of expression to absence of expression) and reverse transition (from absence of expression to presence of expression) were used for estimating the character transition rate. Two different character states were designated: absence of expression (0) and presence of expression (1). The *AddMRCA* function was used to define the state of the MRCA node for two extant duplicated genes with reciprocal expression patterns for each gene family tree (Pagel and Meade 2009). The ancestral state probability was averaged across the 100 bootstrapped trees. If the average of ancestral state probability for absence or presence of

expression was greater than 0.6, it was inferred as the ancestral expression state.

2.2.5 Detection of Asymmetric Sequence Evolution

After the inference of the MRCA expression pattern between extant duplicated gene pairs, I tested for asymmetric protein sequence evolution for these reciprocally expressed gene duplicates, in which one copy has accumulated more amino acid mutations than the other copy after their duplication. The analytical procedure followed the description in Blanc and Wolfe (2004b). To identify an outgroup orthologous sequence, the *Arabidopsis* annotated protein sequences were used as queries to search against other plant annotated protein sequences from four eudicots with available whole genome sequences (*Carica papaya*, *Glycine max*, *Populus trichocarpa*, *Vitis vinifera*) using the BLASP program. I then retrieved the best hit putatively orthologous sequences that match to one copy of *Arabidopsis* duplicated gene pair using the reciprocal best hit method described in Hulsen et al. (2006). Two criteria were used to keep the orthologous sequences for further asymmetric sequence evolution analysis. First, I kept those sequences that shared greater than 80% identity with *Arabidopsis* duplicated genes. Second, I estimated the synonymous substitution rate (d_S) for each triplet of sequences (i.e., two duplicated genes and one best hit orthologous sequence in the outgroup species) using a maximum likelihood method in PAML (Yang 1997). I kept triplets that showed d_S between the *Arabidopsis* duplicated genes that was smaller than that between the *Arabidopsis* duplicated genes and the orthologous sequence in the outgroup species.

To detect if there is significantly asymmetric sequence rate evolution between duplicated genes, protein sequences were aligned using the MUSCLE program with default settings. By using the

Codeml program in the PAML package (Yang 1997), I then obtained maximum likelihood estimates from two different hypotheses [unconstrained rate of evolution (i.e., asymmetric sequence evolution) versus clock-like rate of evolution (i.e., symmetric sequence evolution)] with the Jones substitution matrix (i.e., JTT model) and the gamma correction to accommodate variability in substitution rates. To test if the first hypothesis fits better than the second hypothesis, a likelihood ratio test was applied. Briefly, twice the difference of the likelihood estimate between these two hypotheses [$2\delta L = -2(Ln1-Ln2)$, where $2\delta L$ indicates twice likelihood ratio, *Ln1* indicates the likelihood estimate from the first hypothesis, and *Ln2* indicates the likelihood estimate from the second hypothesis] was compared against a chi-square (x^2) distribution with the degree of freedom equal to 1. The degree of freedom was obtained based on the difference of parameters used in these two different hypotheses. To correct for the issue of multiple testing, a FDR approach described previously was applied to minimize the effect of false positives.

2.2.6 Detection of Asymmetric Expression

To investigate any associations between expression divergence and protein divergence, I examined expression breadth (*EB*) for each copy of gene duplicates and calculated an asymmetric expression index (*Asy*) for gene duplicates.

I defined *EB* by the following equation: $EB_i = a_i/(a_1 + a_2 - b)$, where i = 1, 2 for gene copy 1 and gene copy 2, a_1 indicates the number of organ types, developmental stages, and cell types with expression for copy 1, a_2 indicates the number for copy 2, and b indicates the shared number for both copies.

I defined *Asy* using the following equation: $Asy = | a_1 - a_2 |/(a_1 + a_2 - b)$, where a_1 indicates the number of organ types, developmental stages, and cell types with expression for copy 1, a_2 indicates the number for copy 2, and b indicates the shared number for both copy1 and copy2.

2.2.7 Plant Materials, Nucleic Acid Extraction, and RT-PCR

To confirm organ-specific, reciprocal expression patterns in a pair of class III peroxidase genes (AT3G50990 and AT5G66390), roots and siliques from *Arabidopsis thaliana* ecotype Columbia were collected from four week old plants grown in soil with a 16 hrs/8 hrs day/night photoperiod at 23°C in a growth chamber. Nucleic acid extraction and reverse transcription-polymerase chain reaction (RT-PCR) followed the description in Liu and Adams (2008). Gene-specific primers include AT3G50990-Forward Primer (5'-GGCGGGCATTGTTCTCTCTCAAAT-3'), AT3G50990-Reverse Primer (5'-TCAATGACTTCGAACCCTCGAGCA-3'), AT5G66390-Forward Primer (5'-GAAACCACGGGCTGAGTTTCTGTT-3'), and AT5G66390-Reverse Primer (5'-GCAGCTCTCTCGCTCATTGCATTT-3').

2.2.8 Subcellular Localization Analysis

To examine the subcellular localization in a pair of class III peroxidase genes (AT3G50990 and AT5G55390), full length cDNA products were amplified by RT-PCR using gene-specific primers that include the following underlined restriction enzyme site: AT3G50990F-*KpnI* (5'-CGGGGTACCATGAATACAAAAACGGTGAAG-3'), AT3G50990R-*BamHI* (5'-CGCGGATCCAACATCATGGTTAACCCTCC-3'), AT5G66390F-*KpnI*

(5'-CGGGGTACCATGGCCAAGTCATTGAACATC-3'), and AT5G66390R-*BamHI* (5'-CGCGGATCCATAAGCATGGTTAACCCTCC-3'), with the RT-PCR conditions described above. All PCR products were cloned in frame into pCambia1300 modified vector with CaMV 35S promoter at 5' upstream and GFP 3' downstream. The inserted nucleotide sequence in the resultant plasmid was checked by DNA sequencing. Then, the pCambia1300-AT3G50990-GFP and pCambia1300-AT5G66390-GFP were transformed into *Arabidopsis thaliana* ecotype Columbia. *Agrobacterium*-mediated transformation was conducted using the floral dip method described in Clough and Bent (1998). GFP fluorescence was visualized by using a confocal laser scanning microscope.

2.3 Results

2.3.1 Reciprocal Expression Patterns are Common among Duplicated Genes

To identify cases of reciprocal expression, I analyzed expression patterns of 1538 pairs of genes identified as having been duplicated from the alpha WG duplication event and 466 pairs of tandem duplicates in *A. thaliana* using Affymetrix Ath1 microarray data from 63 different organ types and developmental stages (ADA, *Arabidopsis* Development Atlas; Schmidt et al. 2005), and 20 different cell types and developmental stages in roots (ARA, *Arabidopsis* Root Atlas; Birnbaum et al., 2003; Brady et al. 2007). 63 different organ types and developmental stages, and 20 different cell types and developmental stages are listed in Table 2.1 and Table 2.2, respectively. Each of these two datasets was generated in a single project and had a minimum of three biological replicates. The absence or presence of expression was determined using the Wilcoxon signed rank-based gene expression presence/absence detection algorithm in

Bioconductor (see Materials and Methods). I found that 24% of the WG duplicates in the ADA dataset and 13% in the ARA dataset showed reciprocal expression patterns (Figure 2.2). Among the tandem duplicates, 32% in the ADA dataset and 15% in the ARA dataset show reciprocal expression patterns (Figure 2.2). The lower percentages in the ARA dataset was possibly due to the lower number of data points (20 vs. 63 in the ADA dataset). Seven percent of the WG duplicate gene pairs and nine percent of the tandem duplicates showed reciprocal expression in both the ADA and ARA datasets. The tandem duplicates have a significantly higher frequency of reciprocal expression than the WG duplicates in the ADA dataset (x^2, $P = 0.0003$; Figure 2.2B), but not in the ARA dataset (x^2, $P > 0.05$; Figure 2.2B). When both datasets are considered together, there is a significantly higher frequency of reciprocal expression in the tandem duplicates (38%) than WG duplicates (30%) from the combination of the ADA dataset and the ARA dataset (x^2, $P = 0.0006$; Figure 2.2B), indicative of a higher frequency of expression diversification in tandem duplicates than WG duplicates.

To investigate if certain types of genes more often show a reciprocal expression pattern, I conducted a gene ontology (GO) analysis using the GO annotations from the TAIR website. In this analysis I combined the reciprocally expressed gene duplicates from both the ADA dataset and the ARA dataset and compared duplicated genes with reciprocal expression against all duplicated genes using a chi-square (x^2) analysis. Among the WG duplicates, transcription (x^2, $Q = 0.0007$), transcription factor activity (x^2, $Q = 0.0011$), DNA or RNA binding (x^2, $Q = 0.0490$), and other cellular components (x^2, $Q = 0.0007$) were overrepresented (Figure 2.3A), while other intracellular components (x^2, $Q = 0.0007$), protein metabolism (x^2, $Q = 0.0092$), other cytoplasmic components (x^2, $Q = 0.0007$), organelle-associated genes ((x^2, $Q = 0.0066$-0.0490; e.g., plastids and mitochondria), Golgi apparatus (x^2, $Q = 0.0437$), cytosol (x^2, $Q = 0.0007$), and ribosome (x^2, $Q = 0.0007$) were underrepresented (Figure 2.3A). Among the tandem duplicates,

I did not observe any functional bias for duplicated genes showing reciprocal expression (x^2, $Q >$ 0.05; Figure 2.3B).

2.3.2 Reciprocal Expression Results more from Neofunctionalization than Subfunctionalization

I next applied an integration of expression data and gene family phylogenetic information to infer the putative most recent common ancestral (MRCA) expression pattern of reciprocally expressed gene duplicates. The MRCA expression patterns inferred from other gene members in a gene family using a maximum likelihood algorithm (Gu 2004; Gu et al. 2005; Oakley et al. 2006; Zou et al. 2009; Liu and Adams 2010) can be used to approximate the ancestral state of expression pattern and infer whether the detected cases of reciprocal expression are caused by subfunctionalization (partitioning of ancestral expression patterns) or neofunctionalization (gain of a new expression pattern). The method has been applied to the inference of regulatory sub- or neofunctionalization between duplicated genes in fruit flies (Oakley et al. 2006) and *Arabidopsis* (Zou et al. 2009). To better infer the MRCA expression pattern, phylogenetic distances and the uncertainty of the phylogenetic gene tree topology were also taken into account (Pagel 1999). Among the WG duplicates with reciprocal expression patterns, 46% in the ADA dataset and 36% in the ARA dataset were inferred to be neofunctionalized while only 5% in the ADA dataset and 7% in the ARA dataset were inferred to be subfunctionalized (Figure 2.4A). Among the tandem duplicates, 36% in the ADA dataset and 11% in the ARA dataset were inferred to be neofunctionalized while only 3% in the ADA dataset and 7% in the ARA dataset were inferred to be subfunctionalized (Figure 2.4B). Among those neofunctionalized cases, a small percentage of them (8-14% in WG duplicates and 5% in tandem duplicates) showed gain of a new expression

pattern for both copies (Figure 2.4). The ancestral expression state in some cases could not be assessed due to uncertainty in the phylogenetic topology (labeled as UKW in Figure 2.4) or lack of information such as a small gene family size with only two members (labeled as N.D. in Figure 2.4). The results of the ancestral expression state reconstructions suggested that reciprocal expression patterns between duplicated genes in *Arabidopsis thaliana* are caused more by gain of a new expression pattern (neofunctionalization) than by partitioning of the ancestral expression pattern (subfunctionalization).

2.3.3 Preferential Gain or Loss of Gene Expression in Whole Genome Duplicates and Tandem Duplicates

I next assessed if expression is preferentially gain or loss in particular organ types, developmental stages, and cell types among both WG duplicates and tandem duplicates in both the ADA and ARA datasets (see Materials and Methods for details). In the ADA dataset, a significantly higher percentage of genes with expression gain than expression loss was found in pollen (x^2, $Q = 0.0009$; ca. 11% higher), the shoot apex after bolting (x^2, $Q = 0.0204$; ca. 9% higher), senescing leaf (x^2, $Q = 0.0288$; ca. 9% higher), seeds at developmental stage 9 (x^2, $Q = 0.0288$; ca. 10% higher), and seeds at developmental stage 10 (x^2, $Q = 0.0288$; ca. 10% higher) among the WG duplicates (Figure 2.5A). Among these five organ types/developmental stages, a significantly higher percentage of expression gain than expression loss was only observed in pollen, when a more stringent Bonferroni correction was applied ($P < 0.05$), suggesting that pollen shows a more striking pattern in terms of expression gain after WGD. In contrast, there are not any particular organ types (or developmental stages) showing a significant difference between expression gain and expression loss among the tandem duplicates (x^2, $Q > 0.05$; Figure

2.5A). I then performed gene ontology analysis to see if there is any functional bias for neofunctionalized gene duplicates in shoot apex after bolting, senescing leaf, pollen, and seeds. In shoot apex after bolting, other cytoplasmic components (x^2, $Q = 0.0176$) were overrepresented (data not shown). In pollen, nucleotide binding (x^2, $Q = 0.0334$), transport (x^2, $Q = 0.0125$), mitochondria (x^2, $Q = 0.0289$), and plastid (x^2, $Q = 0.0334$) were overrepresented while transcription (x^2, $Q = 0.0012$) and transcription factor activity (x^2, $Q = 0.0028$) were underrepresented (data not shown). In the other organ structures I did not find any significant difference for each GO category between the neofunctionalized duplicated genes and the total reciprocally expressed duplicated genes (data not shown).

In the ARA dataset, I found that no particular cell types showed a significant difference between expression gain and expression loss among the WG duplicates (x^2, $Q > 0.05$; Figure 2.5B), while two different cell types, phloem and all radial root tissues at stage 3, were found to show a significantly higher percentage of expression loss than that of expression gain among the tandem duplicates (x^2, $Q = 0.0136$-0.0139; Figure 2.5B).

After assessing if particular organ types, developmental stages, and cell types showed preferential expression gain or loss, I next examined if there is any difference in expression gain or loss between WG duplicates and tandem duplicates. In WG duplicates, a significantly higher percentage of expression gain (ca. 10%) than expression loss (ca. 6%) was observed in the ADA dataset (x^2, $Q < 0.0001$; Figure 2.5A), but not in the ARA dataset (x^2, $Q > 0.05$; Figure 2.5B). In contrast, an opposite trend was found in tandem duplicates, where expression loss is significantly more common than expression gain in both the ADA dataset (loss: ca. 11% v.s. gain: ca. 7%) and the ARA dataset (loss: ca. 14% v.s. gain: ca. 3%) (x^2, $Q < 0.0001$; Figure 2.5). This result suggests that gain of expression is more prevalent than loss of expression among WG

duplicates and that an opposite trend exists in tandem duplicates.

2.3.4 Asymmetric Sequence Evolution in Some Pairs with Neofunctionalization of Expression Patterns

To further investigate several cases with neofunctionalization, I performed an analysis of the rate of change of protein sequence for the reciprocally expressed gene duplicates. Asymmetric rate evolution has been proposed as a likely indicator of neofunctionalization, because one copy has accelerated amino acid replacements in comparison to its duplicated partner over evolution (Blanc and Wolfe 2004b). A higher rate of accumulation of replacement mutations is often associated with the gain of new function (Byrne and Wolfe 2007). In my asymmetric rate analysis, the best hit orthologous sequence from an outgroup species was used to polarize the evolutionary rate between gene duplicates (see Materials and Methods for details). Of the WG duplicates, 40 of 265 triplets (15%) showed significantly asymmetric protein sequence divergence ($Q < 0.05$; table 1). Of the tandem duplicates, 8 of 55 triplets (15%) showed significantly asymmetric protein sequence divergence ($Q < 0.05$; Table 2.3). Among them, there are 15 cases (classified as group 1) that showed both asymmetric sequence rate evolution and expression gain as determined by the MRCA expression analysis (Table 2.3), further supporting my inference of neofunctionalization. There were five cases (classified as group 2) where one copy showed significantly accelerated rate evolution and both copies were inferred as neofunctionalized by the MRCA analysis (Table 2.3). In two cases (classified as group 3), the inference from MRCA analysis was subfunctionalization but there was asymmetric sequence rate evolution between the duplicates, suggesting neofunctionalization (Table 2.3). Those two pairs might have undergone a transition stage between subfunctionalization and neofunctionalization,

referred to as subneofunctionalization (He and Zhang 2005), via a combination of regulatory subfunctionalization and sequence neofunctionalization. The remaining cases were inferred to be neofunctionalized only by asymmetric sequence rate analysis because of the uncertain inference by MRCA or the lack of inference without the phylogenetic information (Table 2.3). To test if older duplicated genes tend to show asymmetric rate evolution, I conducted a comparison of synonymous substitution rate (d_S) between pairs with symmetric evolution and asymmetric evolution. I did not see any significant difference in terms of the age of gene duplicates between the symmetric group and the asymmetric group (Figure 2.6). Tandem duplicates were on average younger than WG duplicates based on their d_S data (Figure 2.6), which is consistent with previous reports (Haberer et al. 2004; Blanc and Wolfe 2004b). Although tandem duplicates did not show a significantly higher frequency of asymmetric rate evolution than WG duplicates (x^2, $P > 0.05$), their protein sequence seemed to diverge faster than WG duplicates when considering that their age is relatively younger than WG duplicates (data available upon request).

2.3.5 Asymmetric Sequence Evolution is Associated with Asymmetric Expression Divergence

After investigating the frequency of asymmetric rate evolution for those duplicates with reciprocal expression, I conducted an analysis to see if there is any association between asymmetric sequence divergence and expression divergence. I first scored the expression breadth (*EB*; i.e., how many conditions in which one gene is expressed) from both the ADA and ARA datasets. I then compared the asymmetric expression index (*Asy*; i.e., expression breadth difference) between gene duplicates with symmetric evolution and asymmetric evolution, as well as the expression breadth between the accelerated copy and the non-accelerated copy if gene

duplicates showed asymmetric rate evolution. In WG duplicates, I observed that gene duplicates with asymmetric rate evolution have significantly higher *Asy* value between duplicated genes in comparison to those with symmetric rate evolution (*t*-test, $P = 0.0038$; Figure 2.7A), suggesting that asymmetric rate evolution is often associated with asymmetric expression divergence. When comparing with expression breadth (*EB*), I also found that the copy with accelerated amino acid replacements often showed a lower *EB* value in comparison to its non-accelerated duplicated partner (*t*-test, $P < 0.0001$; Figure 2.7B), suggesting that the copy with accelerated amino acid evolution tends to lose expression across multiple organ types (or cell types or developmental stages) and gain expression pattern in a limited number of organ types conditions (i.e., organ types, cell types, or developmental stages). A similar trend was also found in the tandem duplicates (Figures 2.7C-2.7D); however, the trend was not statistically supported perhaps due to the small sample size (n = 8; *t*-test, $P > 0.05$).

2.3.6 Potential Cases of Neofunctionalization involving Pollen

Among the reciprocally expressed WG duplicates, a significantly higher percentage of expression gain was found in pollen than in other organ types. There were 43 gene pairs that showed expression gain in pollen (Table 2.4). Among them, seven pairs of duplicated genes showed striking reciprocal expression patterns that involved pollen (Figures 2.8, 2.10A). These seven pairs of duplicated genes all showed that one copy has gained expression in pollen; in contrast, its duplicated partner had broad expression across different organ types but no expression in pollen. Three of them only showed expression gain from MRCA expression analysis, including: a pair of hexokinase-like genes (AT3G20040 and AT1G50460), a pair of calcium-dependent lipid-binding genes (AT5G37740 and AT1G66360), and a pair of

O-flucosyltransferase genes (AT1G11990 and AT1G62330) (Figures 2.8A-2.8C). Four of the gene pairs showed both gain of expression from MRCA expression analysis and asymmetric sequence evolution, including: a pair of *GDSL*-motif lipase/hydrolase genes (AT5G03610 and AT3G09930), a pair of dynamin-related genes (AT3G60190 and AT2G44590), a pair of calcium-dependent protein kinase genes (AT3G10660 and AT5G04870), and a pair of trichome birefringence-like genes (AT5G06700 and AT3G12060) (Figures 2.8D-2.8F, 2.10A). The calcium-dependent protein kinase genes are differentially targeted to the peroxisomes and the endoplasmic reticulum (see details in below). The functions for most of the above gene pairs remain uncharacterized. Dynamin-like proteins and calcium-dependent protein kinases have been shown to be involved in pollen tube development (Moutinho et al. 1998; Konopka et al. 2008; Backues et al. 2010), although it is not known if the gene pairs studied here have those functions. The previously reported *SSP* and *BSK1* gene pair (Liu and Adams 2010), which showed pollen-specific reciprocal expression and asymmetric sequence evolution, were not identified in this study because *SSP* has undergone a subsequent duplication and such genes were excluded from this study (see Materials and Methods for details). Thus there may be additional duplicated genes in the *Arabidopsis thaliana* genome that show reciprocal expression involving pollen and asymmetric sequence rate evolution.

2.3.7 Neofunctionalization and Differential Subcellular Localization in a Pair of Class III Peroxidase Proteins

When I examined some cases of organ-specific reciprocal expression, I found a possible example of neofunctionalization: a pair of class III peroxidase genes, the WG duplicates *AtPrx36* (AT3G50990) and *AtPrx72* (AT5G66390). From both microarray expression data and RT-PCR

assays, *AtPrx36* and *AtPrx72* showed reciprocal expression between siliques, where only *AtPrx36* is expressed, compared with roots where only *AtPrx72* is expressed (Figures 2.9A-2.9B). Reciprocal expression between roots and siliques was inferred to be neofunctionalization from MRCA analysis, where *AtPrx36* gained a new expression pattern in siliques (Table 2.3). A second line of evidence for neofunctionalization of *AtPrx36* was found from the asymmetric sequence rate analysis, where *AtPrx36* showed accelerated sequence evolution, suggestive of neofunctionalization (Figure 2.9C; Table 2.3). I also noticed that *AtPrx36* has a more divergent N-terminal region than *AtPrx72* when comparing with other homologous/orthologous sequences in outgroup species (Figure 2.9D), suggesting that *AtPrx36* changed its subcellular localization by acquiring a new targeting sequence.

I then performed a green fluorescent fusion protein (GFP) subcellular localization assay to see if the products of *AtPrx36* and *AtPrx72* are targeted to different subcellular locations (see Materials and Methods for details). I found that the product of *AtPrx36* is targeted to the cell wall (Figure 2.9E) whereas the product of *AtPrx72* is located in the cytosol (Figure 2.9F). The targeting results further support the N-terminal sequence analysis that suggests that *AtPrx36* has changed its subcellular localization after duplication. These results show that the subcellular localization of *AtPrx36* changed from the cytosol to the cell wall. Overall, *AtPrx36* shows three lines of evidence for neofunctionalization: a change in subcellular localization, asymmetric sequence rate evolution, and evidence for acquisition of a new expression pattern from the ancestral expression pattern reconstruction analysis.

2.3.8 Neofunctionalization and Differential Subcellular Localization in a Pair of Calcium-dependent Protein Kinase Proteins

Another example of differential subcellular localization among the genes showing reciprocal expression in this study is a pair of calcium-dependent protein kinase genes *AtCPK2* (AT3G10660) and *AtCPK1* (AT5G04870) (Figure 2.10A). Previous studies using GFP localization experiments have shown that they possess different subcellular localization ability: *AtCPK2* is localized in the endoplasmic reticulum (ER) (Lu and Hrabak 2002) and *AtCPK1* is localized in peroxisomes and lipid bodies (Dammann et al., 2003; Coca and San Segundo 2010). In those studies, the authors showed that the N-terminal region is the critical region determining their subcellular localization ability. Lu and Hrabak (2002) experimentally showed that the first 10 amino acids of *AtCPK2* are readily to direct *AtCPK2* to the ER. Dammann et al. (2003) showed that replacing the first seven amino acids of *AtCPK1* resulted in loss of peroxisome targeting; thus the peroxisome targeting function is located in the first eight amino acids that are highly conserved in other eurosids (Figure 2.10B). This implies that *AtCPK1* retains the ancestral subcellular localization and that *AtCPK2* acquired its subcellular localization to the ER after duplication. It appears that mutations in amino acids four and nine of *AtCPK2* have allowed for targeting to the ER, and the substitution of thymine to alanine at position four abolished targeting to the peroxisome (Figure 2.10B).

My MRCA analysis suggests that *AtCPK2* acquired expression in pollen and loss of expression in many other organ types, while *AtCPK1* retains the ancestral expression pattern of most organ types but not pollen (Figure 2.10A; Table 2.3). From the asymmetric sequence analysis, *AtCPK2* was observed to evolve much faster than its duplicated partner, *AtCPK1* (Figure 2.10C; Table

2.3), suggesting that *AtCPK2* has experienced accelerated evolution. Results from both MRCA and sequence rate analysis provide evidence to support the inference of neofunctionalization for *AtCPK2* including: (1) gaining a pollen-specific expression, (2) evolving in an accelerated fashion in its protein sequence, and (3) acquiring a new subcellular localization.

2.4 Discussion

2.4.1 Reciprocal Expression and Regulatory Neofunctionalization are Common among Duplicated Genes

My study provides some new insights into the evolutionary importance of reciprocal expression patterns between duplicated genes in plants. First, reciprocal expression in different organ types, tissues, cell types, and developmental stages is common among duplicated genes in *Arabidopsis thaliana*. I have shown that 30 to 38% of the duplicated genes in *Arabidopsis* that were examined in this study are reciprocally expressed in different organ types, cell types, and developmental stages**.** This result contrasts to the results of Duarte et al. (2006) who found few cases of reciprocal expression of duplicated genes in *Arabidopsis thaliana* among the six organ types that they examined. Considering that my study examined data from 83 different organ types, cell types, and developmental stages, it is not surprising that I found a much higher number of gene pairs with reciprocal expression patterns. Second, I found that transcription factors are overrepresented among the reciprocally expressed WG duplicates compared to the entire set of WG duplicates, which in itself is overrepresented with transcription factors (Blanc and Wolfe, 2004b). Over-representation of transcription factors among WG duplicates has been explained by the gene dosage balance hypothesis (reviewed in Freeling 2009; Edger and Pires

2009). In the gene dosage balance hypothesis, loss of one copy for highly connected genes such as transcription factors and signaling transduction would lead to a detrimental effect on fitness after whole genome duplication. I propose that after being initially retained by gene dosage, or other reasons, many WG duplicates that are transcription factors underwent regulatory neofunctionalization (or subfunctionalization) that would be associated with functional divergence.

Third, my results indicate that the reciprocal expression patterns of most gene pairs (of those that could be assessed) appear to result from regulatory neofunctionalization instead of regulatory subfunctionalization. This result is consistent with recent proposals that have de-emphasized the importance of subfunctionalization as a factor for the retention of duplicated genes and instead proposed that subfunctionalization is primarily a gene divergence consequence (Freeling 2008). Frequencies of regulatory subfunctionalization or neofunctionalization have been inferred in several different eukaroytes, such as yeast (Tirosh and Barkai 2007), fruit fly (Oakley et al. 2006), and mammals (Farré and Albà 2010). In yeast, 45% of duplicated genes have been shown to experience regulatory neofunctionalization (Tirosh and Barkai 2007). In *Drosophila*, Okaley et al. (2006) inferred that regulatory neofunctionalization (ca. 28%) is more common than regulatory subfunctionalization (ca. 10%). In mammals, Farré and Albà (2010) studied the expression evolution of gene duplicates, and found that 23-25% of them showed regulatory subfunctionalization and 42-52% of them were neofunctionalized, suggesting that regulatory neofunctionalization is more prevalent than regulatory subfunctionalization. My study is consistent with these previous studies conducted in a broad range of eukaryotic organisms, implying that regulatory neofunctionalization plays a more important role than regulatory subfunctionalization in the retention and divergence of duplicated genes over evolution.

Inferring ancestral expression states using maximum likelihood analyses of gene expression within a gene family in a single species can be done computationally for a large number of genes, given the readily available expression data. In addition, the expression data come from one species, allowing for unambiguous comparisons between organ types at the exact same developmental stage. However, the ancestral state reconstruction approach may overestimate the number of genes that have undergone neofunctionalization because subsequent changes in expression of other genes in the family after their common ancestor with the duplicate pair in question could lead to an incorrect inference of neofunctionalization. Additional evidence for neofunctionalization of one copy after gene duplication comes from a combination of evidence for asymmetric sequence rate evolution, along with the ancestral state inference of neofunctionalization, plus information from functional studies if available. Fifteen genes in this study had both asymmetric sequence rate evolution and an ancestral state expression inference of neofunctionalization. In another recent study that used the ancestral state reconstruction approach to study expression patterns of duplicated genes in *Arabidopsis thaliana,* Zou et al. (2009) found that the expression patterns in response to nine abiotic stress treatments indicated that a much higher percentage of genes lost stress responsiveness (up or down regulation under stress) than gained stress responsiveness. Their results are consistent with a larger role for regulatory subfunctionalization than neofunctionalization in the evolution of stress responsiveness of duplicated genes. The results of my study contrast to their observations. However the two data sets are different in type (abiotic stresses vs. organs, developmental stages, and cell types). Another difference is that I did not analyze up and down regulation of expression level in this study, instead focusing on reciprocal expression patterns.

My study showed that a considerable number of duplicate pairs from both WG duplicates and tandem duplicates are reciprocally expressed. What are possible molecular mechanisms causing

reciprocal expression between duplicates genes? One possible mechanism is divergence of *cis*-regulatory element regions between duplicated genes. In *Arabidopsis*, Haberer et al. (2004) found that both segmental duplicates and tandem duplicates showed highly similar *cis*-element regions even though they have high expression divergence, suggesting that minor changes in *cis*-element regions could lead to regulatory neofunctionalization or subfunctionalization in gene duplicates. Another possible mechanism is unequal crossing over. Because tandemly duplicated genes are often derived from unequal crossing over, it is possible that only part of a *cis*-element region is duplicated (Achaz et al. 2000), potentially leading to an instantaneous change in expression pattern after gene duplication such as reciprocal expression patterns.

2.4.2 Expression Gain and Accelerated Sequence Evolution in Pollen

Among the reciprocally expressed WG duplicates, a significantly higher percentage of expression gain was found in pollen (i.e., male gametophyte) than in other organ types from sporophyte and mixture of sporophyte and female gametophyte, including 43 gene pairs that showed expression gain in pollen. The pollen transcriptome has been shown to be distinctive from the transcriptome in other sporophytic organ types, with many genes expressed specifically in pollen (Becker et al. 2003; Honys and Twell 2003). Expression changes after gene duplication may contribute to the distinctiveness of the pollen transcriptome.

I found that four pairs of duplicated genes that showed striking reciprocal expression patterns in pollen had undergone accelerated sequence evolution. Genes that are expressed in reproductive organs sometimes evolve rapidly or undergo positive evolution (reviewed by Swanson and Vacquier 2002). The rapid evolution of traits that are related to reproductive organs has been

considered as an important evolutionary mechanism of speciation (Gavrilets 2000). In plants, a similar trend has been observed in several genes with pollen-specific expression (Fiebig et al. 2004; Schein et al. 2004). The accelerated evolution or positive selection of pollen-specifically expressed genes can be driven by pollen competition and sexual conflict (review by Bernasconi et al. 2004). In addition, the accelerated sequence evolution and positive selection can be casued by species recognition (Ishimizu et al. 1998), or neofunctionalization such as novel phenotypic effects during pollen development (Matsuno et al. 2009) and novel signaling pathway for paternal control of embryogenesis (Liu and Adams 2010).

2.4.3 Differences between Two Different Types of Gene Duplicates

The results from the comparison of reciprocal expression frequency between WG duplicates and tandem duplicates showed that reciprocal expression has occurred more frequently in tandem duplicates. In addition, the results of my ancestral expression pattern analysis indicate that WG duplicates showed more expression gain than expression loss, while tandem duplicates showed more loss than gain. Thus, expression evolution is different between these two different types of duplicates. Casneuf et al. (2006) and Ganko et al. (2007) found that gene duplicates from large scale duplication events (e.g., WG duplicates) largely have highly redundant or overlapping expression pattern, and showed less expression divergence than those from small scale duplication event (e.g., tandem duplicates). One possible explanation is due to the difference of gene duplication mechanisms. Tandem duplication is often derived from unequal crossing over (Achaz et al. 2000). Duplication by unequal crossing over can disrupt the promoter region, whereas that would not occur by whole genome duplication. In contrast, duplicated genes derived from a whole genome duplication event often share almost the entire regulatory region.

Thus, it is expected that expression divergence in tandem duplicates should be greater and that there should be more loss of expression than in WG duplicates.

2.4.4 Accelerated and Asymmetric Sequence Evolution

Duplicated genes can diverge in protein sequences. There are several studies showing a positive correlation between expression divergence and degree of protein divergence in duplicated genes over evolution (e.g., Casneuf et al. 2006; Ganko et al. 2007). My study also showed a significant association between asymmetric expression divergence and asymmetric sequence evolution. In such cases, one copy retained the inference of ancestral expression pattern and had a slower rate of sequence divergence, while the other copy lost expression in some organs and gained expression in others, and evolved in an accelerated manner. Examples include the peroxidase gene pair and the calcium dependent protein kinase gene pair (see Results). The combination of gain of new expression and loss of ancestral expression pattern ultimately leads to a reciprocal expression pattern between duplicated genes. These results are consistent with findings in yeast duplicated genes (Tirosh and Barkai 2007), where asymmetric expression divergence is associated with asymmetric protein divergence. An association between asymmetric expression divergence and asymmetric sequence evolution has not yet reported in duplicated genes in plants.

2.4.5 Asymmetric Sequence Rate Evolution and Neofunctionalization of the *AtRecQ4B* DNA Helicase Gene

A pair of reciprocally expressed DNA helicase genes (AT1G60930 and AT1G10630), in which the function of both copies has been characterized (Hartung et al. 2007), shows asymmetric sequence rate evolution, with the *AtRecQ4B* (AT1G60930) evolving more rapidly (table 1). The products of the two duplicated genes have antagonistic functions where *AtRecQ4B* promotes homologous recombination (HJ) by stabilizing recombination intermediates, whereas *AtRecQ4A* suppresses the frequency of recombination. In comparison to the functions of *recQ*-like genes from other eukaryotes, homologous *recQ* genes in human and yeast mainly perform the function of suppressing recombination that is similar to *AtRecQ4A*, suggesting that neofunctionalization has occurred in *AtRecQ4B* after the gene duplication event within Brassicaceae. Hartung et al. (2007) favored the subfunctionalization model between *AtRecQ4A* and *AtRecQ4B* based on the fact that both the promotion and suppression of recombination were observed in homologous *recQ*-like genes in *E. coli*. However, those antagonistic functions have not been found in any other eukaryote besides *Arabidopsis thaliana*, suggesting that recent evolution of the recombination promotion function of *ARecQ4B* occurred after gene duplication. Further supporting the inference of neofunctionalization is my finding of accelerated and asymmetric sequence evolution in *AtRecQ4B*.

Table 2.1. List of 63 different organ types and developmental stages for ATH1 microarray data in the *Arabidopsis* Development Atlas (Schmid et al. 2005).

No.	Sample ID	Genotype	Tissue	Plant Age	Photoperiod	Substrate
O1	ATGE_3	WT; columbia	root	7 days	continuous light	soil
O2	ATGE_94	WT; columbia	root	8 days	continuous light	1 x MS agar
O3	ATGE_95	WT; columbia	root	8 days	continuous light	1 x MS agar; 1% sucrose
O4	ATGE_93	WT; columbia	root	15 days	continuous light	1 x MS agar; 1% sucrose
O5	ATGE_9	WT; columbia	root	17 days	continuous light	soil
O6	ATGE_98	WT; columbia	root	21 days	continuous light	1 x MS agar
O7	ATGE_99	WT; columbia	root	21 days	continuous light	1 x MS agar; 1% sucrose
O8	ATGE_2	WT; columbia	hypocotyl	7 days	continuous light	soil
O9	ATGE_6	WT; columbia	shoot apex; vegetative	7 days	continuous light	soil
O10	ATGE_8	WT; columbia	shoot apex; before bolting	14 days	continuous light	soil
O11	ATGE_29	WT; columbia	shoot apex; after bolting	21 days	continuous light	soil
O12	ATGE_27	WT; columbia	stem; 2nd internode	21+ days	continuous light	soil
O13	ATGE_28	WT; columbia	1st node	21+ days	continuous light	soil
O14	ATGE_4	WT; columbia	shoot apex; vegetative and young leaf	7 days	continuous light	soil
O15	ATGE_7	WT; columbia	seedling; green part	7 days	continuous light	soil
O16	ATGE_96	WT; columbia	seedling; green part	8 days	continuous light	1 x MS agar
O17	ATGE_97	WT; columbia	seedling; green part	8 days	continuous light	1 x MS agar; 1% sucrose
O18	ATGE_100	WT; columbia	seedling; green part	21 days	continuous light	1 x MS agar
O19	ATGE_101	WT; columbia	seedling; green part	21 days	continuous light	1 x MS agar; 1% sucrose
O20	ATGE_1	WT; columbia	cotyledon	7 days	continuous light	soil
O21	ATGE_5	WT; columbia	leaf 1 + 2	7 days	continuous light	soil
O22	ATGE_91	WT; columbia	leaf	15 days	long day (16/8)	soil
O23	ATGE_19	WT; columbia	leaf 7; petiole	17 days	continuous light	soil
O24	ATGE_20	WT; columbia	leaf 7; proximal half	17 days	continuous light	soil
O25	ATGE_21	WT; columbia	leaf 7; distal half	17 days	continuous light	soil
O26	ATGE_25	WT; columbia	senescing leaf	35 days	continuous light	soil
O27	ATGE_26	WT; columbia	cauline leaf	21+ days	continuous light	soil
O28	ATGE_87	WT; columbia	rossette	7 days	short day (10/14)	soil
O29	ATGE_89	WT; columbia	rossette	14 days	short day (10/14)	soil
O30	ATGE_90	WT; columbia	rossette	21 days	short day (10/14)	soil
O31	ATGE_10	WT; columbia	rossette#4; 1 cm long	10 days	continuous light	soil
O32	ATGE_12	WT; columbia	rossette#2	17 days	continuous light	soil
O33	ATGE_13	WT; columbia	rossette#4	17 days	continuous light	soil
O34	ATGE_14	WT; columbia	rossette#6	17 days	continuous light	soil
O35	ATGE_15	WT; columbia	rossette#8	17 days	continuous light	soil
O36	ATGE_16	WT; columbia	rossette#10	17 days	continuous light	soil
O37	ATGE_17	WT; columbia	rossette#12	17 days	continuous light	soil
O38	ATGE_22	WT; columbia	rossette	21 days	continuous light	soil
O39	ATGE_23	WT; columbia	rossette	22 days	continuous light	soil
O40	ATGE_24	WT; columbia	rossette	23 days	continuous light	soil
O41	ATGE_31	WT; columbia	flower stage 9	21+ days	continuous light	soil
O42	ATGE_32	WT; columbia	flower stage10/11	21+ days	continuous light	soil
O43	ATGE_33	WT; columbia	flower stage 12	21+ days	continuous light	soil
O44	ATGE_39	WT; columbia	flower stage 15	21+ days	continuous light	soil
O45	ATGE_92	WT; columbia	flower	28 days	continuous light	soil
O46	ATGE_40	WT; columbia	flower stage 15; pedicel	21+ days	continuous light	soil
O47	ATGE_34	WT; columbia	flower stage 12; sepal	21+ days	continuous light	soil

No.	Sample ID	Genotype	Tissue	Plant Age	Photoperiod	Substrate
O48	ATGE_41	WT; columbia	flower stage 15; sepal	21+ days	continuous light	soil
O49	ATGE_35	WT; columbia	flower stage 12; petal	21+ days	continuous light	soil
O50	ATGE_42	WT; columbia	flower stage 15; petal	21+ days	continuous light	soil
O51	ATGE_36	WT; columbia	flower stage 12; stamen	21+ days	continuous light	soil
O52	ATGE_43	WT; columbia	flower stage 15l stamen	21+ days	continuous light	soil
O53	ATGE_37	WT; columbia	flower stage 12; carpel	21+ days	continuous light	soil
O54	ATGE_45	WT; columbia	flower stage 15; carpel	21+ days	continuous light	soil
O55	ATGE_73	WT; columbia	pollen	6 weeks	continuous light	soil
O56	ATGE_76	WT; columbia	silique; seed stage 3	8 weeks	long day (16/8)	soil
O57	ATGE_77	WT; columbia	silique; seed stage 4	8 weeks	long day (16/8)	soil
O58	ATGE_78	WT; columbia	silique; seed stage 5	8 weeks	long day (16/8)	soil
O59	ATGE_79	WT; columbia	silique; seed stage 6	8 weeks	long day (16/8)	soil
O60	ATGE_81	WT; columbia	silique; seed stage 7	8 weeks	long day (16/8)	soil
O61	ATGE_82	WT; columbia	silique; seed stage 8	8 weeks	long day (16/8)	soil
O62	ATGE_83	WT; columbia	silique; seed stage 9	8 weeks	long day (16/8)	soil
O63	ATGE_84	WT; columbia	silique; seed stage 10	8 weeks	long day (16/8)	soil

*There are three different biological replicates for each organ type or developmental stage.

Table 2.2. List of 20 different cell types and developmental stages for ATH1 microarray data in the *Arabidopsis* Root Atlas (Birnbaum et al. 2003; Brady et al. 2007)

No; names of .CEL files	Tissue	Line	selection resistance	Replicates #	Ecotype	Background Genotype	Substrate	Plant Age	Reference	Tissues - Radial
Cell Sorting										
C1; LRC_1; LRC_2; LRC_3	LRC (Lateral Root Cap)	J3411	Kan	3	C24	WT	4.5% sucrose	6	Birnbaum et al, 2003	lateral root cap plus epidermis
C2; wol_1; wol_2; wol_3	Stele	WOL	Kan	3	Col	WT	4.5% sucrose	6	Birnbaum et al, 2003	stele
C3; S32_1; S32_2; S32_3	1. Phloem	S32 (JYB697.1)	Basta	3	Col	WT	1% sucrose	6	Brady et al, 2007	protophloem
C4; SUC2_1; SUC2_2; SUC2_3	2. Phloem	SUC2	n.d.*	3	Col	WT	1% sucrose	6	Brady et al, 2007	companion cells
C5; S18_1; S18_2; S18_3	1. Xylem	S18(JYB477.1.3)	Basta	3	Col	WT	1% sucrose	6	Lee et al, 2006	maturing xylem cells
C6; S4_1; S4_2; S4_3	2. Xylem	S4(JYB783.3)	Basta	3	Col	WT	1% sucrose	6	Brady et al, 2007	protoxylem and 2/3 metaxylem
C7; xylem_2501_1; xylem_2501_2; xylem_2501_3	3. Xylem	J2501	Kan	3	C24	WT	4.5% sucrose	6	Brady et al, 2007	metaxylem
C8; J2661_1; J2661_2; J2661_3	Pericycle	J2661	Kan	3	C24	WT	4.5% sucrose	6	Levesque et al, 2006	mature pericyle
C9; J0121_1; J0121_2; J0121_3	1. Pericycle	J0121	Kan	3	C24	WT	4.5% sucrose	6	Brady et al, 2007	xylem pole pericycle
C10; S17_1; S17_2; S17_3	2. Pericycle	S17(JYB408.1.2)	Basta	3	Col	WT	1% sucrose	6	Brady et al, 2007	phloem pole pericycle
C11; J0571_1; J0571_2; J0571_3	G.Tissue (Ground tissue)	J0571	kan	3	C24	WT	4.5% sucrose	6	Birnbaum et al, 2003	ground: endo + cortex + qc
C12; scr5_1; scr5_2; scr5_3	Endodermis	SCR	n.d.	3	1 Ws/2 Ler	WT	4.5% sucrose	6	Birnbaum et al, 2003	endodermis

No; names of .CEL files	Tissue	Line	selection resistance	Replicates #	Ecotype	Background Genotype	Substrate	Plant Age	Reference	Tissues - Radial
C13; E30_1; E30_2; E30_3	M.Endodermis (Mature endodermis)	E30	n.d.	3	Col	WT	1% sucrose	-	-	-
C14; pet111_1; pet111_2; pet111_3	Columella	PET111	Kan	3	Col	WT	4.5% sucrose	6	Nawy et al, 2005	columella - tier 2,3,4
C15; WER_1; WER_2; WER_3	1. Epidermis	WER	n.d.	3	Col	WT	1% sucrose	6	Levesque et al, 2006	LRC and non-hair epidermis
C16; gl2_1; gl2_2; gl2_3	2. Epidermis	GL2	n.d.	3	Col	WT	4.5% sucrose	6	Birnbaum et al, 2003	atrichoblast
C17; CORTEX_1; CORTEX_2; CORTEX_3	Cortex	C1 (JYB315.1.1)	Basta	3	Col	WT	1% sucrose	6	Lee et al, 2006	cortex
Microdissection										
C18; stageI_1; stageI_2; stageI_3; stageI_4	Stage1	n/a	-**	4	Col	WT	4.5% sucrose	6	Birnbaum et al, 2003	all radial tissues
C19; stageII_1; stageII_2; stageII_3; stageII_4	Stage2	n/a	-	4	Col	WT	4.5% sucrose	6	Birnbaum et al, 2003	all radial tissues
C20; stageIII_1; stageIII_2; stageIII_3; stageIII_4	Stage3	n/a	-	4	Col	WT	4.5% sucrose	6	Birnbaum et al, 2003	all radial tissues

*n.d., no data; **, not applicable.

Table 2.3. List of the putative function/function, and the MRCA inference of subfunctionalization and neofunctionalization for reciprocally expressed gene duplicates with asymmetric sequence evolution.

Gene duplicates		Putative function/Function	MRCA		Asymmetric
Gene1	Gene2		ADA	ARA	
WG duplicates					
AT1G07870	AT2G28590	Protein kinase	Neo (2)	-	Neo (2); G1
AT1G55200	AT3G13690	Protein kinase	unknown	-	Neo (1)
AT1G77280	AT1G21590	Protein kinase	-	unknown	Neo (2)
AT4G25160	AT5G51270	Protein kinase	Neo (2)	-	Neo (2); G1
AT5G65600	AT5G10530	Lectin protein kinase	unknown	-	Neo (1)
AT5G03610	AT3G09930	GDSL-motif lipase/hydrolase	Neo (2)	-	Neo (2); G1
AT5G67200	AT3G50230	Leucin-rich repeat transmembrane protein kinase	-	unknown	Neo (2)
AT4G39860	AT2G22270	Unknown protein	Neo (1, 2)	-	Neo (2); G2
AT3G60190	AT2G44590	Dynamin-related protein	Neo (2)	-	Neo (2); G1
AT1G60930	AT1G10930	DNA helicase	Neo (1)	unknown	Neo (1)
AT1G78050	AT1G22170	Phosphoglycerate/biphosphoglycerate mutase	unknown	-	Neo (1)
AT2G02480	AT1G14460	DNA polymerase-related	Neo (2)	-	Neo (2); G1
AT2G18590	AT4G36790	Carbohydrate transmembrane tansporter	Neo (1)	Neo (1)	Neo (1); G1
AT5G44700	AT4G20140	Leucine-rich repeat transmembrane-type receptor kinase	unknown	-	Neo (1)
AT3G59080	AT2G42980	Aspartyl protease	unknown	-	Neo (2)
AT4G28320	AT2G20680	Glycosyl hydrolase	unknown	-	Neo (1)
AT1G35140	AT4G08950	Exordium	Neo (1, 2)	Neo (1)	Neo (1); G2
AT4G14760	AT3G22790	Kinase interacting protein	Neo (1)	-	Neo (1); G1
AT5G66390	AT3G50990	Peroxidase	Neo (2)	-	Neo (2); G1
AT1G70510	AT1G23380	Class I of KN homeodomain transcription factor	Neo (1, 2)	unknown	Neo (1); G2
AT1G02460	AT4G01890	Glycoside hydrolase	-	Sub	Neo (2); G3
AT1G53100	AT3G15350	Acetylglucosaminyltransferase	unknown	-	Neo (1)
AT4G15430	AT3G21620	Unknown protein	-	unknown	Neo (1)
AT1G13270	AT3G25740	Methionine aminopeptidase	Neo (2)	-	Neo (2); G1
AT1G09350	AT1G56600	Galactinol synthase	Sub	Sub	Neo (1); G3
AT2G34940	AT1G30900	Vacuolar sorting receptor	unknown	unknown	Neo (1)
AT5G57580	AT4G25800	Calmodulin-binding protein	unknown	unknown	Neo (2)
AT1G68540	AT1G25460	Oxidoreductase	-	Neo (2)	Neo (2); G1
AT3G10660	AT5G04870	Calcium-dependent protein kinase	Neo (1)	-	Neo (1); G1

Gene duplicates		Putative function/Function	MRCA		Asymmetric
Gene1	Gene2		ADA	ARA	
AT5G14740	AT3G01500	Beta carbonic anhydrase	Neo (1, 2)	unknown	Neo (1); G2
AT1G02050	AT4G00040	Chalcone and stilbene synthase	unknown	Neo (2)	Neo (2); G1
AT1G70710	AT1G23210	Endo-1,4-beta-glucanase	unknown	-	Neo (2)
AT4G24260	AT5G49720	Endo-1,4-beta-glucanase	unknown	-	Neo (1)
AT4G18050	AT5G46540	P-Glycoprotein	Neo (2)	Neo (2)	Neo (2); G1
AT5G06700	AT3G12060	Trichome birefringence-like protein	Neo (1, 2)	-	Neo (2); G2
AT3G53680	AT2G37520	PHD finger transcription factor	-	unknown	Neo (1)
AT1G26310	AT1G69120	MADS-box transcription factor	unknown	-	Neo (1)
AT1G10540	AT1G60030	Xanthine/uracil permease	Neo (1)	Neo (1)	Neo (1); G1
AT2G20340	AT4G28680	Tyrosine decarboxylase	N.D.	-	Neo (2)
AT3G03110	AT5G17020	Exportin protein	N.D.	-	Neo (1)
Tandem duplicates					
AT2G44230	AT2G44260	Unknown protein	unknown	unknown	Neo (1)
AT5G10760	AT5G10770	Aspartyl protease	unknown	-	Neo (1)
AT5G06720	AT5G06730	Peroxidase	-	unknown	Neo (2)
AT3G62000	AT3G61990	O-methyltransferase	unknown	unknown	Neo (2)
AT4G26530	AT4G26520	Fructose-bisphosphate aldolase	Neo (2)	-	Neo (2); G1
AT5G20940	AT5G20950	Glycosyl hydrolase	unknown	-	Neo (1)
AT3G06460	AT3G06470	GNS1/SUR4 membrane protein	-	unknown	Neo (1)
AT5G24900	AT5G24910	Cytochrome P450	-	unknown	Neo (1)

MRCA, the most recent common ancestral expression pattern analysis; ADA, *Arabidopsis* Development Atlas; ARA, *Arabidopsis* Root Atlas; Asymmetric, asymmetric sequence rate analysis; Neo, neofunctionalization; Sub, subfunctionalization; 1, gene1; 2, gene2; -, no detection of reciprocal expression; unknown, unable to infer the most recent common ancestral expression due to an uncertain phylogenetic topology; N.D., the lack of enough information such as a small gene family with two members; G1, group 1 that both MRCA and asymmetric analysis show consistent inference; G2, group 2 that MRCA inferred neofunctionalization for both copies and asymmetric analysis showed neofunctionalization for one copy; G3, group 3 that MRCA inferred subfunctionalization for both copies and asymmetric analysis showed neofunctionalization for one copy.

Table 2.4. List of reciprocally expressed WG duplicates with expression gain in pollen.

No.	Gene1	Gene2	Putative function/Function
1	AT1G07870	AT2G28590	Protein Kinase
2	AT5G03610	AT3G09930	GDSL-like Lipase
3	AT3G20040	AT1G50460	Hexokinase
4	AT4G27730	AT5G53510	Oligopeptide transporter
5	AT5G49180	AT3G06830	Pectin methylesterase
6	AT4G14150	AT3G23670	Microtubule motor kinesin
7	AT4G39860	AT2G22270	Unknown protein
8	AT3G60190	AT2G44590	Dynamin-related protein
9	AT1G18400	AT1G73830	Transcription factor
10	AT2G40300	AT3G56090	Ferritins
11	AT1G52570	AT3G15730	Phospholipase
12	AT5G43900	AT1G04160	Myosin protein
13	AT1G75370	AT1G19650	Phosphatidylinositol transfer protein
14	AT2G44130	AT3G59940	Galactose oxidase
15	AT5G05850	AT3G11330	Leucine rich repeat proteins
16	AT1G21910	AT1G77640	ERF/AP2 transcription factor
17	AT3G19310	AT1G49740	PLC-like phosphodiesterases
18	AT1G53310	AT3G14940	Phosphoenolpyruvate carboxylase
19	AT1G20900	AT1G76500	AT hook domain containing protein
20	AT1G18200	AT1G73640	RAB GTPase
21	AT5G23690	AT3G48830	Polynucleotide adenylyltransferase
22	AT2G23980	AT4G30560	Cyclic nucleotide gated channel protein
23	AT4G01010	AT1G01340	Cyclic nucleotide gated channel protein
24	AT3G08770	AT5G01870	Lipid transfer protein
25	AT4G24970	AT5G50780	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
26	AT1G09600	AT1G57700	Protein kinase
27	AT1G74330	AT1G18670	Protein kinase
28	AT5G37740	AT1G66360	Calcium-dependent lipid-binding protein
29	AT1G18210	AT1G73630	Calcium-binding EF-hand protein
30	AT3G10660	AT5G04870	Calcium-dependent protein kinase
31	AT5G44090	AT1G03960	Calcium-binding EF-hand protein
32	AT1G71050	AT1G22990	Heavy metal transport/detoxification protein
33	AT3G59760	AT2G43750	O-acetylserine (thiol) lyase
34	AT2G41740	AT3G57410	Villin protein
35	AT2G36050	AT3G52540	Ovate protein
36	AT1G75780	AT1G20010	Beta tubulin protein

No.	Gene1	Gene2	Putative function/Function
37	AT4G37760	AT2G22830	Squalene epoxidase
38	AT1G01430	AT4G01080	Trichome birefringence-like protein
39	AT5G06700	AT3G12060	Trichome birefringence-like protein
40	AT1G11990	AT1G62330	O-fucosyltransferase
41	AT1G20550	AT1G76270	O-fucosyltransferase
42	AT1G17790	AT1G73150	DNA-binding bromodomain-containing protein
43	AT1G78090	AT1G22210	Microbial trehalose-6-phosphate phosphatases

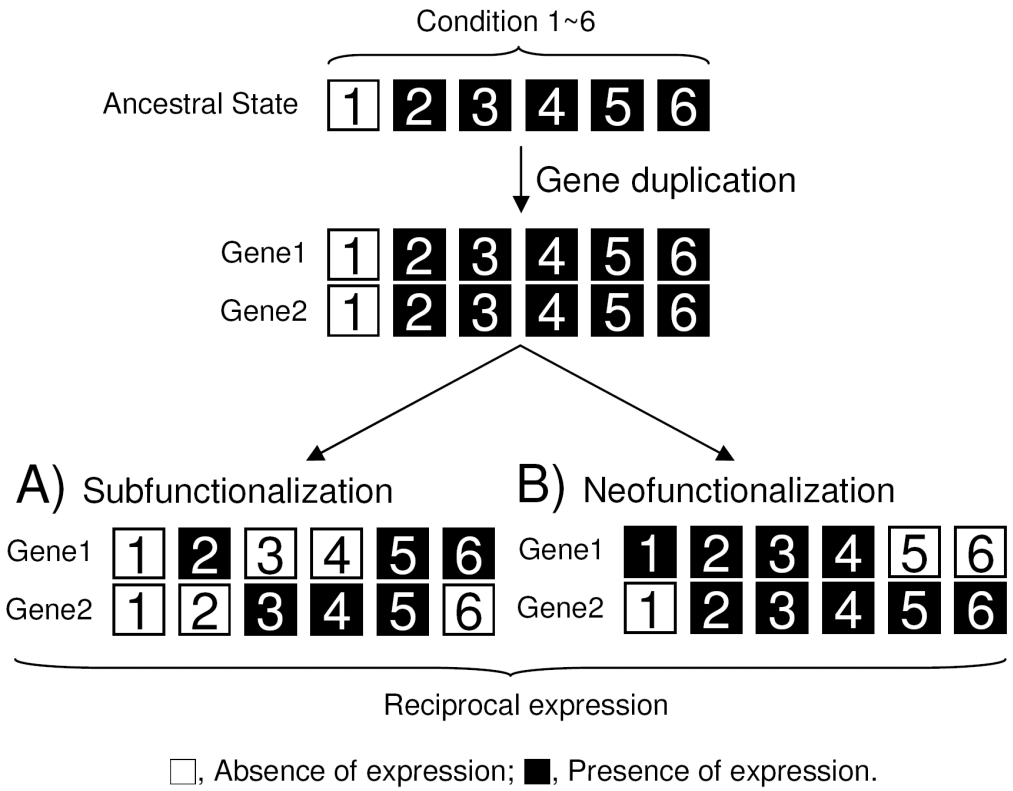
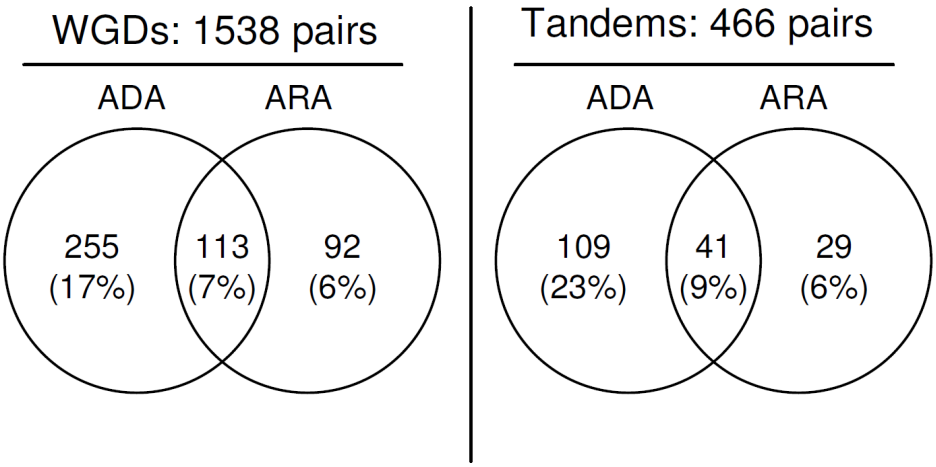


Figure 2.1. Schematics illustrating subfunctionalization and neofunctionalization as evolutionary consequences of reciprocal expression patterns between duplicated genes. Numbers indicate different conditions such as cell types, organ types or developmental stages. (A) Subfunctionalization showing reciprocal expression between the duplicated genes due to the partitioning of the ancestral expression pattern. (B) Neofunctionalization showing reciprocal expression due to the acquisition of a new expression pattern in gene 1 in comparison to the ancestral expression pattern.

A



B

Dataset	WGDs (%)	Tandems (%)	x^2 test (P)
ADA	24	32	0.0003
ARA	13	15	0.3439
Total	30	38	0.0006

Figure 2.2. The frequency of reciprocal expression in WG duplicates and tandem duplicates from the *Arabidopsis* Development Atlas (ADA) dataset and *Arabidopsis* Root Atlas (ARA) datasets. (A) Venn diagram showing the frequency of reciprocal expression among WG duplicates and tandem duplicates. (B) Diagram showing the comparison of reciprocal expression between WG duplicates and tandem duplicates using x^2 test.

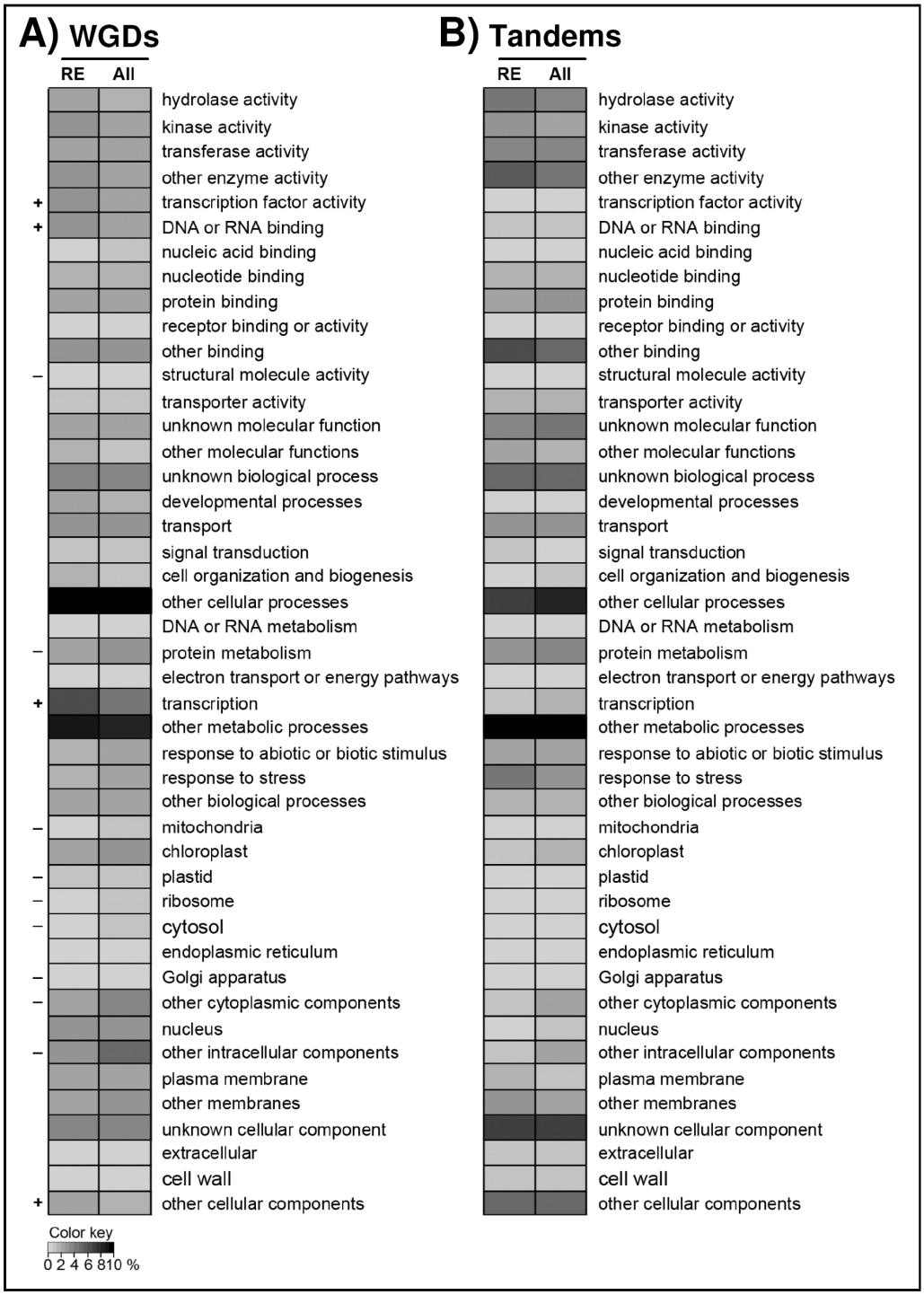


Figure 2.3. A comparison of gene ontology (GO) categories between gene duplicates with reciprocal expression and all gene duplicates among both WG duplicates (A) and tandem duplicates (B). Plus signs and minus signs indicate that the gene ontology is overrepresented and underrepresented, respectively, among gene pairs with reciprocal expression. RE indicates reciprocal expression.

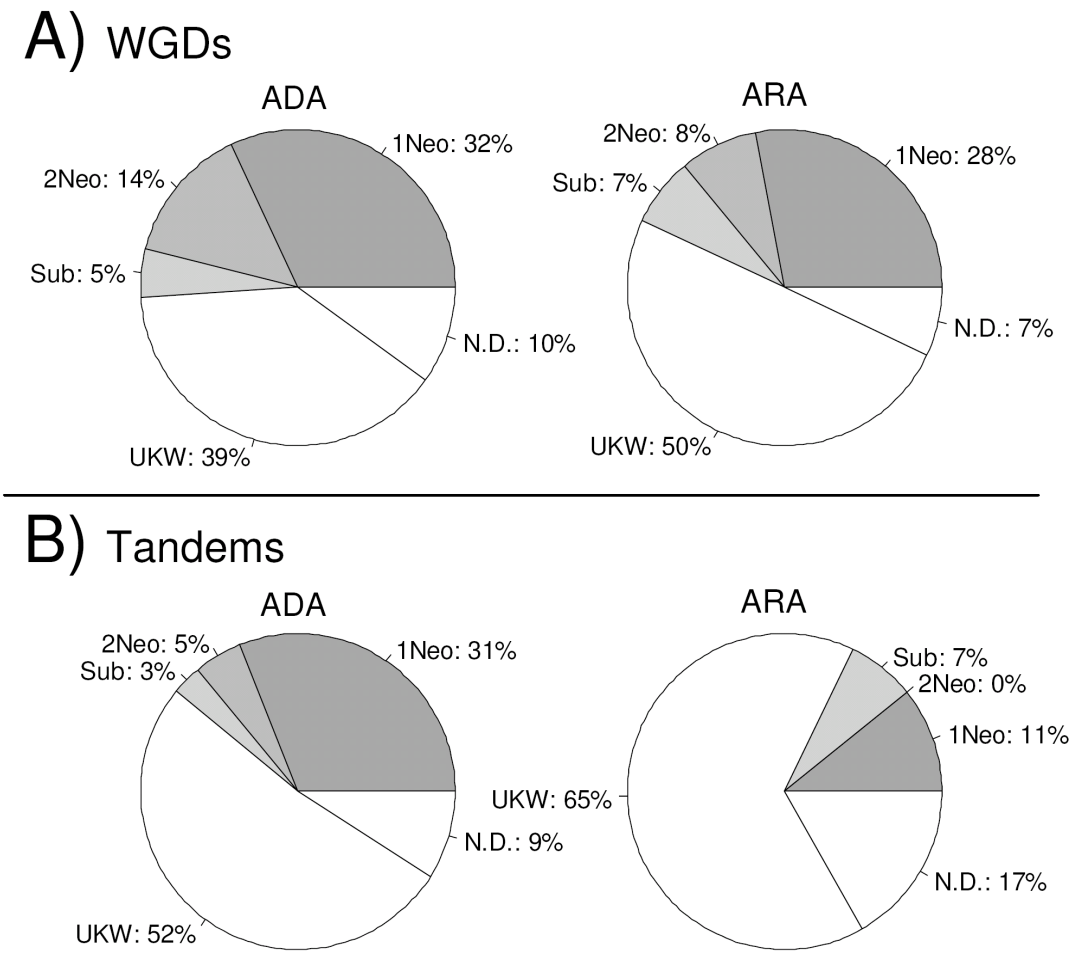


Figure 2.4. The relative frequency of subfunctionalization and neofunctionalization of expression patterns. Neofunctionalization and subfunctionalization were inferred by MRCA analysis in both WG duplicates (A) and tandem duplicates (B) from both the *Arabidopsis* Development Atlas (ADA) and *Arabidopsis* Root Atlas (ARA) datasets. Abbreviations: 1Neo, neofunctionalization for one copy; 2Neo, neofunctionalization for both copies; Sub, subfunctionalization; UKW, unknown due to uncertain tree topology; N.D., not determined due to small gene family size with only two members.

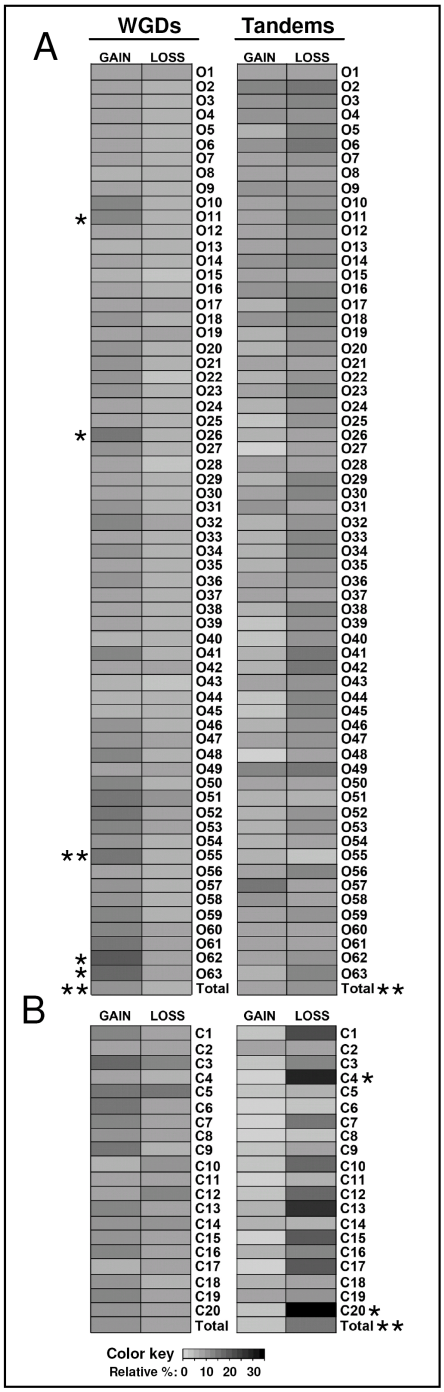


Figure 2.5. Expression gains and losses by organ type, developmental stage, and cell type. Summary of chi square analyses between expression gain and expression loss from MRCA analysis in each condition (i.e., organ type, developmental stage, or cell type). A statistically significant difference after 5% false discovery rate correction is labeled with a star (*: $Q < 0.05$; **: $Q < 0.01$). (A) A comparison of the relative percentage between expression gain and expression loss in 63 different developmental stages and organ types in both WG duplicates and tandem duplicates. The 63 different developmental stages and organ types are listed in Table 2.1. (B) A comparison of the relative percentage between expression gain and expression loss in 20 different developmental stages and cell types in roots. The 20 different developmental stages and cell types of roots are listed in Table 2.2.

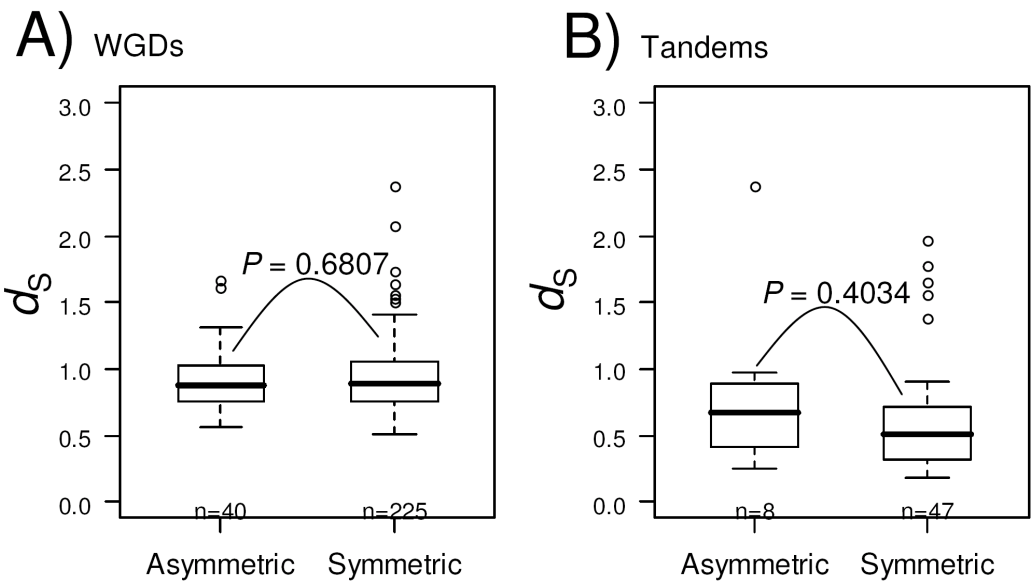


Figure 2.6. Box plots showing a comparison of the synonymous substitution rate (d_S) between gene duplicates with asymmetric sequence evolution and those without asymmetric sequence evolution. (A) WG duplicates. (B) Tandem duplicates.

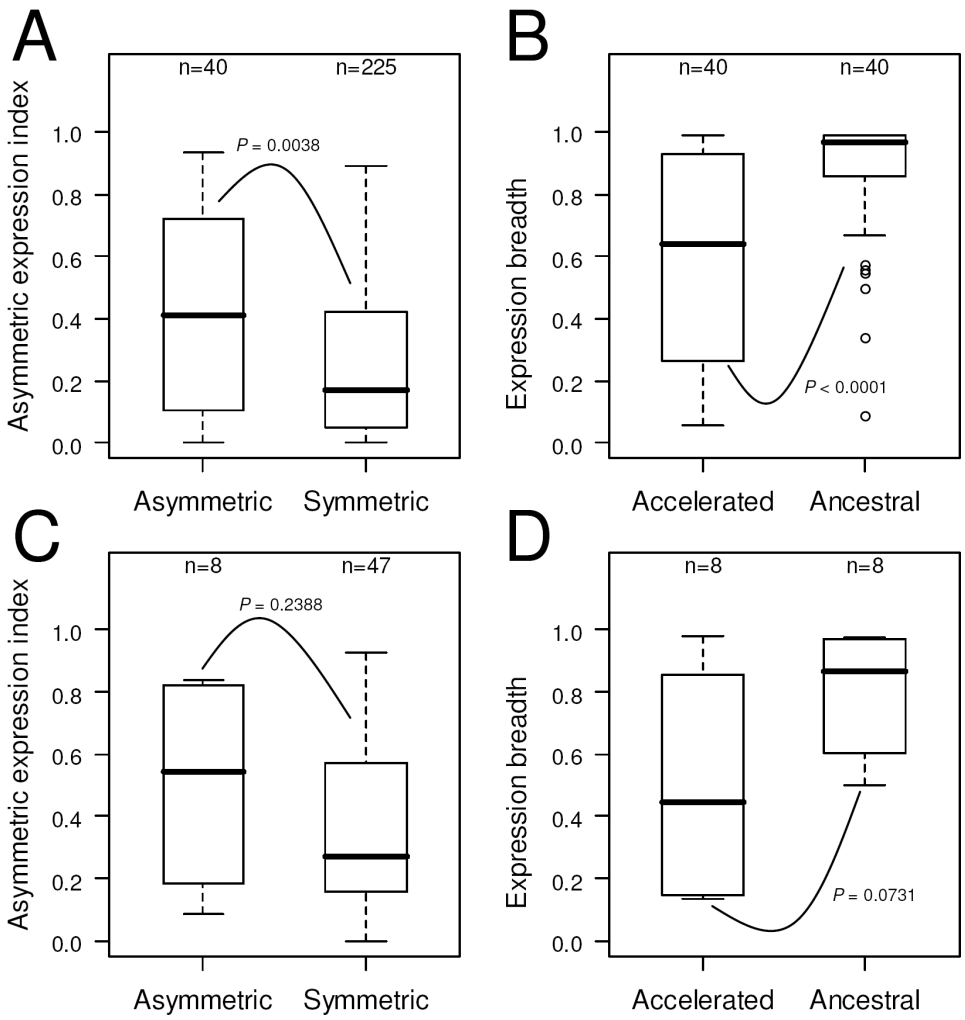


Figure 2.7. Asymmetric sequence evolution is associated with asymmetric expression divergence. (A) A comparison of the asymmetric expression index between asymmetrically evolved pairs and symmetrically evolved pairs among the WG duplicates. (B) A comparison of expression breadth between the accelerated copy and the non-accelerated copy among the WG duplicates. (C) A comparison of the asymmetric expression index between asymmetrically evolved pairs and symmetrically evolved pairs among the tandem duplicates. (D) A comparison of the expression breadth between the accelerated copy and the non-accelerated copy among the tandem duplicates.

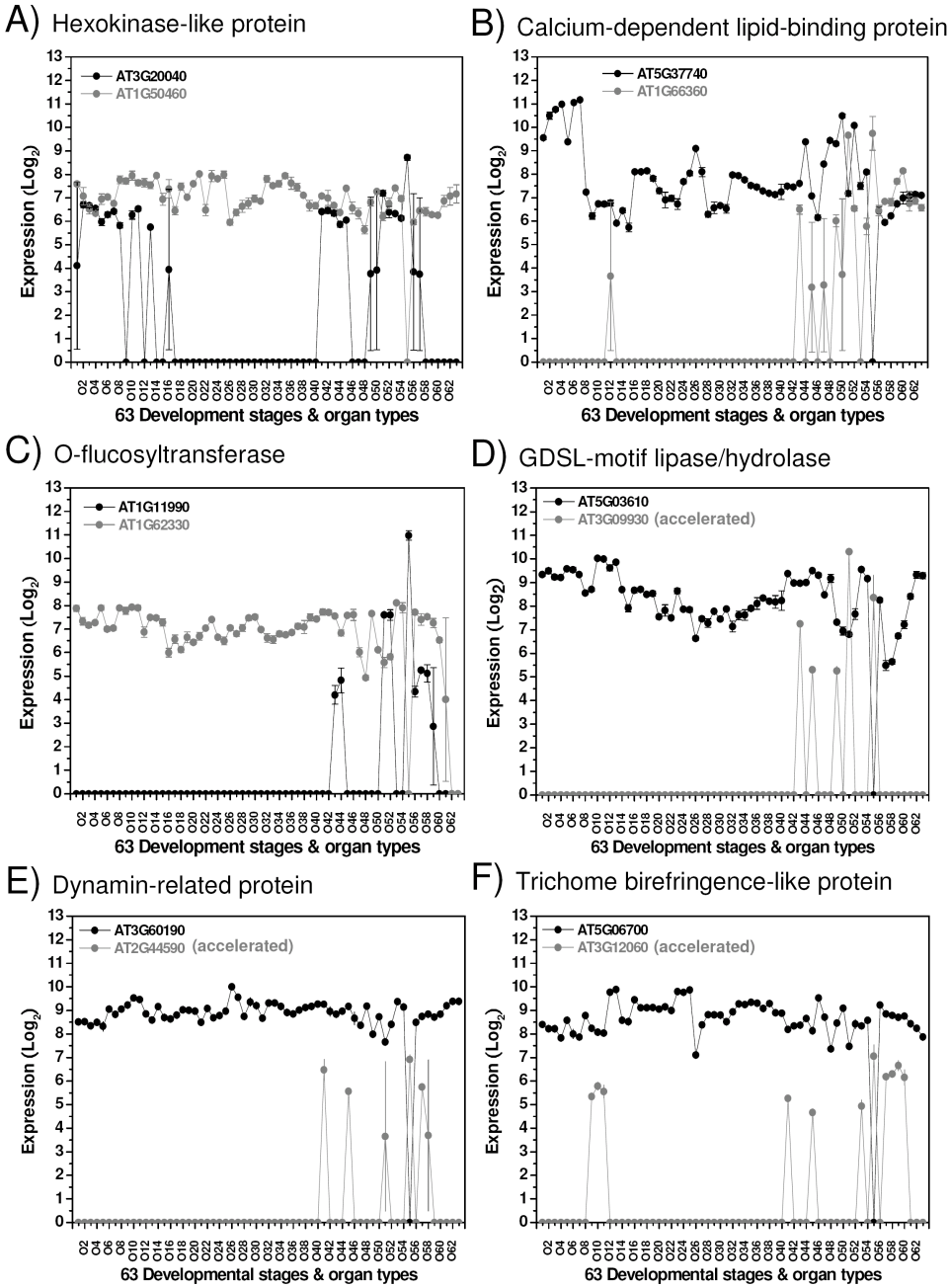


Figure 2.8. Reciprocal expression involving pollen (O55 in x-axis). Diagrams showing some striking reciprocally expressed gene duplicates in which one copy showed a restricted expression pattern and gain of expression in pollen (A-F), plus accelerated sequence evolution (D-F). MAS5-normalized microarray gene expression data from 63 different developmental stages and organ types. Absence or presence of expression was determined by using mas5calls function in software Bioconductor. Error bars indicate standard deviations (n = 3). The 63 different developmental stages and organ types are listed in Table 2.1. (A) A pair of hexokinase-like genes (AT3G20040 and AT1G50460). (B) A pair of calcium-dependent lipid-binding protein genes (AT5G37740 and AT1G66360). (C) A pair of O-flucosyltransferase genes (AT1G11990 and AT1G62330). (D) A pair of GDSL-motif lipase/hydrolase genes (AT5G03610 and AT3G09930). (E) A pair of dynamin-related genes (AT3G60190 and AT2G44590). (F) A pair of trichome birefringence-like genes (AT5G06700 and AT3G12060).

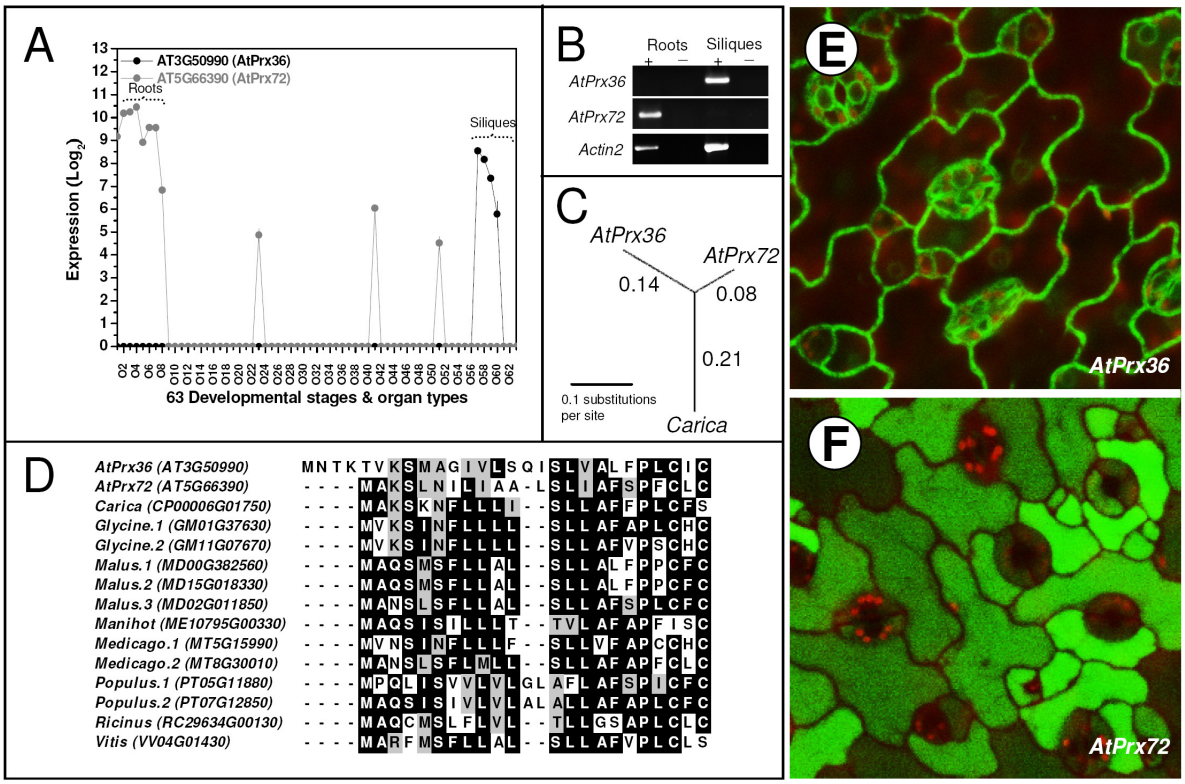


Figure 2.9. Differential subcellular localization and neofunctionalization of a peroxidase gene. (A) MAS5-normalized microarray gene expression data of *AtPrx36* (AT3G50990) and *AtPrx72* (AT5G66390) from 63 different developmental stages and organ types. Absence or presence of expression was determined by using mas5calls function in software Bioconductor. Error bars indicate standard deviations (n = 3). The 63 different developmental stages and organ types are listed in Table 2.1. (B) Reverse transcription (RT)-PCR expression assays of *AtPrx36* and *AtPrx72* in roots and siliques. Plus signs (+) indicate reactions with reverse transcriptase and minus signs (-) indicate reactions without reverse transcriptase. (C) Maximum likelihood tree showing that *AtPrx36* evolved faster than *AtPrx72*. (D) Alignment of N-terminal region of *AtPrx36*, *AtPrx72*, and other homologous/orthologous sequences from outgroup species. (E) GFP subcellular localization of AtPrx36 in cell wall (green fluorescence). (F) GFP subcellular localization of *AtPrx72* in cytoplasm (green fluorescence).

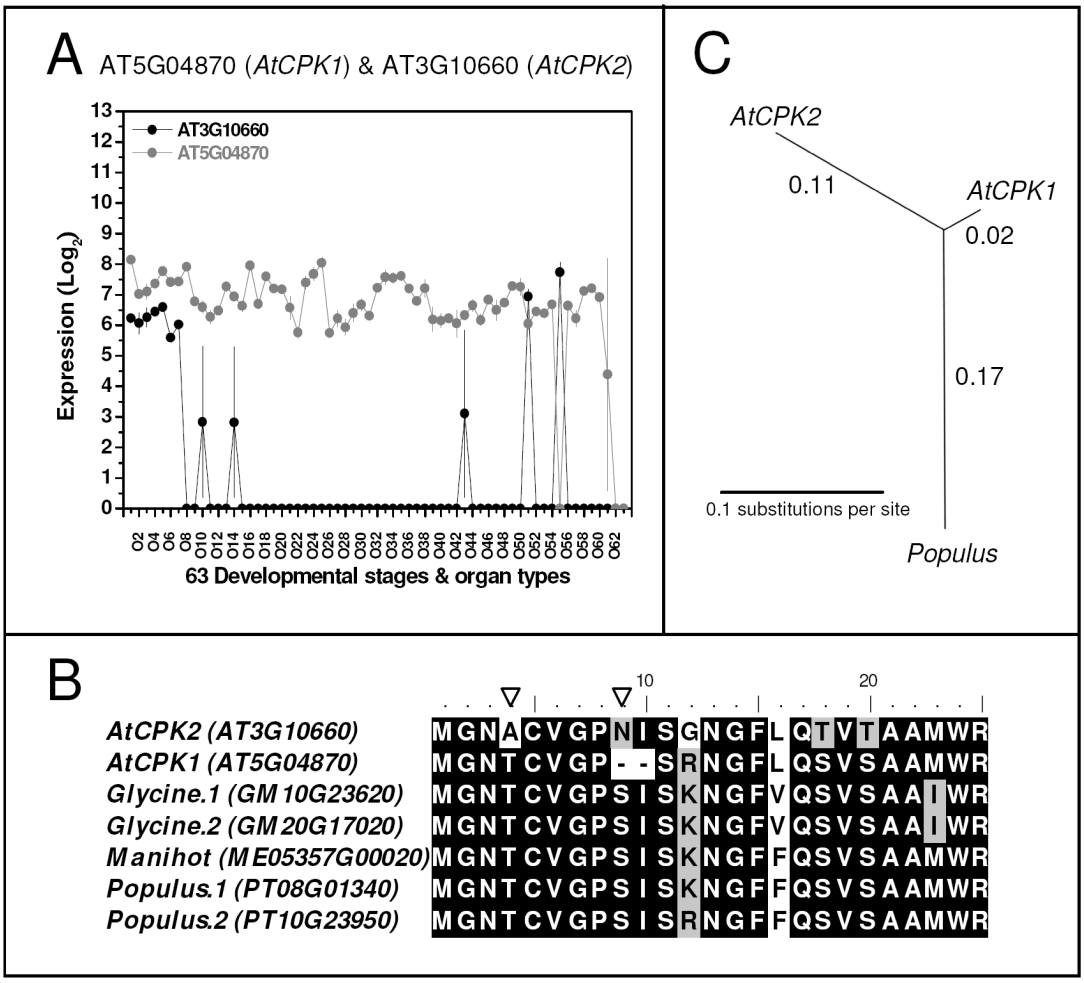


Figure 2.10. Differential subcellular localization and neofunctionalization in a pair of calcium-dependent protein kinase genes. (A) MAS5-normalized microarray gene expression data of *AtCPK1* (AT5G04870) and *AtCPK2* (AT3G10660) from 63 different developmental stages and organ types. Absence or presence of expression was determined by using mas5calls function in software Bioconductor. Error bars indicate standard deviation (n = 3). The 63 different developmental stages and organ types are list in Table 2.1. (B) Alignment showing the N-terminal targeting region. Arrowheads indicate gene-specific amino acid changes in *AtCPK2*. (C) Maximum likelihood tree showing that *AtCPK2* evolved faster than *AtCPK1*.

3 Dramatic Changes in Function and Expression Pattern of a Gene Duplicated by Polyploidy Created a Paternal Effect Gene in the Brassicaceae[1]

3.1 Introduction

Whole genome duplication (WGD), or polyploidy, has been an ongoing process during eukaryotic evolution, with polyploidy events having occurred during the evolution of fish, frogs, yeasts, and flowering plants, among other groups (Otto and Whitton 2000; Wolfe 2001; Seoighe 2003; Jaillon et al. 2009; Van de Peer et al. 2009). Almost all angiosperms show evidence for at least one round of WGD sometime during their evolutionary history, with many plants having had multiple polyploidy events occur during the evolution of their lineage (Cui et al. 2006; Soltis et al. 2009). In addition to polyploidy, duplicated genes can be formed by segmental duplications of multiple genes along one chromosome, tandem duplication of individual genes, and duplicative retroposition. All the types of gene duplication have contributed greatly to the large number of genes in many eukaroytic genomes. After formation by duplication, the functions of duplicated genes can diverge by the acquisition of new function, neofunctionalization (Ohno 1970), or partitioning of ancestral function, subfunctionalization (Hughes 1994; Force et al. 1999). Expression patterns of duplicated genes can diverge by changes in gene regulation, including gain of a new expression pattern relative to the ancestral state or partitioning of an

1 A version of chapter 3 has been published. **Liu, S.-L.**, Adams, K.L. (2010) Dramatic change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the Brassicaceae. Molecular Biology and Evolution 27: 2817–2828.

ancestral expression pattern between the duplicates, also referred to as neofunctionalization and subfunctionalization, respectively (Force et al. 1999). Functional and expression divergence are widely regarded as important mechanisms for the retention of duplicated genes.

In the genome of *Arabidopsis thaliana*, there have been at least three rounds of ancient WGD events during the evolution of its lineage, termed α-, β-, and γ-WGD events. Among them, α is specific to the Brassicaceae family, and its timing is likely at the base of the Brassicaceae family (Barker et al. 2009), β is specific to the Brassicales order after the divergence of the Brassicaceae and Caricaceae from a common ancestor, whereas γ is an older WGD event that is presumably eudicot specific (Blanc et al. 2003; Bowers et al. 2003; Jaillon et al. 2007; Ming et al. 2008). Expression patterns of genes duplicated by WGD and by other smaller scale mechanisms have been examined in a range of organ types, developmental stages, and stress conditions from published microarray data sets. Several studies have shown that there has been considerable divergence in expression patterns across different organs and treatments, with over half of the duplicated pairs examined in each case showing significant divergence in expression patterns between duplicates (Blanc and Wolfe 2004b; Haberer et al. 2004; Casneuf et al. 2006; Duarte et al. 2006; Ganko et al. 2007; Ha et al. 2007; Zou et al. 2009). Expression divergence plus accelerated and asymmetric sequence evolution (i.e., a much faster rate of sequence evolution in one duplicate compared with the other) have been interpreted as evidence for functional divergence (Blanc and Wolfe 2004b). However, there are relatively few cases of experimentally demonstrated gain of a new function of duplicated genes during the evolution of the Brassicaceae family. Examples of neofunctionalization that have been reported include the nitrilase genes *NIT1* and *NIT4*, where *NIT1* has a new function and accelerated rate of sequence evolution, but expression patterns are similar (Blanc and Wolfe 2004b), and the mercaptopyruvate sulfurtransferases *AtMST1* and *AtMST2* that have a different subcellular

localization, to the mitochondria or cytoplasm (Nakamura et al. 2000), but similar expression patterns. Another example is the gene pair *MEDEA* and *SWINGER* that are differentially imprinted in the endosperm (*MEDEA* is paternally imprinted and SWINGER is not imprinted), have largely overlapping, but not identical, expression patterns (Spillane et al. 2007), and different but partially redundant functions (Wang et al. 2006). A case of neofunctionalization after duplicative retroposition is *CYP98A8*/*CYP98A9* compared with *CYP98A3*. Neofunctionalization of *CYP98A8*/*CYP98A9* led to a novel phenolic pathway in pollen of *A. thaliana*, and the genes' expression patterns are mostly limited to flowers, in contrast to *CYP98A3* which is expressed in most organs but not pollen (Matsuno et al. 2009). Less well documented are cases of neofunctionalization that show gain of a new function, elimination of the old function, gain of expression in new organ types, and loss of expression in other organ types by one of the duplicates, yet such cases likely involve some of the most dramatic changes in function after gene duplication.

In this study, I identified that the *SHORT SUSPENSOR* (*SSP*) gene (Bayer et al. 2009) and the *Brassinosteroid Kinase 1* (*BSK1*) gene (Tang et al. 2008) are paralogs derived by the α-WGD at the base of the Brassicaceae family. I present analyses of gene expression and sequence evolution indicating that *SSP* has undergone neofunctionalization from being involved in brassinosteroid signal transduction to regulating the timing of zygote elongation by a unique paternal effect mechanism involving transcription in sperm cells of the pollen and translation in the zygote. In addition, I analyzed a duplicated copy of *SSP*, *SSP-like1*, which also has undergone neofunctionalization, and another duplicated copy of *SSP*, *SSP-like2*, which might become a psudogene.

3.2 Materials and Methods

3.2.1 Microarray Data Analysis

The Arabidopsis ATH1 microarray data from 63 different developmental stages and organ types (#ME00319) (Schmid et al. 2005) and the Soybean Genome Array data (#GSE12286) (Haerizadeh et al. 2009) were obtained from the Gene Expression Omnibus at National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). The 63 different developmental stages and organ types used in *Arabidopsis* ATH1 microarray data are listed in Table 2.1. Raw CEL files were processed and normalized using the MAS5.0 algorithm in Bioconductor (http://www.bioconductor.org/). To determine the absence or presence of expression, the "mas5calls" function in Bioconductor was implemented in the statistical package R (http://www.r-project.org/). This statistical procedure performs the Wilcoxon signed rank-based gene expression presence/absence detection algorithm and yields a detection call (i.e., a detection P value) to access if the detected transcript is significantly greater than background signal noise. A gene with a P value less than 0.05 is marked as a "presence," whereas a gene with a P value equal to or greater than 0.05 is marked as an "absence." Because each gene involves three replicates, only those assigned as absence of expression in at least 2 of 3 replicates were given an absence call.

3.2.2 Phylogenetic Analysis of the *BSK* Gene Family

To identify the orthologous *BSK1*-like genes in outgroup species, I implemented a phylogenetic analysis of the *BSK* gene family with a focus on species with genome sequence data available,

including 4 eudicot species (*A. thaliana*, *Carica papaya*, *Populus trichocarpa*, and *Vitis vinifera*), 2 monocot species (*Oryza sativa* and *Sorghum bicolor*), and 1 moss species (*Physcomitrella patens*). The sequences were retrieved from a Blast search with default settings in the Web site Plaza (http://bioinformatics.psb.ugent.be/plaza/) (Proost et al. 2009). In addition, I also included orthologous *BSK1*-like genes from *Brassica rapa* subsp. *pekinensis*, *Cleome spinosa*, *Gossypium hirsutum*, *Glycine max*, and *Helianthus annuus* in order to increase taxon sampling for the orthologous gene expression assay. Using *BSK1* in *A. thaliana* as query, the orthologous *BSK1*-like sequences from *Brassica*, *Cleome*, *Gossypium*, *Glycine*, and *Helianthus* were obtained from Blast searches (at least greater than 80% identity) of GenBank at NCBI (http://www.ncbi.nlm.nih.gov/). The *SSP* and *BSK1* genes from *A. lyrata* were obtained from Blast searches of Phytozome v4.0 at the Joint Genome Institute (http://www.phytozome.net/). Prior to phylogenetic analysis, sequences were aligned using TransAlign with the ClustalW program (Bininda-Emonds 2005) and manually checked using the program Bioedit (Hall 1999). A very divergent, short, sequence region at the 5′ end was removed and then the remainder of the gene sequence was used for further phylogenetic analysis. Phylogenetic analysis was performed with a Bayesian method using MrBayes v3.1.2, described in section 3.2.4, and with a maximum likelihood (ML) method using Garli (Zwickl 2006). The nucleotide substitution model was automatically estimated from the empirical data. Statistical support for nodes was determined using bootstrapping with 100 ML replicates from Garli and 50% majority rule of 250 sampled trees (standard deviation of split frequency < 0.01) from MrBayes v3.1.2.

3.2.3 Plant Materials, Nucleic Acid Extraction, and RT-PCR

RNA was extracted from roots, stems, rosettes, leaves, flowers, ovules, or pollen from the following species: *Arabidopsis thaliana* (ecotype Columbia), *B. rapa* subsp. *pekinensis* (Chinese cabbage variety, MU525B, West Coast Seeds), *C. papaya* (cultivar Sun-Up), *H. annuus* (wild population from Utah), *G. hirsutum* (cultivar Maxxa), and *V. vinifera* (cultivar Pinot Noir). *Brassica rapa* plants were subjected to 4 °C for 2 months to stimulate bolting and flower production. Pollen from *Carica* and *Helianthus* was collected by tapping the flowers on a piece of paper and pollen from *Brassica* and *Vitis* was collected by using a vacuum cleaner method (Johnson-Brousseau and McCormick 2004). Collected pollen materials were examined for purity under light microscopy. Nucleic acid extraction and reverse transcription-polymerase chain reaction (RT-PCR) conditions followed those in Liu and Adams (2008). For RT-PCR, 25–35 reaction cycles were applied to assay gene expression level differences in different organ types. Gene-specific primers are listed in Table 3.1. New sequences determined in this study were deposited in GenBank: *G. hirsutum* partial *BSK1.3* cds. and *C. papaya* partial *BSK11* cds. with accession numbers GU321198 and GU321199, respectively.

3.2.4 Reconstruction of the Most Recent Common Ancestral (MRCA) Expression Pattern

Prior to the reconstruction of the most recent common ancestral (MRCA) expression state between *SSP* and *BSK1*, I first obtained other members of the *BSK* gene family in *Arabidopsis thaliana* by performing an all-against-all BLASTn search with the following criteria: >50% identity, > 200 nt length, and e value < 1e-02. To retrieve all members of *BSK* gene family from

the BLASTn search, I then clustered them based on the transformed e-value using a Markov Clustering program (http://micans.org/mcl/). In total, I identified 14 genes in the family, including all of those from Shiu et al. (2004) plus *SSP-like2* (AT2G17160) which is missing much of the coding region and likely is a pseudogene (see Results section). I then implemented a phylogenetic analysis by using the 13 genes (excluding AT2G17160) to infer the MRCA expression state of *SSP* and *BSK1*. Nucleotide sequences were aligned using TransAlign with the ClustalW program (Bininda-Emonds 2005), and manually checked using the program Bioedit (Hall 1999). A very divergent sequence region at the 5' end was removed and the well-conserved remainder of the gene was used for further phylogenetic analysis. A gene family tree was generated using a Bayesian analysis with the program MrBayes v.3.1.2 (Huelsenbeck and Ronquist 2001). Briefly, two parallel runs were executed, and each run consisted of four chains (MCMC sampling; one hot and three cold) and one tree was sampled every 1,000 generations for 1,000,000 generations. In total, 1,000 samples (or trees) were obtained and 25% of samples were the "burn-in" of the chain. After 250,000 generations, trees were saved and 250 of them with standard deviation of split frequency < 0.01 were used later for the MRCA analysis. For the MRCA analysis, I followed the analytical procedure as described in Zou et al. (2009). Reconstruction of the MRCA expression state was conducted with the maximum likelihood algorithm using the program MultiState in the package BayesTraits v.1.0 (Pagel and Meade 2009). It has been shown that using a maximum likelihood method, with a gene family phylogeny, is useful to reconstruct ancestral expression pattern between duplicated genes (Gu et al. 2005; Oakley et al. 2006; Zou et al. 2009). To take the uncertainty of the phylogenetic tree topology and branch length into account, 250 trees (standard deviation of split frequency < 0.01) deduced from previous Bayesian analysis and 500 bootstrapping trees from maximum likelihood analyses using the software Phylip v.3.68 (Felsenstein 2008) were imported into BayesTraits. Prior to the analyses, each tree was rooted at the midpoint using the program Reroot in the

software Phylip v.3.68. Two evolutionary transition rates comprising forward and reverse transition were used for estimating the character transition rate. When applying the reconstruction of the most recent ancestral expression state, 0 is designated as the absence of expression and 1 is designated as the presence of expression. The "*AddMRCA*" function was used to define the MRCA node of *BSK1* and *SSP* (Pagel and Meade 2009). The ancestral state probability was averaged across 250 different phylogenetic trees. The average of ancestral state probability that is greater than 0.60 was considered as a convincing inference of the ancestral expression state because that will follow the 60% majority consensus rule.

3.2.5 Selection Analysis

To test if there has been evidence of accelerated sequence evolution or positive selection acting on *SSP*, I used a phylogeny-based approach to examine if there has been any sequence rate acceleration or positive selection acting on *SSP* or *BSK1* using PAML (Yang 2007). Orthologous sequences from *P. trichocarpa* and *V. vinifera* were retrieved from CoGe (http://synteny.cnr.berkeley.edu/CoGe) (Lyons et al. 2008), and orthologous sequences from *C. papaya* were obtained from the Web site Plaza (http://bioinformatics.psb.ugent.be/plaza/) (Proost et al. 2009) based on collinear syntenic analyses. Pairwise ω (d_N/d_S)-ratio analysis, protein sequence rate acceleration, and positive selection were implemented using the program Codeml in PAML. For pairwise ω-ratio analysis, nonsynonymous (d_N) and synonymous (d_S) nucleotide substitution analysis was implemented using maximum likelihood in Codeml. For the detection of sequence rate acceleration, I followed the analytical procedure described in Spillane et al. (2007). One ω-ratio model, two ω-ratio model, three ω-ratio model, and free ω-ratio model branch models were implemented. The first model assumes that only one ω-ratio leads to whole

phylogeny branches; the second model assumes that one ω-ratio leads to the pro-ortholog branches and the *BSK1* branch, and another ω-ratio leads to the *SSP* branch; the third model assumes that three different ω-ratios lead to pro-ortholog branches, *BSK1* branch, and *SSP* branch, respectively; and the last model allows different ω-ratios for each branch of the phylogeny. Then, twice the difference of their likelihood ratio between any two models (likelihood ratio test [LRT]) was compared against a chi-square (x^2) distribution. The degree of freedom (d.f.) was obtained based on the difference of parameters used in any two models. In the detection of positive selection, two branch-site models, model A test1 and model A test2, were implemented, and the LRT was conducted against a chi-square (x^2) distribution with the 50:50 mixture of df = 0 and df = 1. Results from the branch-site model can allow us to evaluate which specific codons along sequence underwent positive selection ($\omega > 1$). In this study, I applied a branch-site model to detect positive selection on the *SSP*, *SSP-like1*, and *BSK1* genes. To correct multiple testing, a 5% false discovery rate control was used (Anisimova and Yang 2007).

3.2.6 *Cis*-Regulatory Element Analysis

Sequence up to 723 bases upstream of *SSP*, *SSP-like2*, and *BSK1* from *A. thaliana* were searched against the PlantCARE cis-element database (Lescot et al. 2002) to identify predicted cis-regulatory elements and compare them among the three genes. To determine if there has been any insertion of transposable elements in the cis-element region, I performed Blast searches of CENSOR in the Repeat Masking of giri (http://www.girinst.org/censor/index.php) (Kohany et al. 2006).

3.3 Results

3.3.1 *SSP* and *BSK1* Are Whole Genome Duplicates with Completely Different Organ-Specific Expression Patterns

While studying genes duplicated by the most recent WGD during the evolutionary history of the Brassicaceae (Blanc et al. 2003; Bowers et al. 2003), also known as the α-WGD, I noticed that the *SSP* gene (locus AT2G17090) and the *BSK1* gene (locus AT4G35230) are paralogs. *SSP* is in a duplicated block on chromosome 2 and *BSK1* is in a duplicated block on chromosome 4, shown in Figure 3.1. The two genes have very different functions: *SSP* regulates the timing of elongation of the embryo by activating the *YODA* signaling pathway, using a paternal control mechanism involving transcription in sperm cells of the pollen followed by translation only in the embryo (Bayer et al. 2009). In stark contrast, *BSK1* is part of the brassinosteroid signal transduction pathway (Tang et al. 2008). It is phosphorylated by the brassinosteroid receptor *BRI1* and it phosphorylates *BSU1* (Kim et al. 2009). Thus, there has been a change in function in one or both genes after gene duplication.

To determine how the expression patterns of *SSP* and *BSK1* have evolved since gene duplication, I compared the expression patterns of *BSK1* and *SSP* using Affymetrix ATH1 microarray data from 63 different organ types and developmental stages in *A. thaliana* (Schmid et al. 2005) and RT-PCR expression assays in *B. rapa*. In *Arabidopsis*, *BSK1* was highly expressed in every organ type and developmental stage except for pollen where it is not expressed (Figure 3.2A). In complete contrast, *SSP* showed expression above background only in pollen (Figure 3.2A); the *SSP* results are consistent with data and analysis from Bayer et al. (2009). In *B. rapa*, one copy

of *SSP* and two copies of *BSK1* were identified based on my phylogenetic analysis (Figure 3.3). Both copies of *BSK1* were highly expressed in every organ type that I examined except for pollen where neither is expressed, whereas *SSP* showed expression only in pollen (Figure 3.2B). Overall, the organ-specific expression patterns of *SSP* and *BSK1* are completely different and exactly opposite, and they are consistent between *Arabidopsis* and *Brassica*.

3.3.2 *BSK1* Reflects the Ancestral Expression Pattern and Function

The dramatic difference in expression patterns between *SSP* and *BSK1* could be due to partitioning of the ancestral, preduplication, expression pattern between *BSK1* and *SSP*, if the ancestral state was expression in all organs, that would be an example of subfunctionalization of expression patterns. Alternatively, either *BSK1* or *SSP* could retain the ancestral expression pattern, with the other gene having undergone a complete change to gain a new expression pattern (i.e., neofunctionalization). To distinguish among these possibilities, I assayed expression of orthologous genes from outgroup species that diverged before the Brassicaceae-specific WGD. To identify the orthologs, I first reconstructed the phylogenetic relationships of the *BSK* gene family among the sequenced genomes from *A. thaliana*, *C. papaya* (papaya), *P. trichocarpa* (poplar), *V. vinifera* (grape), *O. sativa* (rice), and *S. bicolor* (sorghum) by using *P. patens* (a moss) as an outgroup (Figure 3.3). Sequences from additional eudicots were included for putative orthologs of *BSK1* and *BSK11*. The *BSK* gene phylogeny showed that there have been several rounds of gene duplication events during *BSK* gene family evolution in angiosperms. There are two major clades of genes related to *BSK1* that formed after the divergence of monocots and eudicots: the *BSK1* group and the *BSK11* group (Figure 3.3). The *BSK11* group is well supported as a clade to the exclusion of the *BSK1* sequences, but some relationships within the *BSK1* group

are not well resolved. Nevertheless, within the *BSK1* clade, I identified orthologous genes from *C. spinosa*, *C. papaya*, *G. hirsutum*, *G. max*, *P. trichocarpa*, *V. vinifera*, and *H. annuus* (Figure 3.3).

After identifying *BSK*1-orthologous genes from outgroup species, I then assayed their expression pattern in pollen and in multiple organ types in *C. papaya* (papaya), *G. hirsutum* (cotton), *V. vinifera* (grape), and *H. annuus* (sunflower) by using RT-PCR. All the genes showed expression in various organ types but no expression in pollen (Figure 3.4A). In addition, I analyzed Affymetrix microarray data from *G. max* (soybean) (Haerizadeh et al. 2009). Expression of the *BSK1*-like gene in soybean was below background in pollen but at relatively high levels in all other organ types (Figure 3.4B). Thus, the *BSK1*/*SSP* orthologs in papaya, cotton, grape, soybean, and sunflower all show similar expression patterns to *BSK1* not to *SSP*. These observations indicate that the preduplication expression pattern of *SSP* and *BSK1* is no expression in pollen but expression in other organs. To infer the preduplication expression pattern of *SSP* and *BSK1* using another approach, I examined their most recent common ancestral (MRCA) expression state by implementing a ML method using the phylogeny of the *BSK* gene family in *A. thaliana* to reconstruct the MRCA expression state (Figure 3.5). The results were consistent with the outgroup expression analysis.

The above results indicate that *BSK1* exhibits the ancestral expression state and potentially the ancestral function. Further support for the ancestral function comes from examining other members of the BSK gene family in a phylogenetic context. The *BSK* genes *BSK2*, *BSK3*, and *BSK5* have been functionally characterized as being involved in brassinosteroid signal transduction (Tang et al. 2008). All of those genes branch as an outgroup to the clade containing *BSK1* and *SSP* (Figure 3.3), strongly suggesting that an ancestral function of the BSK genes was

involvement in brassinosteroid signal transduction, not in regulating embryo elongation and division.

3.3.3 Loss of the Original Function of *SSP* by Mutations in the Kinase Activation Domain

SSP and *BSK1* have two major functional domains: a protein kinase domain and a tetratricopeptide repeat (TPR) domain (Figure 3.6). The protein kinase domain is responsible for the catalytic activity, and the TPR domain is involved in protein–protein interactions. Mutations in these two regions have been shown to be detrimental for the function of *SSP* (Bayer et al. 2009). When compared with other orthologs, there is a deletion in the activation loop of the protein kinase domain in *SSP*, which resides in the substrate-binding pocket (Figure 3.6). In addition, *SSP* has a one nucleotide deletion in a codon corresponding to a serine in *BSK1* that is critical for its function (Figure 3.6). Serine-230 in *BSK1* is the major site of phosphorylation by the brassinosteroid receptor kinase *BRI1* and mutations result in an 82% reduction in phosphorylation (Tang et al. 2008). *BRI1* phosphorylation of *BSK1* at Ser-230 promotes *BSK1* binding to the *BSU1* phosphatase and continuation of the signal transduction cascade (Kim et al. 2009). Mutation of the serine abolished binding of *BSK1* to *BSU1* (Kim et al. 2009). Those analyses indicate that deletion mutations in *SSP* have resulted in loss of the original *BSK1* function in brassinosteroid signal transduction. The complete change in expression pattern of *SSP* also probably contributed to loss of the original function because *SSP* is not expressed in most of the organs where the brassinosteroid receptor *BRI1* that starts the signal transduction cascade is expressed (*BRI1* does not appear to be expressed in pollen, from microarray data determined by Schmid et al. [2005]).

3.3.4 Rapid Amino Acid Sequence Evolution in *SSP*

Considering that *SSP* has changed in function after its formation by gene duplication, I then asked if there is any acceleration in sequence evolution or positive selection acting on the protein sequence of *SSP*. To answer this question, I implemented sequence rate analysis by using PAML. The *BSK1*-like genes from *Carica*, *Populus*, and *Vitis* were regarded as pro-orthologs based on their chromosomal syntenic collinear relationship with *SSP* and *BSK1* duplication block (Lyons et al. 2008; Proost et al. 2009). The ratio ($\omega = d_N/d_S$) of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitution for the lineage ancestral to *SSP* is significantly higher than the one to *BSK1* (Figure 3.7A). A two ω-ratio model that allows one ω-ratio leading to pro-orthologs and the *BSK1* branch (reflecting functional constraint) and another ω-ratio leading to the *SSP* branch (allowing functional diversification) fits better than the one ω-ratio model that only applies one ω-ratio for the whole phylogeny based on a LRT ($P < 0.0001$; Table 3.2). In addition, the three ω-ratio model that assumes three different ω-ratios leading to the pro-ortholog branch, the *BSK1* branch, and the *SSP* branch, respectively, fits better than two different ω-ratios, suggesting that the *BSK1* branch evolves much slower than the pro-ortholog branch and the *SSP* branch (Table 3.2). However, a free ω-ratio model fits marginally only better than the three ω-ratio model, suggesting that the majority of changes in the protein sequence occurred after divergence of *SSP* and *BSK1* ($P < 0.05$; Table 3.2).

I found that the sequence rate acceleration in *SSP* has continued over the recent evolutionary history of the group. From pairwise d_N/d_S ratio comparisons among two *Arabidopsis* species and from *B. rapa*, *SSP* showed a 30 times significantly higher d_N/d_S ratio than *BSK1* (Figure 3.7B),

suggesting that the sequence rate acceleration was an ongoing process after the speciation event leading to the *Arabidopsis* and *Brassica* lineages and also between *A. thaliana* and *A. lyrata*. The higher d_N/d_S ratio in *SSP* is largely caused by the elevation of nonsynonymous nucleotide substitutions (Table 3.3). I next performed a branch-site model test to examine if there has been any positive selection acting on specific sites in *SSP*, but I did not detect any evidence of positive selection acting on specific sites (Table 3.4). Both the free-ratio model and the branch-site model failed to detect evidence of positive selection, suggesting that the accelerated sequence evolution of *SSP* is due to the relaxation of purifying selection after gene duplication (Kondrashov et al. 2002; Jordan et al. 2004).

3.3.5 Neofunctionalization of *SSP-like1* After Duplication of *SSP*

In addition to the WGD event that created *SSP*, I also observed two other duplications that created two loci closely related to *SSP*: AT2G17170 - here referred to as *SSP-like1*, and AT2G17160 - here referred to as *SSP*-like2. The two genes are close to *SSP* on chromosome 2 and represent tandem duplicates (Figure 3.1). Phylogenetic analysis shows that *SSP*-like1 branches with *SSP* (Figure 3.3) and *SSP*-like2 branches with *SSP*-like 1 (data not shown). The finding that *SSP* and *SSP*-like1 are paralogs is consistent with the results of a large-scale analysis of receptor-like kinase genes in *A. thaliana* and *O. sativa* where AT2G17090 and AT2G17170 branched together (Shiu et al. 2004). Evidence from their chromosome locations and my phylogenetic trees indicate that *SSP-like1*, *SSP-like2*, and *SSP* are derived by two tandem duplication events: *SSP* duplicated in tandem to create the ancestral *SSP-like1*/*SSP-like2* sequence, that gene was transposed a few genes downstream on the chromosome (Figure 3.1), and then there was either a partial duplication to create *SSP-like2* or a complete duplication

followed by a deletion at the 5′ end of *SSP*-like2 (Figure 3.8A).

The *SSP-like1* gene compared with *SSP* is missing two exons and part of a third exon at the 3′ end of the gene including the region corresponding to the TPR protein-binding domain (Figure 3.8A). The TPR domain is essential for the function of *SSP* (Bayer et al. 2009), and thus, its absence is highly suggestive that *SSP-like1* does not have the same function as *SSP*. The function of *SSP-like1* is currently unknown. *SSP-like2* only contains about one-fourth of the protein kinase domain, in addition to lacking the TPR domain (Figure 3.8A). Analysis of microarray data from 63 different organ types and developmental stages of *A. thaliana* (Schmid et al. 2005) indicated that *SSP-like1* shows expression above background only in unopened flowers (stage 9-11), 28-day whole flowers, and sepals (stage 15), whereas *SSP*-like2 shows no expression above background across any of the developmental stages and organ types, and it likely is a pseudogene fragment (Figure 3.8B). I verified some of the microarray results using RT-PCR with six different organ types. In *SSP-like1*, expression was seen in unopened flowers of stage 9 and earlier among the organs types examined (Figure 3.8C). In contrast to *SSP-like1*, no expression of *SSP-like2* was observed (Figure 3.8C), further suggesting that *SSP-like2* is a pseudogene fragment. The expression pattern of *SSP-like1* contrasts greatly to *SSP*, and thus, the expression pattern of *SSP-like1* has considerably changed after gene duplication.

To test if there has been adaptive evolution acting on *SSP-like1*, I performed sequence rate and positive selection analysis using PAML. Compared with *SSP* and *BSK1*, *SSP-like1* has experienced rate acceleration after its formation, especially at nonsynonymous sites (Figures 3.9, 3.10). The rate acceleration is comparable in scale with the sequence evolution of *SSP* (Figure 3.9). Although the free ratio did not show any evidence of positive selection (i.e., $d_N/d_S > 1$) on the branch leading to *SSP-like1* (Figure 3.9), the branch-site model suggests that *SSP*-like1

shows evidence for positive selection at many sites (Figure 3.10; Table 3.5), suggesting that *SSP*-like1 might have undergone adaptive evolution after gene duplication.

3.4 Discussion

3.4.1 Neofunctionalization of *SSP* and *SSP*-like1 by Complete Changes in Expression Pattern, Function, Amino Acid Changes, and Deletions

SSP shows several hallmarks of a gene that has undergone neofunctionalization after duplication, and it is uncommon to find all of these features in a single neofunctionalized gene: 1) The function of *SSP* has changed from being a component of the brassinosteroid signal transduction pathway to regulating elongation of the embryo by an intriguing paternal effect mechanism. The dramatic functional change is surprising, although neofunctionalization of a duplicated gene sometimes produces a paralog with a very different function. 2) Functional divergence was caused in part by deletions in the kinase activation domain that abolished kinase activity and binding of the *SSP* predecessor to its interaction partner *BSU1*. Not only has *SSP* gained a new function but the gene also has lost its original function, unlike some duplicated and neofunctionalized genes that are still partially redundant. 3) An accelerated rate of amino acid changes in *SSP*, relative to *BSK1*, also probably was involved in functional divergence of *SSP*. 4) The organ-specific expression pattern of *SSP* has changed and it is completely opposite from its duplicated partner *BSK1* in pollen compared with 62 other organs and developmental stages. Such a drastic change in expression pattern has been found rarely, if at all, in duplicated genes, although expression data sets of comparable sizes are available only in a very small number of multicellular eukaryotes.

I hypothesize that the expression pattern of *SSP* changed before the functional change, from expression in all organs except pollen to expression only in pollen. That might have occurred by gain of expression in pollen followed by loss of expression in all other organs, or perhaps expression in all organs (except pollen) was lost followed by gain of expression in pollen before the gene could suffer a pseudogenization mutation. If instead the function of *SSP* changed first to its current function before the expression pattern changed, it would hyperactivate the *YODA* pathway in green tissues and result in developmental defects, including lack of stomata, as inferred by data from seedlings expressing *SSP* from a strong, broadly active promoter (Bayer et al. 2009). Alternatively, the deletions in the kinase catalytic domain of *SSP* that abolished the original function could have occurred first, followed by changes in expression pattern and gain of the new function. After becoming expressed in pollen *SSP* was free to evolve rapidly in amino acid sequence.

Why would paternal control of zygote elongation and division evolve from a duplicated gene whose original function immediately upon WGD was involvement in brassinosteroid signal transduction? One possibility is that both the *SSP* and *BSK1* proteins are plasma membrane bound and contain a TPR domain that is important for mediating protein–protein interactions (Tang et al. 2008; Bayer et al. 2009). It is hypothesized that SSP may exert its function in regulating the timing of zygote elongation by recruiting an unidentified pathway activator and thus the importance for protein–protein interactions (Bayer et al. 2009). Alternatively, *SSP* may have evolved into a regulator of zygote elongation by chance co-option of a duplicated gene that had undergone accelerated amino acid sequence evolution and deletions.

In addition to neofunctionalization of *SSP*, I found evidence for neofunctionalization of

SSP-like1 after forming by duplication of *SSP*. *SSP-like1* has a different organ-specific expression pattern from *SSP*, most notably that it does not appear to be expressed in mature pollen where *SSP* is exclusively expressed (except for the *SSP* transcripts provided by the sperm to the zygote). Thus, *SSP-like1* functions in different organ types from *SSP*, and it probably has a different function. *SSP-like1* has a greatly accelerated rate of amino acid substitutions, even more so than *SSP*, and it shows evidence for positive selection at specific sites. However, it is not known if any of those sites are amino acids critical for function. *SSP-like1* has lost the TPR domain at the C-terminus. The TPR domain is essential for the function of *SSP* (Bayer et al. 2009) and its loss in *SSP-like1*, in combination with the accelerated amino sequence evolution and positive selection, further indicate that *SSP* and *SSP-like1* have diverged in function. *SSP-like1* probably has lost its original function and gained a new function as has *SSP*. The sequence of events that created *SSP-like1* and *SSP* by gene duplication and the events involved in neofunctionalization are summarized in Figure 3.11.

I hypothesize that *SSP-like1* was created by duplication of *SSP* after the divergence of the *Arabidopsis* and *Brassica* lineages from a common ancestor, based on phylogenetic evidence. My phylogenetic analysis shows that *SSP*-like1 in *A. thaliana* branches with *SSP* in *A. thaliana* and *A. lyrata* instead of basal to *SSP* in *Brassica* (Figure 3.3). However, the statistical support level for the branch separating *SSP* in *Brassica* from *SSP* and *SSP-like1* in *Arabidopsis* is relatively low, so my inference of the timing of the duplication should be regarded as tentative. Once the *B. rapa* genome is mostly or fully sequenced, it should be possible to determine if there is or is not a homolog of *SSP-like1* in *Brassica*.

The very different organ-specific expression patterns of *SSP*, *BSK1*, and *SSP-like1* suggest that changes have occurred in cis-regulatory elements of *SSP* and *SSP-like1*. I used the cis-regulatory

element prediction program PlantCARE (Lescot et al. 2002) to predict and compare *cis*-regulatory elements among the three genes. Numerous putative *cis*-regulatory elements were detected, with eight being unique to *SSP* and five being unique to *SSP-like1* (Figure 3.12). The unique *cis*-regulatory elements might contribute to their organ-specific expression organs. However, it is difficult to say how many of the predicted *cis*-regulatory elements are actually acting as regulatory elements and which ones are spurious matches to potential *cis*-regulatory elements. Further analysis with experimental constructs would be necessary to determine which regulatory elements have changed to give *SSP* and *SSP-like1* their unique expression patterns. In addition, I found a 769-bp helitron 1,660-bp upstream of the start codon of *SSP* in *A. thaliana*, but the helitron was not present in *SSP* in *B. rapa*. Considering that both *Brassica* and *Arabidopsis* show expression of *SSP* in pollen, the helitron does not appear to be involved in the pollen-specific expression of *SSP* in *A. thaliana*.

3.4.2 Neofunctionalization after Gene Duplication in Plants

SSP and *SSP-like1* add to the small number of cases of gain of a new function after gene duplication and loss of the old function during the evolution of a plant family. Studies of neofunctionalization of genes duplicated by WGD, as well as other types of gene duplication, have revealed several types of neofunctionalization involving regulation and/or sequence and structural changes. Changes in protein function can occur by mutations in the amino acid sequence or by structural changes in the sequence including deletions and insertions, especially in functional domains. Some genes show either amino acid changes or structural changes, whereas other genes like *SSP* and *SSP-like1* show both. Duplicate genes that evolve new functions can either lose their old function, like *SSP*, or retain the old function with

neofunctionalization having the effect of diversifying the gene's function. Gain of a new function by a duplicated gene can be accompanied by changes in expression patterns, as with *SSP* and *SSP-like1*, or instead expression patterns can remain largely the same. Likewise, regulatory neofunctionalization (new expression patterns) can occur with or without changes in the function of the protein coded by the gene. Regulatory neofunctionalization has been proposed to act in either a qualitative manner, with gain of a completely new expression pattern after duplication, or a quantitative manner, with changes in the expression level of one copy after gene duplication (Force et al. 1999; Duarte et al. 2006).

Polyploidy events provide a large number of new genes that could potentially undergo neofunctionalization. Neofunctionalization might occur relatively soon (within about 1 Myr) after polyploidy in plants that are still cytologically polyploids, or it may be a process that mostly happens several million years later, during or after cytological diploidization. The only currently known cases of neofunctionalization in an evolutionarily recent plant polyploid, to my knowledge, are 15 genes in *G. hirsutum* (tetraploid cotton) that show regulatory neofunctionalization (Chaudhary et al. 2009). In contrast, there are numerous potential cases of regulatory neofunctionalization (neofunctionalization of expression patterns) after the α-WGD event in the Brassicaceae, based on a combination of expression divergence and asymmetric sequence evolution between the duplicates (Blanc and Wolfe 2004b; Ganko et al. 2007) or expression divergence and ancestral state inference (Duarte et al. 2006; Zou et al. 2009). However, changes in function have not been studied or shown for most of those cases of regulatory neofunctionalization. Neofunctionalization of expression patterns can be detected much more readily than the evolution of new functions because the latter requires experimentally determined functional information.

An alternative fate of duplicated genes is escape from adaptive conflict (EAC) and sometimes it may be mistaken for neofunctionalization (discussed in Des Marais and Rausher 2008). In the EAC model, a single (preduplication) gene undergoes selection to perform a new function in addition to its original function. However, the gene is constrained from improving either function because of detrimental effects on the other function. After duplication, one copy improves one function and the other copy improves the other function. In the case of *SSP* and *BSK1*, the ancestral function was the current function of *BSK1* and not the current function of *SSP*; that is inconsistent with EAC. In addition, the EAC model predicts that both duplicates will undergo adaptive change instead of showing purifying selection. *BSK1* is undergoing purifying selection within the protein sequence, whereas *SSP* exhibits relaxation of selection but it does not show evidence for positive selection. I conclude that *SSP* and *BSK1* have not undergone EAC, and instead, *SSP* has experienced neofunctionalization.

3.4.3 Recent Evolutionary Origin of Paternal Control of Embryonic Patterning

In addition to being a dramatic example of neofunctionalization, my study of *SSP* also provides insights into the timing of the evolution of the gene's intriguing and novel paternal effect mechanism for control of zygote elongation after fertilization. *SSP* regulates the *YODA* pathway that activates elongation and asymmetric division of the zygote after fertilization to create the embryo precursor and the elongated suspensor cell, using a paternal control mechanism of translation of *SSP* transcripts that were provided by the sperm cells in the pollen instead of there being maternal *SSP* expression (Bayer et al. 2009). My study shows that *SSP* originated at the base of the Brassicaceae family, and thus the *SSP*-mediated paternal control of embryonic

patterning is restricted to the Brassicaceae. Thus, other angiosperms must use a different mechanism to regulate elongation of the zygote after fertilization. One possibility would be transcripts from another gene provided by the pollen, using a similar paternal effect mechanism as does *SSP*. Another possibility would be expression of a gene in the zygote only after fertilization that regulates the *YODA* pathway. The mechanism involving *SSP* has replaced the ancestral mechanism, as the zygotes in *SSP* mutants do not undergo normal elongation (Bayer et al. 2009).

Yet a different mechanism for regulating the timing of zygote elongation may occur in apomictic plants in the genus *Boechera*, which undergo embryogenesis without fertilization by pollen, and lie within the lineage encompassed by the α-WGD (Bailey et al. 2006). Apomictic *Boechera* plants would lack paternally supplied *SSP* transcripts, and thus *Boechera* probably has an alternative genetic basis for controlling zygote elongation. One possibility would be expression of the maternal allele of *SSP* in the zygote at the proper time for elongation.

In addition to *SSP*, uniparental expression of genes involved in embryo and endosperm development also includes imprinted genes where only one allele is expressed and the other allele is epigenetically silenced in a parent-of-origin specific manner. Two of the paternally imprinted genes, *MEDEA* and *FWA*, arose from the α-WGD at the base of the Brassicaceae from a nonimprinted ancestral gene, and each gene has a nonimprinted paralog (Nakamura et al. 2006; Spillane et al. 2007). Interestingly, *SSP*, *MEDEA*, and *FWA* all originated from the same WGD by neofunctionalization and gain of uniparental expression, albeit with uniparental expression being accomplished with different mechanisms. Thus, the creation of imprinting and other parent-of-origin expression effects during seed development are ongoing evolutionary processes in plants.

Table 3.1. Gene-specific primers.

Species	Gene	Direction	Primer (5'-->3')	GenBank No.
Arabidopsis thaliana	*SSP-like1*	Forward	TGTTGCTTCTCGACGCCTCTTGAT	AT2G17170, Proost et al. 2009
		Reverse	TATGATACGCCACTCGCAACCTCA	
	SSP-like2	Forward	ATGGGTACAAAAGCTAGGAGACCAAA	AT2G17160, Proost et al. 2009
		Reverse	ATGCCTCCAGATTCTGCTTTGTCAT	
	UBQ10	Forward	TCACCGGAAAGACAATCACC	BP860797
		Reverse	ACGTACGGCCATCCTCTAG	
Brassica rapa subsp. *pekinensis*	*BSK1.1*	Forward	GTTGTTGCCAATCCTTGTGTTCGG	FP340665
		Reverse	AGCTGCTTTGAGGTCGGAGAAAGA	
	BSK1.2	Forward	CTTCAACGCCACAGAAGCCACTTT	EX076408
		Reverse	CCAAACACAGTTGGCGACACCATT	
	SSP	Forward	TGTTGCCATTCATTGTCTTCCGGG	AC232545
		Reverse	ATCAAGCCACGATTCTGCAAACGG	
	Actin1	Forward	TCGAGACTTTCAATGTCCCTGCCA	EX135734
		Reverse	ACGGAATCTTTCAGCTCCGATGGT	
Carica papya	*BSK1*	Forward	ACCCACCCAGAAAGAGACCAAACT	CP00020G01800, Proost et al. 2009
		Reverse	TAGGCATGTACTCAGCAACGAGCA	
	BSK11	Forward	TTCGACTGAAGAGGCAACTGTGGT	CP00026G02150, Proost et al. 2009; GU321199
		Reverse	CCCTACGTCTATGAACTGAGAATAACAC	
	Actin1	Forward	AGACACACAGGTGTCATGGTTGGA	EL784289
		Reverse	GGCAGTTTCAAGCTCCTGCTCAAA	
Gossypium hirsutum	*BSK1.1*	Forward	AACAACAGGAGCCACAGAACCGTA	DT555127, DW480267, DW480268
		Reverse	GCATTGCCCACTCGATGGTTTGAT	
	BSK1.2	Forward	ATCTTTCCAAGAGTGGACCCAGCA	EX168588, EY196709
		Reverse	GCATGTCTAGTTTGGCAAGGGCAA	
	BSK1.3	Forward	ATCAAACCATCGAGTGGGCAATGC	GU321198
		Reverse	ATCTGCTGGGTCCACTCTTGGAAA	
	Actin1	Forward	ACTGGTGTTATGGTTGGGATGGGT	ES812773
		Reverse	AGCTTGGATGGCAACATACATGGC	
Vitis vinifera	*BSK1*	Forward	ACAATACCCAGAACCCACCCTTCA	VV03G03010;, Proost et al. 2009
		Reverse	CTCTCAAACGCATGGCCCATTCAA	
	Actin1	Forward	TGCCTGCCATGTATGTTGCCATTC	CF513819, FC070998
		Reverse	CCACCACTAAGCACAATGTTGCCA	
Helianthus annuus	*BSK1*	Forward	GGGTCTTTGTGCATCAGCTCCAAA	DY923044, DY92629, GE495420
		Reverse	ACCCGATCAAATTCGCTAGTCGCT	
	Actin1	Forward	AAGGCTGGATTCGCTGGAGATGAT	DY910078
		Reverse	ACGATTTCCCGTTCTGCTGACGTA	

Table 3.2. ω (d_N/d_S)-ratio values and LRT statistics under different branch models.

Model	ω-ratio			NP	*l*
	Pro-orthologs	***BSK1***	***SSP***		
	Vitis*, *Populus*, *Carica	***A. lyrata*, *A. thaliana*, *Brassica***	***A. lyrata*, *A. thaliana*, *Brassica***		
One-ratio[a]	0.11			19	-6649.75
Two-ratio[b]	0.04		0.42	20	-6519.05
Three-ratio[c]	0.05	0.02	0.43	21	-6510.08
Free-ratio[d]	See Figure 3.7A			35	-6497.97

Likelihood ratio tests of models			
Comparisons	**2δ*l***	**d.f.**	***P*-value**
One-ratio v.s. two-ratio	261.40	1	<0.0001
Two-ratio v.s. three-ratio	17.94	1	<0.0001
Three-ratio v.s. free-ratio	24.22	14	0.0431

[a]The one-ratio model assumes only one ω-ratio for the whole phylogeny.
[b]The two-ratio model assumes one ω-ratio for pro-orthologs and *BSK1*, and another ω-ratio for *SSP*.
[c]The three-ratio model assumes three different ω-ratio for pro-orthologs, *BSK1*, and *SSP*, respectively.
[d]The free-ratio allows each branch of the phylogeny to have a different ω-ratio.
Abbreviations: NP, number of parameters; *l*, likelihood estimate of model; d.f., degree of freedom.

Table 3.3. Pairwise comparison of d_N value, d_S value, and d_N/d_S ratio in *SSP* and *BSK1*.

Variable	Taxon	***SSP***			***BSK1***		
		A. lyrata	***A. thaliana***	***Brassica***	***A. lyrata***	***A. thaliana***	***Brassica***
d_N	*A. lyrata*	-			-		
	A. thaliana	0.045	-		0.001	-	
	Brassica	0.193	0.190	-	0.007	0.006	-
d_S	*A. lyrata*	-			-		
	A. thaliana	0.140	-		0.104	-	
	Brassica	0.438	0.506	-	0.475	0.475	-
d_N/d_S	*A. lyrata*	-			-		
	A. thaliana	0.319	-		0.010	-	
	Brassica	0.441	0.375	-	0.015	0.013	-

Table 3.4. LRT statistics of branch-site models for *SSP* and *BSK1* branches.

Gene	Model A_{test2}[a] v.s. Model A_{test1}[b]						
	Purifying selection (%, $d_N/d_S < 1$)	**Neutral evolution (%, $d_N/d_S = 1$)**	**Positive selection (%, $d_N/d_S > 1$)**	**$2\delta l$**	**d.f.**	***Q*-value**	**Selected site[c]**
SSP (*A. lyrata*, *A. thaliana*, *Brassica*)	37	63	0	0.00	50:50 mixture of 0 and1	0.5000	-
BSK1 (*A. lyrata*, *A. thaliana*, *Brassica*)	98	2	0	0.00	50:50 mixture of 0 and1	0.5000	-

[a]The model A_{test2} is a null hypothesis that assumes no positive selection (ω-ratio = 1) on the foreground branch.
[b]The model A_{test1} is an alternative hypothesis that assumes positive selection (ω-ratio > 1) on the foreground branch.
Abbreviations: *l*, likelihood estimate of model; d.f., degree of freedom.

Table 3.5. LRT statistics of branch-site models for *SSP-like1* and *SSP* branches.

Gene	Model A_{test2}[a] v.s. Model A_{test1}[b]						
	Purifying selection (%, $d_N/d_S < 1$)	**Neutral evolution (%, $d_N/d_S = 1$)**	**Positive selection (%, $d_N/d_S > 1$)**	**$2\delta l$**	**d.f.**	***Q*-value**	**Selected site[c]**
SSP-like1 (At2g17170)	59	2	39	4.60	50:50 mixture of 0 and1	0.0480	See Figure 3.4.
SSP (*A. lyrata*, *A. thaliana*)	60	40	0	0.00	50:50 mixture of 0 and1	0.5000	-
SSP *(Brassica rapa)*	69	31	0	2.54	50:50 mixture of 0 and1	0.0833	-

[a]The model A_{test2} is a null hypothesis that assumes no positive selection (ω-ratio = 1) on the foreground branch.
[b]The model A_{test1} is an alternative hypothesis that assumes positive selection (ω-ratio > 1) on the foreground branch.
[c]Only those with posterior probability greater than 0.95 are shown.
Abbreviations: *l*, likelihood estimate of model; d.f., degree of freedom.

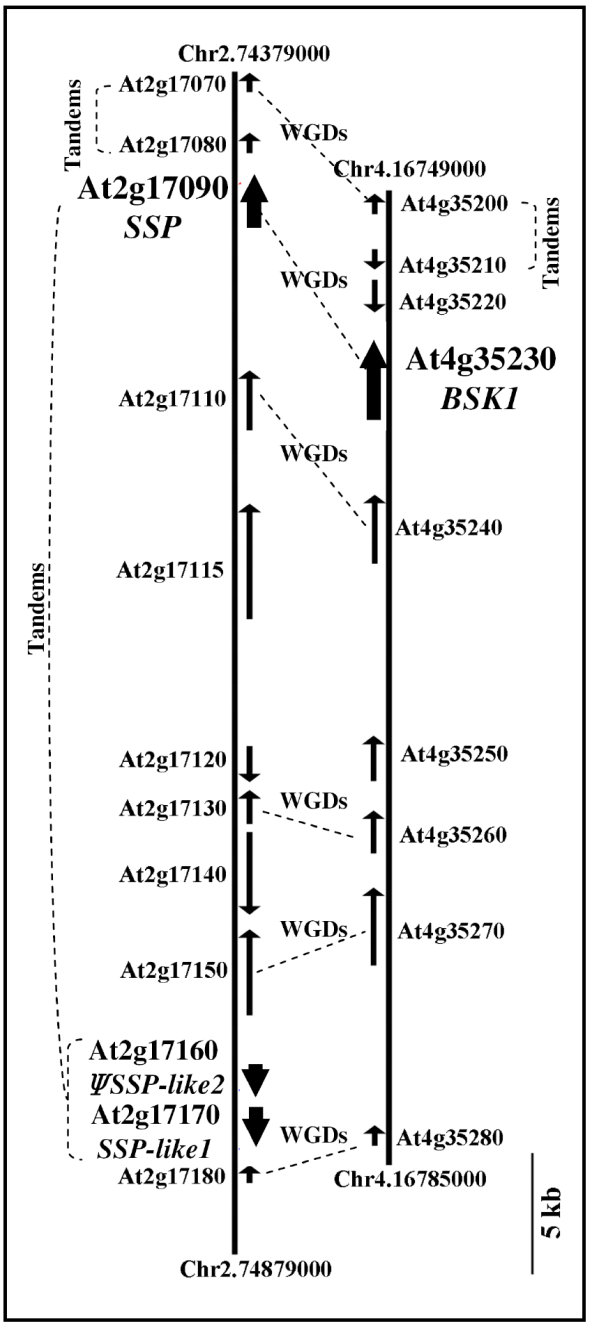


Figure 3.1. Duplicated blocks on chromosomes 2 and 4 containing *BSK1*, *SSP*, *SSP-like1*, and *SSP-like2*. Scale bar indicates the nucleotide length. Abbreviations: Chr2, chromosome 2; Chr4, chromosom4; WGDs, genes duplicated by the most recent WGD; Tandems, genes formed by tandem duplication; W, putative pseudogene. Information about SSP-like1 and SSP-like2 is presented in the last section of the Results.

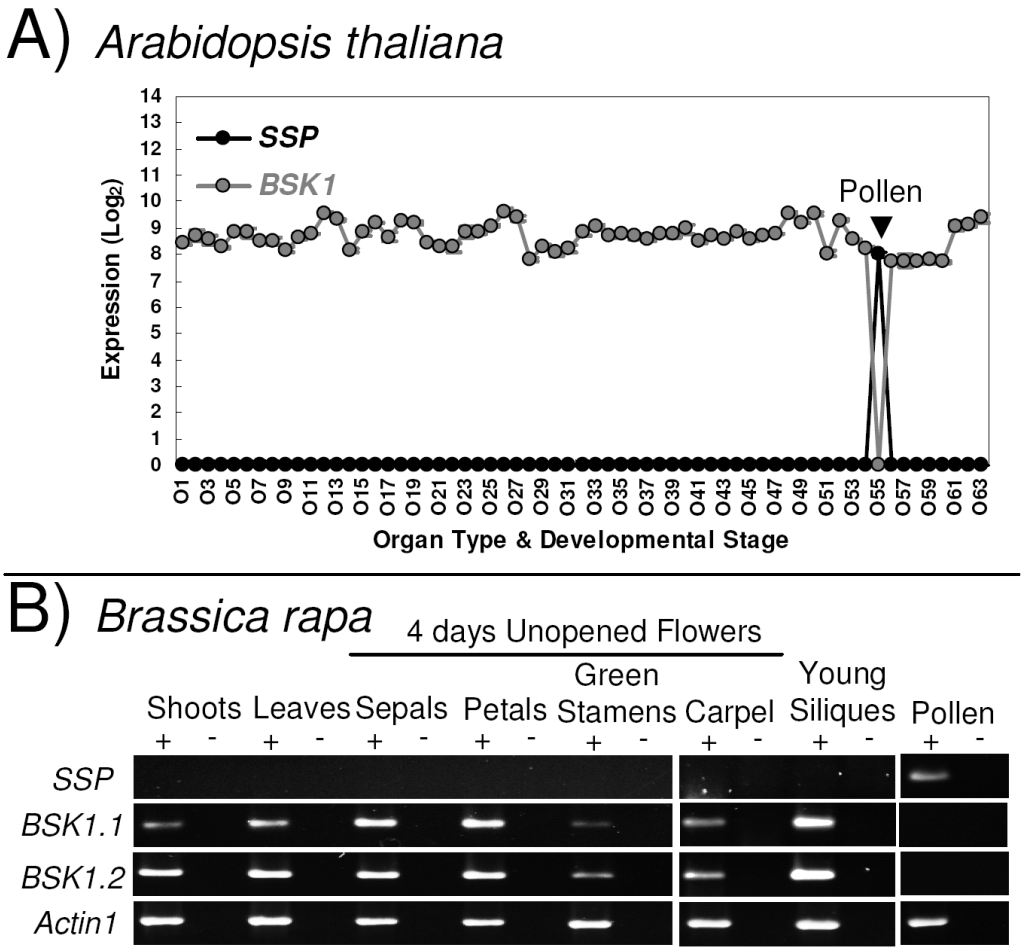


Figure 3.2. *SSP* and *BSK1* show opposite organ-specific expression patterns. (A) Gene expression of *SSP* and *BSK1* in *Arabidopsis thaliana*. The MAS5-normalized microarray data were obtained from 63 different developmental stages and organ types. Absence and presence of expression above background was determined using mas5calls. Error bars show variance among three biological replicates. The different developmental stages and organ types are listed in Table 2.1. (B) RT-PCR expression assays of *SSP*, *BSK1.1*, and *BSK1.2* in *Brassica rapa*. Plus signs indicate reactions containing reverse transcriptase and minus signs indicate reactions without reverse transcriptase.

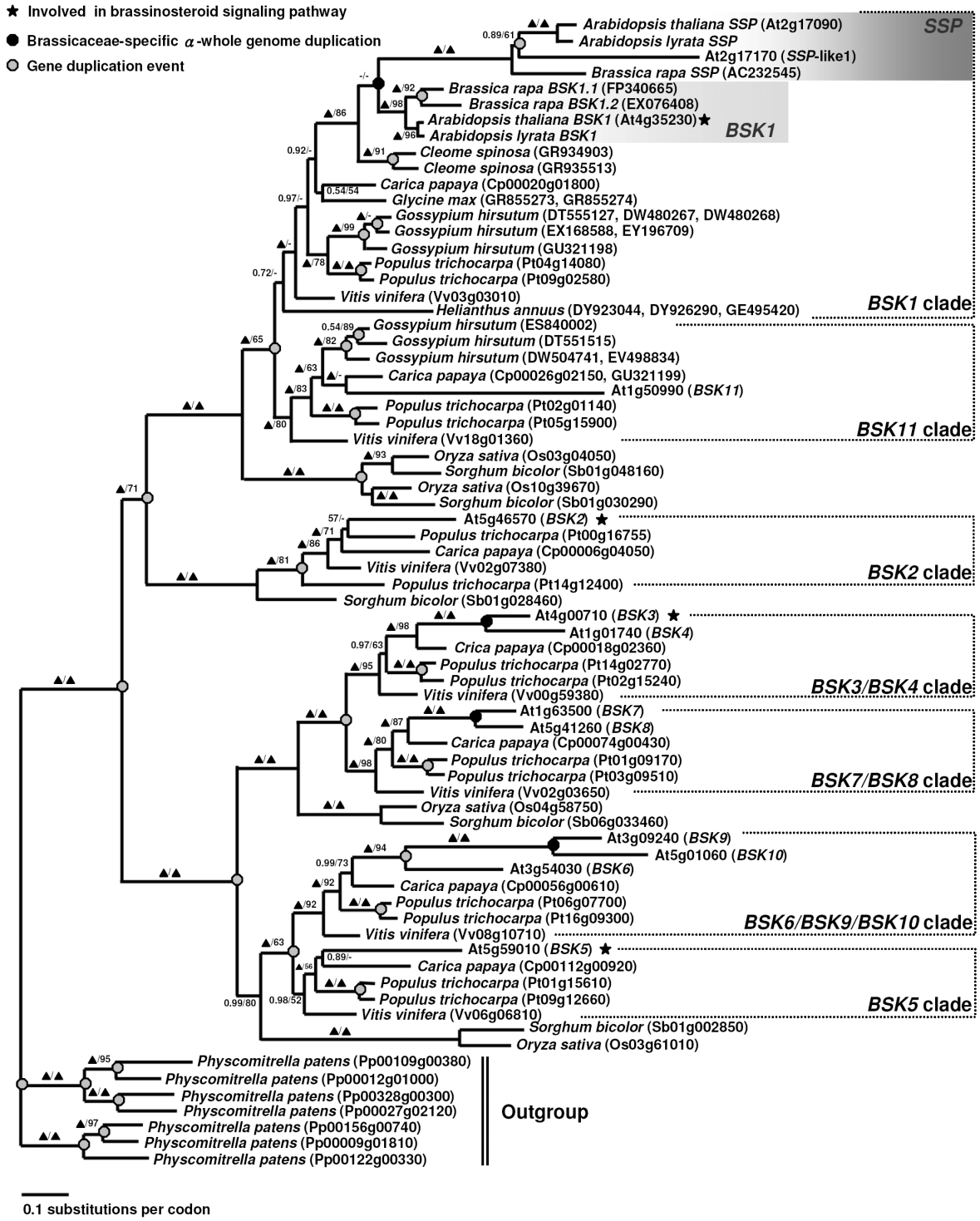


Figure 3.3. Phylogenetic tree of the *BSK* gene family. Topology is inferred from a Bayesian method. Statistical support is indicated on branches. The first value is posterior probability from Bayesian analysis (▲=1) and the second value is 100 bootstrapping replicates using ML analysis (▲=100). Dash indicates either posterior probability < 0.5 or bootstrap value < 50. Stars indicate genes that have been shown to function in brassinosteroid signal transduction (Tang et al. 2008). Dark gray shading indicates *SSP* genes and *SSP-like1* gene, whereas light gray shading indicates *BSK1* genes.

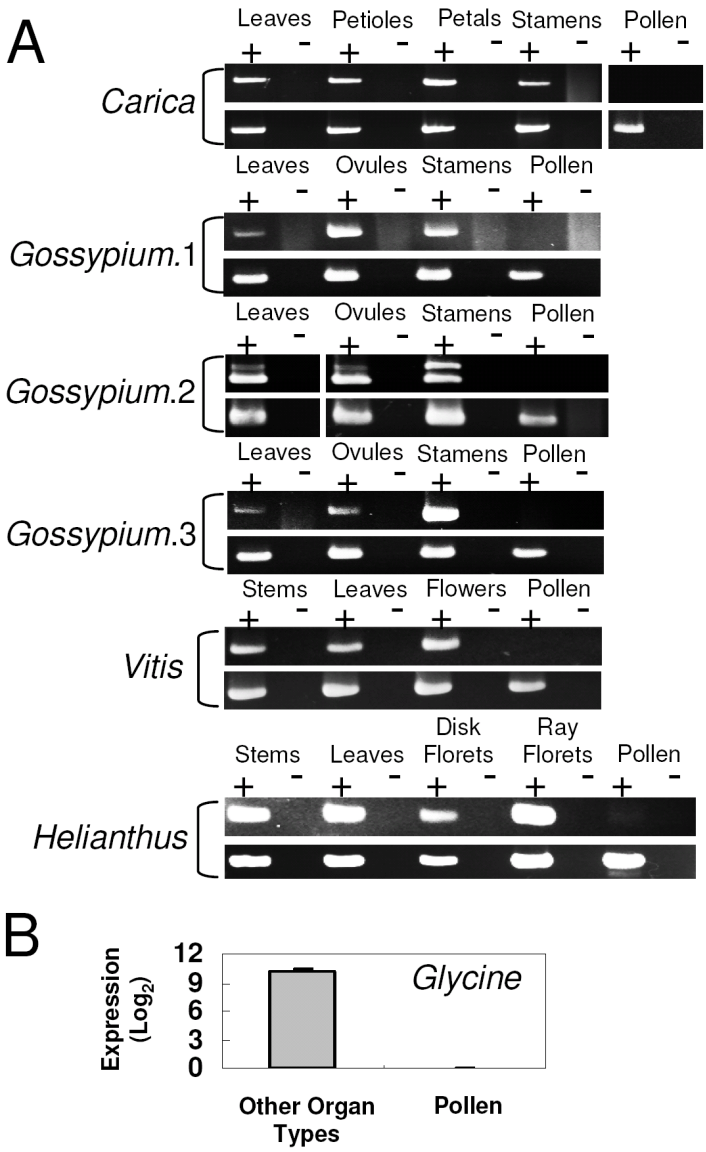


Figure 3.4. Gene expression analyses of *BSK1* orthologs from outgroup species. (A) RT-PCR expression assays of *BSK1* orthologs in the upper panels of each pair, and actin genes in the lower panel of each pair were used as a control for DNA template concentration. See Figure 3.3 and Table 3.1 for gene accession numbers. Plus signs indicate reactions containing reverse transcriptase (RT) and minus signs indicate reactions without RT. Two bands are present in *Gossypium hirsutum BSK1.2* because of an intron retention alternatively spliced variant. (B) Microarray expression analysis of the *BSK1* ortholog in *Glycine max*. Error bars show variance among different biological replicates (two biological replicates in other organ types and three biological replicates in pollen).

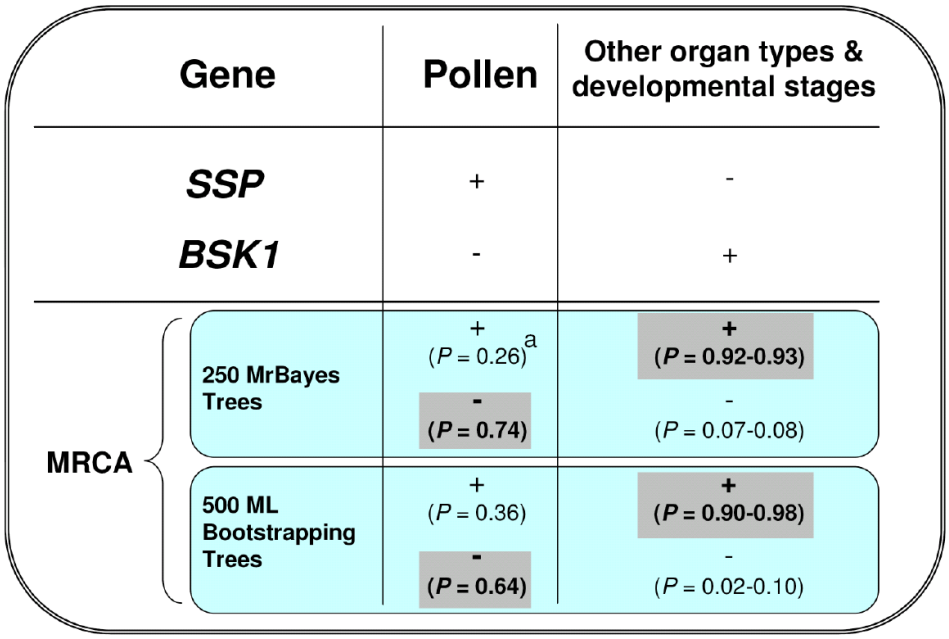


Figure 3.5. Reconstruction of the most recent common ancestral (MRCA) expression state between *SSP* and *BSK1* by using a maximum likelihood method with gene family phylogenies. Plus sign (+) indicates presence of expression; minus sign (-) indicates absence of expression. [a]*P* indicates the probability of absence or presence of expression. Grey shading indicates that the probability is greater than 60% following the 60% consensus rule. From my analysis, the MRCA expression state of *SSP* and *BSK1* is no expression in pollen ($P = 0.64$-0.74) but expression in all of the other organ types and developmental stages ($P = 0.90$-0.98). These results suggest that *SSP* became expressed in pollen after gene duplication, and they are consistent with the results of the outgroup RT-PCR analysis.

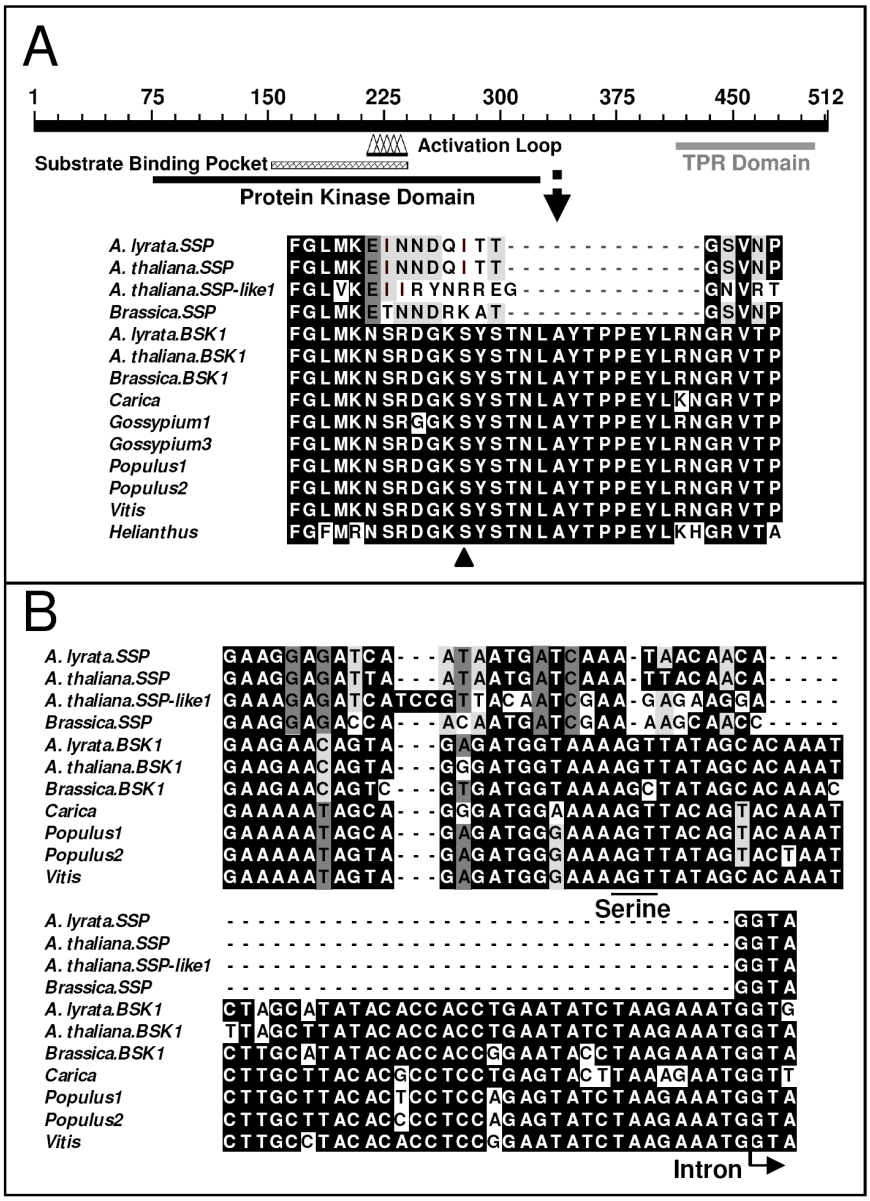


Figure 3.6. Sequence analysis of functional domains in *SSP*, *SSP-like1*, and *BSK1*. Conserved functional domains were identified by using the Conserved Domain Database (CDD) with an interactive domain family analysis (Marchler-Bauer et al. 2007). (A) Diagram from the CDD showing two functional domains: the protein kinase domain and the TPR domain. The effects are illustrated of nucleotide substitutions and deletions in exon 4 on the activation loop of the protein kinase domain in *SSP* and *SSP-like1*. The location of serine-230, that is essential for *BSK1* function, is marked with a triangle. (B) Nucleotide alignment of exon 4 showing the locations of the nucleotide substitutions and deletions. The positions corresponding to serine-230 are indicated with a line and ‘‘serine.’’ The start of intron 4 is indicated with an arrow. Shading: black, .45% shared identity; dark gray, 35–45% shared identity; light gray, 25–35% shared identity; and white, <25% shared identity.

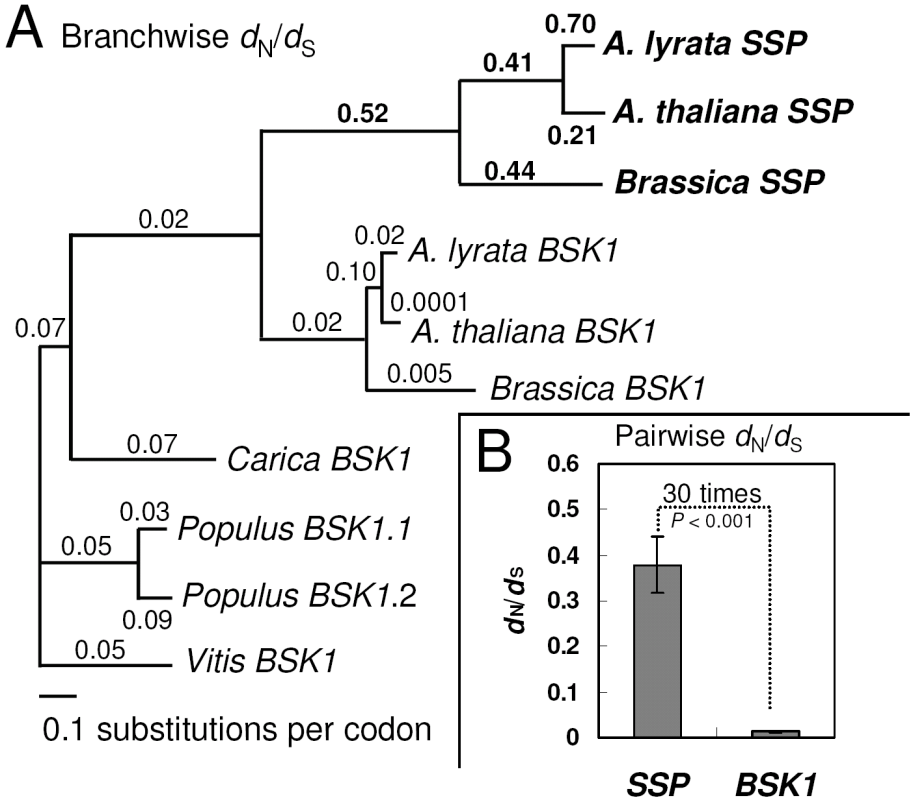


Figure 3.7. Sequence rate and selection analysis of *SSP* and *BSK1*. (A) Phylogenetic tree of *SSP* genes from *Arabidopsis thaliana*, *A. lyrata*, and *Brassica rapa*, and *BSK1* genes from *A. lyrata*, *A. thaliana*, and *B. rapa*, *Carica papaya*, *Populus trichocarpa*, and *Vitis vinifera*. Tree branch length and branchwise d_N/d_S ratios above the branches were estimated with codeml using a free-ratio model. A higher evolutionary rate is found in *SSP*. The tree is unrooted. (B) Pairwise d_N/d_S ratios, d_N, and d_S analysis among *A. lyrata*, *A. thaliana*, and *B. rapa* showing that *SSP* has a d_N/d_S ratio 30 times higher than BSK1.

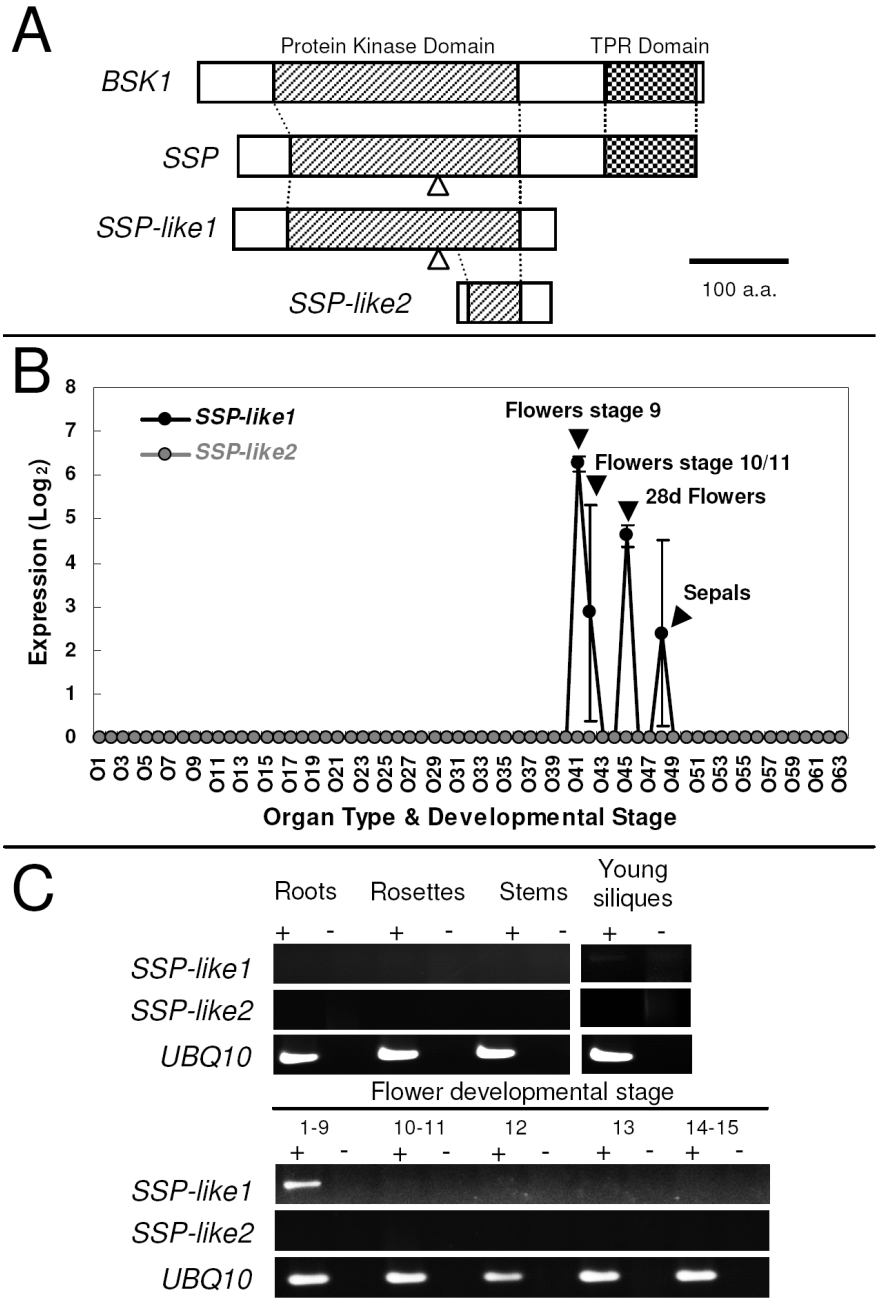


Figure 3.8. Gene structure and expression of *SSP-like1* and *SSP-like2*. (A) Gene structure of *SSP-like1* and *SSP-like2* in comparison with *BSK1* and *SSP*. Arrowheads indicate the deletion in the protein kinase domain. (B) The MAS5-normalized microarray data from 63 different developmental stages and organ types. Expression values were background corrected. Error bars indicate variance among replicates. The developmental stages and organ types in the microarray data are listed in Table 2.1. (C) Expression assay by RT-PCR verifying that *SSP-like1* is expressed in the early stage of unopened flower but not in other organs examined, and *SSP-like2* is not expressed across different organ types. *UBQ10* was used as a control for cDNA template concentration.

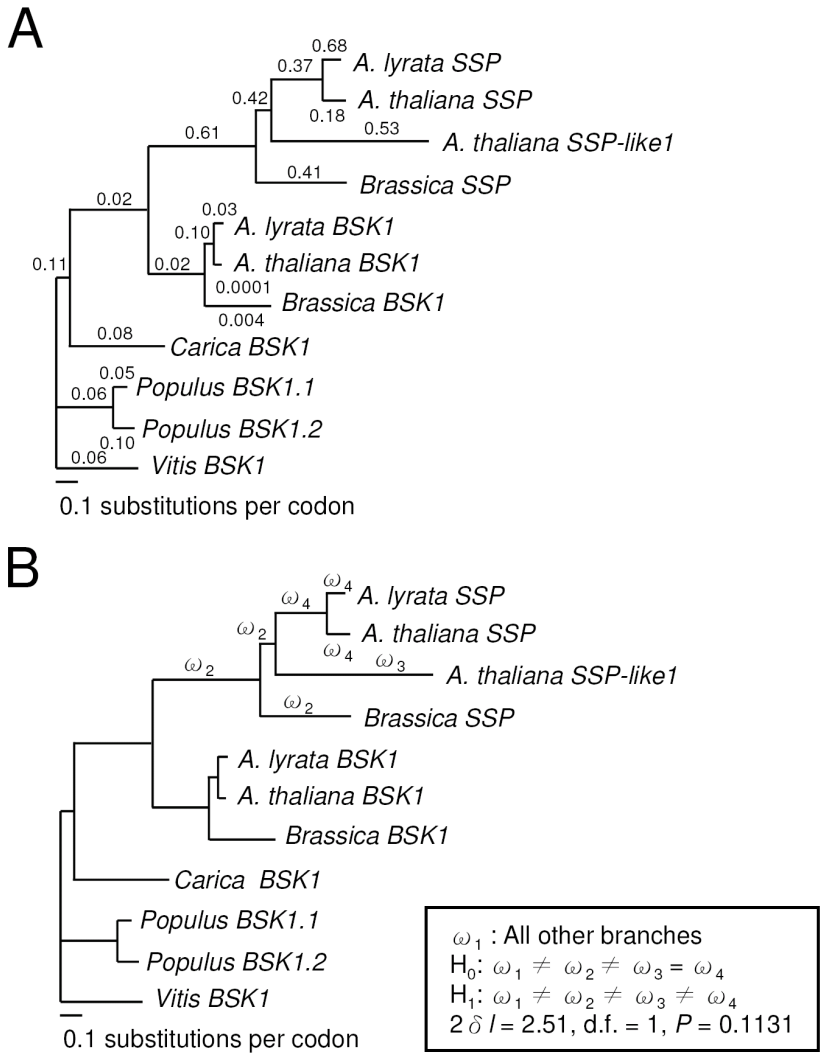


Figure 3.9. Sequence rate analysis of *SSP-like1*. (A) Phylogenetic tree of *SSP-like1* from *Arabidopsis thaliana*, *SSP* from *A. thaliana, A. lyrata*, *Brassica rapa* subsp. *pekinensis,* and *BSK1* genes from *A. lyrata*, *A. thaliana*, *Brassica rapa* subsp. *pekinensis*, and *BSK1* orthologs from *Carica papaya*, *Populus trichocarpa*, and *Vitis vinifera*. Tree branch lengths and branchwise d_N/d_S ratios above the branches were estimated with Codeml using a free-ratio model. The tree is unrooted. *SSP-like1* shows a higher evolutionary rate of sequence changes than the *BSK* genes. (B) Likelihood ratio test showing that *SSP-like1* did not undergo a much higher rate of sequence evolution than its duplicated partner, *SSP*. Abbreviations: *2δl*, twice difference of likelihood values from two models; d.f., degree of freedom; H0, null hypothesis; H1: alternative hypothesis.

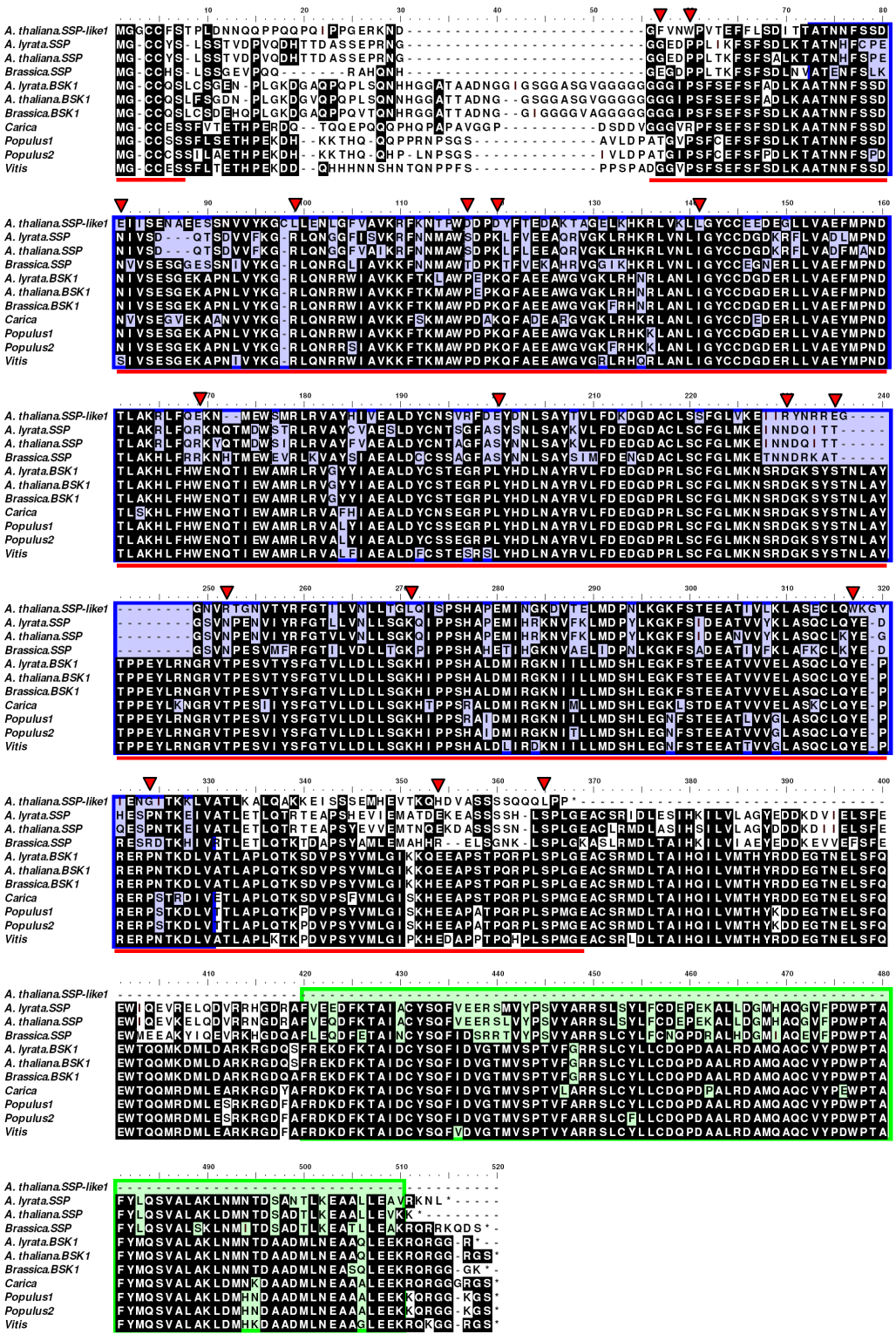


Figure 3.10. Sequence alignment for positive selection analysis. The region used for analysis is underlined in red. Arrowheads indicate the positively selected sites (posterior probability > 0.95) in *SSP-like1* (AT2G17170) determined by a branch-site model in PAML. The blue box indicates the protein kinase domain and the green box indicates the TPR domain.

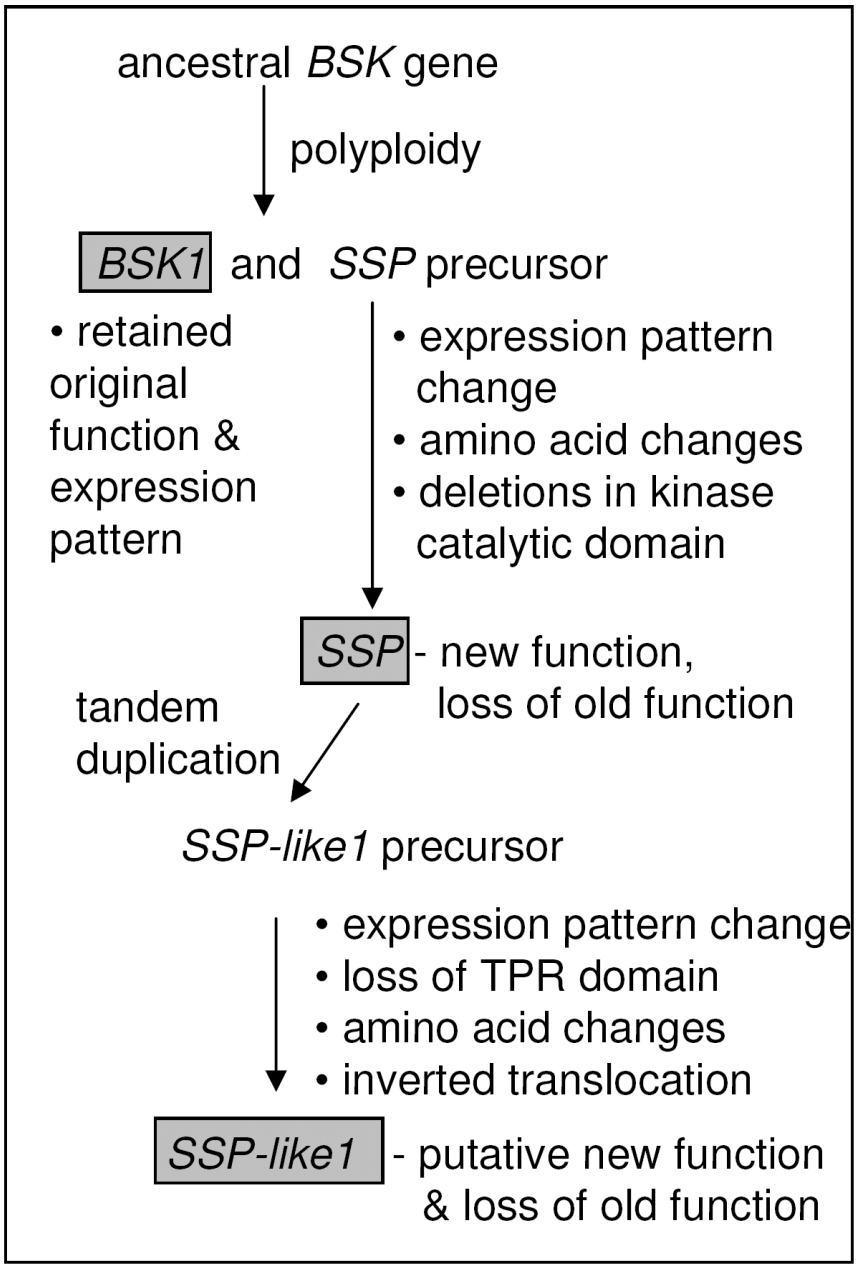


Figure 3.11. Model for duplication and neofunctionalization of *SSP* and *SSP-like1*. See text for explanation of the steps.

A

Cis -element	*SSP* 723 nt upstream	*SSP-like1* 723 nt upstream	*BSK1* 723 nt upstream
5UTR Py-rich stretch		5UTR Py-rich stretch	5UTR Py-rich stretch
AAGAA-motif	AAGAA-motif	AAGAA-motif	AAGAA-motif
ACE	ACE		
AE-box			AE-box
ARE			ARE
as-2-box		as-2-box	
AT1-motif			AT1-motif
ATCT-motif		ATCT-motif	ATCT-motif
Box 4	Box 4	Box 4	
Box I	Box I	Box I	Box I
Box-W1		Box-W1	
CAAT-box	CAAT-box	CAAT-box	CAAT-box
CGTCA motif	CGTCA-motif		
circadian	circadian	circadian	
CTAG-motif		CTAG-motif	
EIRE			EIRE
ERE		ERE	ERE
GARE-motif		GARE-motif	GARE-motif
G-box	G-box		
G-Box	G-Box		
GT1 motif	GT1-motif	GT1-motif	
HSE	HSE	HSE	
MRE	MRE		
P-box			P-box
RY-element		RY-element	
Skn-1 motif	Skn-1_motif		
TATA-box	TATA-box	TATA-box	TATA-box
TC-rich repeats	TC-rich repeats	TC-rich repeats	TC-rich repeats
TCT-motif	TCT-motif		TCT-motif
TGACG-motif	TGACG-motif		
Unnamed__2	Unnamed__2		
Unnamed__4	Unnamed__4		Unnamed__4
W-box		W box	

B

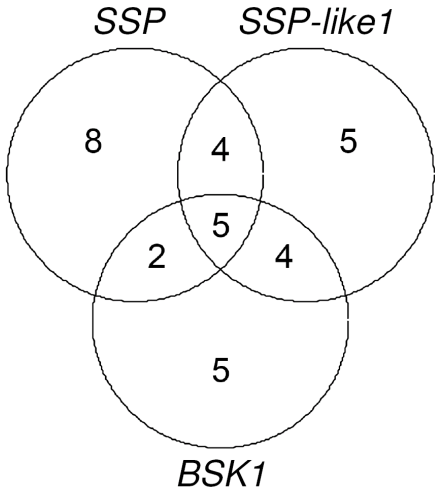


Figure 3.12. Analysis of the regions upstream of *SSP, SSP-like1,* and *BSK* coding regions for potential *cis*-regulatory elements. To understand if there have been changes in the sequences of *cis*-regulatory elements contributing to the acquisition of organ-specific expression, I surveyed the number of unique potential *cis*-regulatory elements in *SSP*, *SSP-like1*, and *BSK1* from *A. thaliana* (A) by using the plant *cis*regulatory element database PlantCARE. *SSP* and *SSP-like1* contain four unique *cis*-elements in comparison to *BSK1* (B). *SSP* has eight unique predicted *cis*-regulatory elements, and *SSP-like1* has five unique predicted *cis*-regulatory elements. It is unclear if any these unique predicted *cis*-regulatory elements play a role in the organ-specific expression patterns of *SSP* and *SSP-like1*.

4 Molecular Adaptation and Expression Evolution Following Duplication of Genes for Organellar Ribosomal Protein S13 in Rosids[1]

4.1 Introduction

Gene duplication has been an ongoing process during eukaryotic evolution that has provided genetic raw material for the evolution of new gene functions that can lead to morphological and physiological novelty. Duplicated genes can undergo sequence divergence caused by positive selection or neutral drift and divergence in expression patterns and function (Wagner 2002; Zhang 2003; Li et al. 2005). Two common fates of retained duplicated genes are neofunctionalization – gain of a new function or expression pattern by one copy (Ohno 1970) and subfunctionalization – partitioning of ancestral function or expression pattern between both copies (Force et al. 1999; Lynch and Force 2000b). Plant genomes contain large numbers of duplicated genes, derived by polyploidy, segmental duplications, tandem duplications, and retroposition of cDNAs. Many duplicated genes in plant genomes have been preserved and undergone purifying selection, a few have undergone positive selection and functional diversification, and some have experienced subfunctionalization (Moore and Purugganan 2003; Lawton-Rauh 2003; Adams and Wendel 2005; Moore and Purugganan 2005).

---

1 A version of chapter 4 has been published. **Liu, S.-L.**, Adams, K.L. (2008) Molecular adaptation and expression evolution following duplication of genes for organellar ribosomal protein S13 in rosids. BMC Evolutionary Biology 8: 25 (16 pages).

There are three genes that code for organellar ribosomal S13 genes among rosid species. Analysis of *rps13* genes in the rosid species *Arabidopsis thaliana*, cotton (*Gossypium arboreum*), and soybean (*Glycine max*) revealed the presence of two expressed copies of *rps13* in the nucleus that were derived by gene duplication (Adams et al. 2002a; Mollier et al. 2002). Both in vitro and in vivo RPS13 protein import experiments indicated that one copy encodes the chloroplast-imported RPS13 protein (nucp *rps13*) while the other encodes mitochondrion-imported RPS13 protein (numit *rps13*) (Adams et al. 2002a; Mollier et al. 2002). It was inferred that the missing mt *rps13* gene product has been functionally replaced by the product of numit *rps13* in a common ancestor of *Arabidopsis*, cotton, and legumes. Thus the function of numit *rps13* has been modified after gene duplication, and one could argue that numit *rps13* has gained a new function because it is operating in a new cellular context (the mitochondrial ribosome instead of the chloroplast ribosome). Subsequently mt *rps13* was lost from mitochondrial DNA many times during the evolutionary history of rosids, as inferred from a Southern blot hybridization survey (see modified Figure 4.1 based on Adams et al. 2002b). Surprisingly, however, there were many species of rosids that do appear to retain *rps13* in the mitochondrion, based on Southern blot hybridizations, but the gene in some species may not be intact or functional.

The organellar *rps13* genes in rosids provide an intriguing system to study gene duplication because the subcellular location and site of action of numit RPS13 has changed after gene duplication from the chloroplast to the mitochondrion. I have studied sequence evolution of numit *rps13* among rosids to determine what kinds of amino acid changes have taken place and where those amino acids are located in the tertiary structure, as well as to test the hypothesis that there has been adaptive evolution. Also I have examined the divergence of expression patterns between numit *rps13* and nucp *rps13*. After finding intact and expressed numit *rps13* and mt

rps13 genes in *Malus* I tested the hypothesis that there has been expression partitioning of the two genes in different organ types and/or stress conditions to preserve both genes.

4.2 Materials and Methods

4.2.1 Database Searches

The following expressed sequence tags (EST) for numit *rps13* were obtained from GenBank by BLAST searches using numit *rps13* from *Arabidopsis thaliana* (DR380621) as a query: *Citrus sinensis* (CX672493), *Malus domestica* (DR995890, CN494589, CN925904, and CX023021), *Populus trichocarpa* (DT496554), *Medicago truncatula* (BI272420), *Glycine max* (BM188020), and *Gossypium hirsutum* (DW488419). The following ESTs for nucp *rps13* were obtained from GenBank by BLAST searches using nucp *rps13* from *Arabidopsis thaliana* (DR367899) as a query: *Citrus sinensis* (CX053776), *Glycine max* (EH261685), *Gossypium hirsutum* (DW226103), *Malus domestica* (DT000985), *Medicago truncatula* (BI265445) and *Populus trichocarpa* (DT486949). Mt *rps13* ESTs from *Malus domestica* (CN871057 and CN875489) and *Prunus persica* (AJ873687) were identified by BLAST searches of GenBank using mt *rps13* from *Beta vulgaris* as a query. Mt *rps13* genes used in this study include: *Triticum aestivum* (Y00520), *Zea mays* (AF079549), *Magnolia* spp. (Z49799), *Helianthus* annuus (AJ243789), *Daucus carota* (X54417), *Oenothera berteriana* (X54416), *Beta vulgaris* (DQ381464) and *Marchantia polymorpha* (M68929).

4.2.2 Sequence Alignment/Analysis and Phylogenetic Analysis

Sequence alignment was done using transAlign, which aligns protein-coding DNA sequence based on the alignment of amino acids (Bininda-Emonds 2005). Aligned sequences were refined with BioEdit (Hall 1999) for further phylogenetic and codon substitution analysis (available upon request). For phylogenetic analysis, maximum likelihood (ML) analysis was conducted with MultiPhyl v1.0.6 (Keane et al. 2007) with SPR (subtree pruning and regrafting) branch swapping. The optimal evolutionary model selected for ML was K81uf+I+G (Kimura three parameter with unequal base frequencies + proportion of sites + gamma distribution) using the following parameters: assumed nucleotide frequencies A = 0.36736, C = 0.20950, G = 0.27698, T = 0.14616; expected transition/transversion ratio = 0.94; expected pyrimidine transition/purine transition ratio = 0.30; proportion sites assumed to be invariable = 0.29; rates for variable sites assumed to follow the gamma distribution with shape parameter = 4.28. Bootstrapping was performed using MultiPhyl v1.0.6 with neighbor-joining algorithm and 100 replicates (Keane et al. 2007).

4.2.3 d_N/d_S Analyses and Likelihood Ratio Tests (LRT) for Positive Selection

Nonsynonymous (d_N) and synonymous (d_S) nucleotide substitution rates were calculated by using program yn00 in PAML v3.15 (Yang 1997). The *t*-test was applied to test if there is any significant difference of pairwise d_N, d_S, and d_N/d_S ratios between the nucp *rps13* and numit *rps13* cDNA sequences. The statistically significant level was set at 95%. The statistical analysis in this study was implemented by using the statistical package R (http://www.r-project.org/). For detection of positive selection, codon-based analysis was implemented using codeml in PAML

v3.15 (Yang 1997). The mature coding region, without the mitochondrial targeting presequences, was included in my analysis. Site-specific models were used for testing positive selection on numit *rps13* and nucp *rps13* (Yang 2007). Two LRTs were used for the detection of positive selection: M7-M8 and M8a-M8 (Yang and Nielson 2002; Swanson et al. 2003; Yang 2007). Because comparison of M7 and M8 provides a more powerful test of positive selection and has less sensitivity to large evolutionary distances and G+C content than comparison of M1 and M2 (Yang et al. 2000; Friedman and Hughes 2006), only comparison of M7 and M8 was used for the detection of positive selection in this study. M7 and M8a models are the null models without positive selection (no codon with $d_N/d_S > 1$) and the M8 model is the alternative model with positive selection. Branch site-specific model was used to test if there has been positive selection for *Malus* branch in numit *rps13*, nucp *rps13*, and mt *rps13* (Yang 2007). For this analysis, the branch which I am interested in testing positive selection was assigned as the foreground lineage. Therefore, the branch that leads to *Malus* in numit *rps13*, nucp *rps13*, and mt *rps13* were set as the foreground lineage, and branches that lead to the rest of other rosid species were designated as background lineages. For the branch site-specific model, two LRTs were used for detection of positive selection as follows: M1-Model A test1 (MA_{test1}) and Model A test2 (MA_{test2})-MA_{test1} (Yang and Nielson 2002; Yang 2007). M1 and MAtest2 models are the null hypothesis without positive selection (no codon with $d_N/d_S > 1$) and the MAtest1 model is the alternative selection with positive selection. For site-specific and branch site-specific models used for detection of positive selection, M8a is more stringent than M7, and MA_{test2} is more stringent than M1 because their ω2 (d_N/d_S) is fixed at 1. For all LRTs, the first model is simpler than the second one, with fewer parameters and a poorer fit to the data. Therefore, the first model has a lower maximum likelihood index. To test if there is statistically better maximum likelihood for the second model, twice the difference in log maximum likelihood values between the two compared models [$2\delta L$ = -2(*Ln1-Ln2*), where *Ln1* and *Ln2* represent for log of maximum likelihood value in the first

model and the second model] was compared against a chi-square (x^2) distribution. The degrees of freedom (d.f.) equal the additional parameters used in the more complex model. The d.f. is 2 for M7-M8 and M1-MA_{test1}, while the d.f. is 1 for M8a-M8 and MA_{test2}-MA_{test1}. However, it has been argued that the appropriate model comparison for M8a-M8 would be to use 50:50 mixture of d.f. = 0 and d.f. = 1 (Swanson et al. 2003). In this study, the calculation of *P* values for M8a-M8 was followed as described in Kapralov and Filatov (2007). The *P* value was first obtained using d.f. = 1 and then divided by 2. Bayes empirical bayes (BEB) approach was used to determine positively selected sites in the M8 model and MA_{test1} model. In BEB analysis, the posterior probabilities were calculated to select the codon with d_N/d_S greater than 1. Because BEB analysis is more powerful at detecting positive selection than naive empirical Bayes (NEB) analysis (Yang et al. 2005), only BEB analysis was considered in this study.

4.2.4 Structural Analysis of RPS13

Using the search function "first approach mode" in SWISS-MODEL (Schwede et al. 2003), nucp, numit, and mt RPS13 amino acid sequences in *Malus* were selected to search for a suitable template for further structural analysis. Based on the results, RPS13 of *Escherichia coli* and *Thermus thermophilus* were chosen for the following structural analysis. RPS13 structural data file for *E. coli* (PDB ID: 2gy9M) (Mitra et al. 2006) and *T. thermophilus* (PDB ID: 1fjgM) (Carter et al. 2000) were obtained from RCSB Protein Data Bank (http://www.rcsb.org/pdb). Location and property of each amino acid in *E. coli* and *T. thermophilus* RPS13 was analyzed using DeepView-Swiss-PdbViewer v.3.7 (Schwede et al. 2003). Selected positive sites obtained by BEB analysis and sites shared with similarity between numit RPS13 and mt RPS13 were plotted in the relative position of RPS13 in *E. coli* and *T. thermophilus* according to amino acid

alignment.

4.2.5 Microarray Data Analysis

Raw Affymetrix ATH1 microarray data (Schmid et al. 2005) were downloaded from the TAIR website (TAIR accession number: ME00319) (http://www.arabidopsis.org/). Raw data were processed and normalized based on the GC-RMA method (Wu et al. 2004). The expression values were converted into log2 numbers by only considering values of perfect-match probes (Wu et al. 2004). Normalization and analysis of the microarray data were implemented using Bioconductor. Pearson correlation analysis was conducted to statistically compare the similarity of expression profile among *rps9*, *rps10*, *rps11*, and numit *rps13*, and between numit *rps13* and nucp *rps13*. The correlation coefficient (R) values correspond to the similarity of the expression profile between two genes. A one-way ANOVA was used to test if there is a significant expression difference between two different genes or any two different organ types. All statistical analyses were implemented using statistical package R (http://www.r-project.org/). The statistically significant level was set at 95%.

4.2.6 Plant Materials and Abiotic Stress Treatments

Several organ types, including stems, leaves, peduncles, sepals, petals, stamens, stigmas+styles, and young fruit of apple (*Malus domestica* Borkh.) were collected from the University of British Columbia's Botanical Garden. Seedlings were used for abiotic stress experiments. Prior to sowing, apple seeds were soaked in distilled water at 4°C for 6 weeks for vernalization. Seedlings were cultivated in a peat-vermiculite soil mixture under fixed day/night (16 h day/8 h

night) and temperature of 20–23°C. After germination and emergence from the soil for 7 days, the plants were subjected to five different abiotic stresses: 4°C for 7 days (cold treatment), 37–40°C for 12 hours (heat stress), 100 mM NaCl solution for 7 days (salt treatment), submerged in distilled water for 4 days (water submersion treatment), and dark for 7 days (dark treatment). After stress treatments, roots, hypocotyls, cotyledons, and leaves were collected and immediately frozen in liquid nitrogen and stored at -80°C until nucleic acid extraction.

4.2.7 Nucleic Acid Extraction, Gene Amplification, and Sequencing

Genomic DNA extraction was done using the DNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. Total RNA extraction was performed as described previously (Adams et al. 2003). The extracted nucleic acid concentrations and purities were determined by using a NanoDrop spectrophotometer. The quality was checked by running on 1.5% agarose gels. Before reverse transcription, 3 µg of RNA (500 µg/µl) was treated with 1 unit of DNaseI (New England Biolabs) and incubated at 37°C for 30 min twice. Then 4 µg of DNase-treated RNA was reverse-transcribed by using M-MLV reverse transcriptase (Invitrogen). The reverse transcription conditions were 25°C for 10 min, 37°C for 60 min, and 70°C for 15 min. Finally, the reverse-transcribed samples were treated with RNase (Invitrogen) at 37°C for 20 min.

The *rps13* genes were amplified from genomic DNA and cDNA by polymerase chain reaction (PCR) using gene-specific primers (Table 4.1). The PCR was performed in a reaction mixture (10 µl) consisting of 4.88 µl of ddH2O, 1 µl of genomic DNA/cDNA solution, 1 µl of PCR buffer, 1 µl of 2.5 mM MgCl2 solution, 1 µl of 0.2 mM dNTPs, 0.5 µl of 0.4 µM each primer, and 0.12 units of Taq DNA Polymerase (Sigma). The PCR conditions were 96°C for 4 min, and 30 cycles

of 96°C for 40 s, 60°C for 40 s, 72°C for 1 min, and 72°C for 10 min. The PCR products were run on a 1.5% agarose gel and extracted from the gel using QIAquick Gel Extraction Kit (Qiagen). PCR products amplified from genomic DNA and cDNA were sequenced directly. The sequencing was performed in a reaction mixture containing 0.4 µl of ABI BigDye Version 3.1 (Applied Biosystems), 3.6 µl of BigDye buffer, 5.5 µl of 50 ng template, and 0.5 µl of 0.4 µM forward or reverse primer. The sequencing reaction was carried out with the following program: 1 min at 96°C, and 25 cycle of 10 s at 96°C, 5 s at 50°C, and 4 min at 60°C. The sequencing products were run on an ABI 377 DNA Sequencer (Applied Biosystems) at the UBC Centre for Plant Research, or on an ABI 3730 Sequencer at the Nucleic Acids Protein Service unit at UBC. The GenBank accession number for mt *rps13* in *Malus*, including information about RNA editing sites, is EU084692.

4.2.8 Real-time qRT-PCR

Quantitative real-time RT-PCR was performed with a BioRad iQ5 system using SYBR green master mix (BioRad) following the manufacturer's instructions, except that 25µl total reaction volumes were used. The PCR conditions were 96°C for 3 min, and then 35 cycles of 96°C for 10 s, 58°C for 30 s, and 72°C for 30 s. Gene-specific primers are listed in Table 4.1. For each sample, two technical replicates were performed. Reactions for a standard curve were run with each set of experimental reactions. After the completion of PCR, the melting curves were analyzed to distinguish the true product from artifacts such as primer dimers. The iQ5 software and Microsoft Excel were used for data analysis. Normalization was done using the actin gene *ACT2*. GenBank accession numbers for sequences used to design primers are: *Malus ACT2*: CN903171, CN902302, and N917499; *Malus cob*: CN872477, CN856986, CN856316; *Malus*

rps10: CV882925.

4.3 Results

4.3.1 Identification of Numit *rps13* in *Malus*, *Populus*, and *Citrus*, and Phylogenetic Analysis

To study the evolution of numit *rps13* sequences in the rosids, I identified sequences homologous to numit *rps13* in *Malus*, *Populus*, and *Citrus* by BLAST searches of the NCBI expressed sequence tag (EST) database using the numit *rps13* from *Arabidopsis* as a query. ESTs (see Materials and Methods) were aligned and the open reading frames were identified. The genes in each species are predicted with high probability to encode mitochondrial proteins by four prediction programs: MitoProt (Claros and Vincens 1996), TargetP v1.1 (Emanuelsson et al. 2000), Predotar v1.03 (Small et al. 2004), and iPSORT (Bannai et al. 2002). Three of the four programs discriminate mitochondrial from chloroplast proteins and no support was obtained for chloroplast targeting. Alignment of the N-termini with mitochondrial targeting presequences from numit *rps13* in other rosid species shows sequence conservation and indicates that they are homologous genes (Figure 4.2). Because numit *rps13* in *Arabidopsis*, *Gossypium*, and *Glycine* have been experimentally shown to be imported into mitochondria but not chloroplasts (Adams et al. 2002a), I infer that the products from the homologous genes in *Malus*, *Populus*, and *Citrus* also are targeted to mitochondria. I found numit *rps13* sequences only in species belonging to the rosid lineage, suggesting that the gene duplication event that created numit *rps13* likely occurred after the emergence of the rosid lineage. I did not find a numit *rps13* sequence in *Vitis*, a group at the base of rosids (Soltis et al. 2005; Jansen et al. 2006), despite the mostly sequenced genome

(Jaillon et al. 2007) and a large EST collection available at NCBI. I infer that the gene duplication event likely occurred at the base of the eurosids after separation from the Vitaceae lineage (Figure 4.1).

Nucp *rps13* sequences from seven rosid species were obtained from GenBank by BLAST searches using the previously characterized nucp *rps13* from *Arabidopsis* (Kumar et al. 1995) and aligned with the numit *rps13* sequences. Phylogenetic analysis of the sequences verified the orthologous relationships of numit *rps13* and nucp *rps13* sequences (Figure 4.3). The phylogenetic relationships inferred from nucp *rps13* and numit *rps13*, however, did not follow established relationships of rosid species (Soltis et al. 2005). For numit *rps13*, the positions of *Populus* and *Gossypium* in the tree are switched; for nucp *rps13* the gene positions in the tree are completely scrambled. The lack of congruence between the gene trees and organismal phylogeny is probably due to the short sequences being analyzed (about 242 bp after excluding the third codon position). Alternatively, it might be caused by the complex history of multiple polyploidy events and subsequent loss of duplicated genes during rosid evolution. For example, after polyploidization the lineage containing legumes and *Malus* might have retained copy 1 while *Arabidopsis*, *Citrus*, *Gossypium* and *Populus* retained copy 2 (Figure 4.3). Comparisons of branch lengths showed that numit *rps13* sequences have diverged rapidly with much longer branch lengths than the nucp *rps13* sequences, suggesting there has been accelerated evolution of numit *rps13* in each lineage, and consistent with results from Adams et al. (2002a) that included fewer species of rosids.

4.3.2 Accelerated Nucleotide Substitution Rates and Positive Selection on Numit *rps13*

To better quantify the rate increase and to determine if both d_N and d_S (or just d_N) are accelerated in numit *rps13* sequences, the d_N, d_S values, and d_N/d_S ratios were compared between nucp *rps13* and numit *rps13* in the rosids. Significantly higher d_N values and d_N/d_S ratios were observed in the numit *rps13* genes than in the nucp *rps13* genes ($P < 0.05$; Table 4.2). The d_N values in the numit *rps13* genes were 3–4 times greater than in the nucp *rps13* genes. No significant difference in d_S values, however, was found between the nucp *rps13* and numit *rps13* copies ($P > 0.05$; data not shown). These results show that numit *rps13* has been experiencing a considerably accelerated rate of non-synonymous substitutions compared with the nucp *rps13* in all seven lineages of rosids.

Although simple pair-wise comparison of d_N and d_S analysis provides some insights into the accelerated amino acid substitution in numit *rps13*, branchwise estimation of d_N, d_S values, and d_N/d_S ratios can provide additional information such as positive selection and adaptive molecular evolution along certain branches and clades (Yang 1997; Yang and Nielson 2002; Yang and Swanson 2002; Swanson et al. 2003; Yang 2007). To detect if there has been positive selection acting on numit *rps13* and *nucp rps13* in different rosid species, site specific model analysis was conducted using PAML (Yang 1997). Both the *rps13* tree (Figure 4.3) and the species tree that reflects our current understanding of rosid phylogeny (Soltis et al. 2005) were used for detection of positive selection to determine if the tree topology influenced detection of positively selected sites, and no differences were found in this regard.

One ratio model (M0) shows that the d_N/d_S ratio of numit *rps13* is about 4 times higher than nucp *rps13* (Table 4.3), congruent with observations from simple d_N and d_S analysis. For each dataset, likelihood ratio tests (LRT) for detection of positive selection were examined using M7-M8 and M8a-M8 comparisons. For numit *rps13*, M8 is significantly better than M7 and M8a in numit *rps13* ($P < 0.05$) and 2.3% of sites are inferred to undergo positive selection (Table 4.3), suggesting there has been positive selection acting on numit *rps13*. In addition, based on Bayes Empirical Bayes (BEB) analysis, two codon sites in numit *rps13* were identified as showing strong positive selection (posterior probability > 0.8; Table 4.3). However, there is no evidence for sites under positive selection in nucp *rps13*. In nucp *rps13*, M8 is not significantly better than M7 and M8a (Table 4.3). These results indicate that strong positive selection acts on the evolution of numit *rps13* among the rosid species, particularly at codons 28 and 114.

4.3.3 Structural Location of Positively Selected Amino Acids

Relative locations of the two positively selected sites were plotted on the tertiary amino acid structures of RPS13 from *Escherichia coli* and *Thermus thermophilus* (Figure 4.4) to infer approximate locations in numit RPS13. Except for the C-terminal end which is longer and contains one more α-helix structure in *T. thermophilus*, the overall structure of RPS13 from *E. coli* and *T. thermophilus* are relatively similar (Figure 4.4). RPS13 residues at positions 28 and 114 in *E. coli* (28 and 116 in *T. thermophilus*) correspond to the positively selected sites in numit RPS13 based on the amino acid alignment (Figures 4.4, 4.5). Previous structural and functional studies of *Thermus* RPS13 showed that there are two structurally important regions where RPS13 interacts with the 16S rRNA (Carter et al. 2000; Yusupov et al. 2001; Broderson et al. 2002). One of them is the loop region between helix 1 and turn 1 (residues 22–25) (Broderson et

al. 2002). One of the positively selected residues found in my analysis is close to that region. The second positively selected residue in numit RPS13 is located in the highly basic COOH-terminal extended region (Figures 4.4). This region is virtually devoid of secondary structure and found to interact with the 16S rRNA at the P-site and A-site (residues 116–120 for P-site and residues 120–122 for A-site) (Carter et al. 2000; Yusupov et al. 2001; Broderson et al. 2002). Positively selected sites in or near regions that interact with the 16S rRNA suggest that there has been functional refinement of numit RPS13 to interact better with the 16S rRNA.

4.3.4 Amino Acid Changes in Numit *rps13* that Increase Identity to Mt *rps13*

Because numit *rps13* was derived from nucp *rps13* and encodes a RPS13 protein that functions in the mitochondria, I was interested in determining if the numit RPS13 has become more like the mt RPS13 amino acid sequence. I determined if there have been mutations in the numit *rps13* genes in any of the seven rosid species that change an amino acid to the residue that is present in mt RPS13. In my alignment of nucp RPS13, numit RPS13, and mt RPS13 across different plant species, sixteen sites were identified where numit RPS13 in one or more species has the same amino acid at the corresponding site in mt RPS13 from seven angiosperms, and the amino acid is different from the amino acid(s) present in nucp RPS13 (Figure 4.5A). The sixteen sites are relatively evenly distributed in the RPS13 tertiary structure (Figures 4.4). Although little is known about the exact functions of those regions, some of the mutations might help improve the function of numit RPS13 in the mitochondrial ribosome. I plotted the sixteen amino acid changes on the phylogeny of the seven rosid species to infer when they might have occurred (Figure 4.5B). Mutations are inferred to have occurred along most of the branches, suggesting continuous refinement of the numit rps13 sequence. I infer that three mutations occurred in the

common ancestor of all the species, with a subsequent mutation at site 80 in Arabidopsis from D to E (a conservative substitution) and at site 110 in Gossypium from A to S, although scenarios with multiple recent mutations cannot be ruled out.

4.3.5 Expression Evolution of Numit *rps13*

After I studied sequence evolution of numit *rps13*, I tested the hypothesis that there have been changes in expression patterns and levels of expression of numit *rps13* relative to nucp *rps13*. Extensive microarray data are available for *Arabidopsis thaliana* including a single study that examined expression in 51 organs and developmental stages using the ATH1 array (Schmid et al. 2005). I analyzed expression data for numit *rps13* and nucp *rps13* using an ANOVA approach (see Methods). Nucp *rps13* is expressed at a higher level than numit *rps13* in most organs ($P < 0.05$; Figure 4.6). Notable exceptions were roots, senescing leaves and pollen where numit *rps13* is more highly expressed ($P < 0.05$; Figure 4.6A). When comparing expression levels of the two genes among organs, sometimes the levels of both genes go up or down together, but sometimes the levels go in opposite directions. Overall the organ-specific expression patterns between the two genes show both similarities and differences, depending on the organs compared. The expression correlation between these two genes are positively correlated, but at a moderate degree (Pearson correlation coefficient, $R = 0.31$, $P < 0.05$).

To compare expression patterns of numit *rps13* to other nuclear encoded ribosomal protein genes, I analyzed expression data for three other nuclear encoded mitochondrial ribosomal protein genes: *rps9*, *rps10*, and *rps11* (Bonen and Calixte 2006). The expression levels and patterns of all four genes were highly similar (Pearson correlation coefficient, $R = 0.73–0.94$, $P < 0.0001$),

although *rps11* expression levels were lower in some organs ($P < 0.05$; Figure 4.6B). Thus numit *rps13* has evolved an expression pattern similar to that of other nuclear-encoded genes for mitochondrial ribosomal proteins, and its expression has diverged from that of nucp *rps13*.

4.3.6 *Malus* Contains an Expressed and RNA-edited Copy of *rps13* in the Mitochondrion

Having studied the sequence and expression evolution of numit *rps13* I next consider the fate of mt *rps13* in rosids. A large number of rosids, including *Arabidopsis*, *Gossypium*, *Glycine*, and *Citrus* have lost mt *rps13*, as judged by DNA gel blot hybridization (Adams et al. 2002b) (Figure 4.1), but some rosids retain *rps13* in mitochondrial DNA. I identified a transcribed copy of *rps13* in *Malus domestica* (apple) from BLAST searches of the NCBI EST database that is 90–92% identical with the *rps13* gene in the mitochondrion of several eudicots. The sequence was derived from a study of ESTs in *Malus* (Newcomb et al. 2006). A similar gene (97% identical) was found in *Prunus persica*, another member of the Rosaceae family. I evaluated the *rps13* sequence from *Malus* for sites of C-to-U RNA editing by PCR amplifying and sequencing *rps13* from genomic DNA and comparing the gDNA sequence to the EST sequences. C-to-U RNA editing plays an important role in the expression of plant mitochondrial genes to restore certain amino acids to those that are evolutionarily conserved (Shikanai 2006). Among angiosperm *rps13* genes, nine possible edited sites have been identified (Figure 4.7). Four of those sites are already T's instead of C's in the genomic DNA sequence of mt *rps13* from *Malus*, and thus RNA editing might be expected at five sites in the *Malus* cDNAs. No editing was observed in the two ESTs from *Malus*, and the EST from *Prunus* had editing at only one site (the 100th base). To verify the lack of RNA editing, I amplified and directly sequenced mt *rps13* cDNAs from leaves

and petals of *Malus*. Unexpectedly five RNA editing sites were discovered in both leaves and petals at five sites (Figure 4.7). The RNA editing efficiency has been shown to vary in different organ types, developmental stages, and environmental conditions (Kurihara-Yonemoto and Handa 2001; Peeters and Hanson 2002). The discrepancy between ESTs and my analysis might be due to the effects of different organ types, developmental stages, or environmental conditions on the RNA editing efficiency of mt *rps13*. After transcription, RNA editing converts codons from serine to leucine (the 26th, 56th, and 287th bases) and from arginine to cysteine (the 100th and 256th bases). The changes made by RNA editing would restore an evolutionarily conserved amino acid sequence and make the resulting protein likely to be functional, should the transcripts be translated.

4.3.7 Co-expression of Numit *rps13* and Mt *rps13* in 14 Different Organ Types and under Five Different Stress Conditions

I conducted expression assays of numit *rps13* and mt *rps13*, and RNA editing examination of mt *rps13*, in *Malus* to test the hypothesis that expression patterns have been partitioned between the two genes in different organ types. RT-PCR was performed with (RT+) or without (RT-) reverse transcriptase to check for DNA contamination (Figure 4.8). Transcripts of numit *rps13* and mt *rps13* were observed in roots (from seedlings), stems, stigmas + styles, and ovaries. Transcripts of these two genes were also observed in 10 additional organ types: hypocotyls, cotyledons, young leaves (from seedlings), mature leaves, peduncles, petals, seeds, sepals, stamens, and young fruit. Mt *rps13* cDNAs were sequenced from each organ type to determine if any sites were edited. Different organ types in apple all showed the five RNA editing sites mentioned above. The results of the RT-PCR experiments show that both copies are co-expressed in all

examined organ types.

Although both copies of *rps13* are expressed in many organ types it is possible that there might be partitioning of expression between the two genes under stress conditions. I tested the hypothesis by examining the expression patterns of numit *rps13* and mt *rps13* under five different stresses including cold, dark, heat, salt, and water submersion treatments to determine if expression partitioning occurs under different environmental stresses. Apple seedlings were independently subjected to each of these five stresses (see Materials and Methods). Four organ types, including roots, hypocotyls, cotyledons, and leaves were examined for expression of numit *rps13* and mt *rps13*. Based on RT-PCR results, transcripts of numit *rps13* and mt *rps13* were observed under each of the five stress conditions: cold, dark, heat, salt, and water submersion treatments. Sequencing of RT-PCR products showed that the transcripts were RNA edited at 5 sites, although in some cases there was partial editing at one or more sites.

I have shown transcription of both numit *rps13* and mt *rps13* in many organs and under several stress conditions. However it is possible that transcripts from one gene might be present at low levels. PCR of cDNA template (RT-PCR) is not a quantitative technique and low levels of transcripts might not be distinguishable from high levels. To determine if abundant levels of steady-state transcripts are derived from both genes, I used real-time PCR (qRT-PCR) to assay transcript levels of numit *rps13* and mt *rps13*. I compared levels of numit *rps13* to those of another nuclear gene for a mitochondrial ribosomal protein, *rps10* (for ribosomal protein S10), and I compared levels of mt *rps13* to those of another mitochondrial gene, *cob* (for cytochrome b). I assayed expression levels in a subset of 9 organs and two stress conditions that were assayed above. Numit *rps13* transcripts were present at or above the levels of *rps10* in ovaries and roots, and most organs showed at least half as many transcripts from numit *rps13* as *rps10*

(Figure 4.9A). Cold stress was an exception, where the numit *rps13*/*rps10* ratios were less than 0.4 (Figure 4.9A). Mt *rps13* transcripts were present at slightly over twice the levels of cob transcripts in leaves (Figure 4.9B). The mt *rps13*/*cob* transcript ratio was above 0.5 in most other organs and under cold stress. Overall there is evidence for relatively high levels of steady-state transcripts from numit *rps13* and mt *rps13* in most organs.

4.3.8 Purifying Selection Acting on Numit *rps13*, Nucp *rps13*, and Mt *rps13* in *Malus*

Detecting purifying selection, neutral evolution, and positive selection on each of the three *rps13* genes would provide insights into the selective forces at work and could indicate if one copy is a pseudogene. A d_N/d_S ratio less than one would indicate purifying selection and be evidence for a functional gene, and a d_N/d_S ratio of about one would be evidence of neutral evolution and thus probably pseudogenization. The branch site-specific model in PAML was used to detect adaptive molecular evolution in numit *rps13*, nucp *rps13*, and mt *rps13* for the *Malus* branches by comparing with other plant species. For each LRT, M1 is the model assuming neutral evolution among sites, MA_{test1} assumes that there are positively selected amino acid sites for the foreground lineage, and MA_{test2}, is used to test if the detection of positive selection is an artifact. LRTs showed that MA_{test1} is not significantly better than M1 and MA_{test2} in *Malus* numit *rps13* and *Malus* mt *rps13*, indicating that there is no positive selection for these two branches (Table 4.4). For nucp *rps13*, MA_{test1} is significantly better than M1, but not better than MA_{test2} (Table 4.4), suggesting that there is no positive selection in nucp *rps13*. In addition, branch site-specific models can provide detailed information about how many sites are undergoing purifying selection, neutral evolution, and positive selection. For numit *rps13*, 65% of the sites are under

purifying selection with d_N/d_S less than one, 28% of sites are under neutral evolution with d_N/d_S close to one, and 7% of sites are under positive selection with d_N/d_S greater than one (Table 4.4) without statistical support by LRTs. For nucp *rps13*, 90% of sites are under purifying selection, 7% of sites are under neutral selection, and 3% of sites are under positive selection (Table 4.4) although the positively selected sites are not supported by LRTs. For mt *rps13*, 66% of sites are under purifying selection and 34% of sites are under neutral selection (Table 4.4). The number of sites experiencing each type of selection in numit *rps13* in *Malus* is similar to numit *rps13* in other rosids that do not have mt *rps13* (Table 4.5). Overall my results reveal that the three ribosomal protein S13 genes in *Malus* are undergoing higher purifying selection than neutral evolution, suggesting that their functionality is maintained by purifying selection.

4.4 Discussion

4.4.1 Continuous Accelerated Evolution and Molecular Adaptation of Numit *rps13* among Rosids

After gene duplication some retained duplicated genes undergo asymmetric rate divergence and the faster evolving copy experiences relaxed constraint or positive selection (Conant and Wagner 2003; Byrne and Wolfe 2007). The increased d_N rate, d_N/d_S ratio, and positively selected sites detected in numit *rps13*, compared with nucp *rps13*, across different rosid species show there has been accelerated sequence evolution of numit *rps13*. The increased d_N rate is likely to be correlated with the modified function of numit *rps13* - from encoding a chloroplast ribosomal protein to a mitochondrial ribosomal protein. Some amino acid sequence changes were probably necessary for numit RPS13 to interact well with other proteins in the mitochondrial ribosome.

The increased rate of non-synonymous substitutions is a continuing process instead of there being a burst of sequence change upon formation of numit *rps13* followed by a prolonged period of lower rates of non-synonymous substitutions. The continued higher d_N rate and the two positively selected codons in numit *rps13* near the positions where RPS13 interacts with 16S RNA domain suggest that the sequence or perhaps the function of numit RPS13 in mitochondria is continuing to be refined. Site-specific positive selection has been detected in other duplicated genes (e.g., Sun et al. 2006; Thomas 2006; Johnson and Thomas 2007). For example, in duplicated *AP3* and *PI* genes there has been functional diversification driven by positive selection acting on different sites within a functional domain involved in heterodimerization (Hernandez-Hernandez 2007).

The sixteen amino acids in one or more rosid species that have changed to the amino acid present in mt RPS13 (Figure 4.5) indicate that numit RPS13 is becoming more like mt RPS13. I propose that this is a type of convergent sequence evolution of numit *rps13* that possibly improves the function of numit RPS13 in the ribosome. The amino acid changes appear to have been taking place continuously during rosid evolution (Figure 4.5B), with the largest number (five) having occurred on the branch leading to the legumes *Glycine* and *Medicago*. I speculate that some of the amino acid changes that have taken place along the terminal or subterminal branches of the tree might have allowed numit *rps13* to be selected for, over mt *rps13*, and allowed multiple independent losses of mt *rps13* in different lineages (Figure 4.1). In species where mt *rps13* has been lost the product of numit *rps13* presumably functions as well as, or perhaps better than, mt *rps13* or else there would have been selection for retention of mt *rps13*. Overall the adaptive and convergent evolutionary forces that seem to be acting on numit *rps13* have continued during rosid evolution instead of being factors only soon after gene duplication in a common ancestor of most rosids.

4.4.2 Expression Evolution of Numit *rps13* and Nucp *rps13* in *Arabidopsis*

Expression patterns of some duplicated genes have been shown to evolve in a divergent and sometimes asymmetric manner (Gu et al. 2002; Gu et al. 2005; Chung et al. 2006). For example, many of the genes derived by an ancient polyploidy event in the *Arabidopsis* lineage have undergone considerable expression divergence (Blanc and Wolfe 2004b; Casneuf et al. 2006; Ganko et al. 2007), some of which may have been asymmetric between the two duplicates. Considering that numit *rps13* and nucp *rps13* show asymmetric divergence in sequence, and different subcellular locations of their protein products, I predicted that they also would show asymmetric divergence in expression, with the expression pattern of numit *rps13* being similar to that of other nuclear-encoded genes for mitochondrial ribosomal proteins. Numit *rps13* and nucp *rps13* did show some differences in expression patterns although the differences were only dramatic in roots, senescing leaves, and pollen. Numit *rps13* has evolved a similar expression pattern to three other genes for mitochondrial ribosomal proteins, perhaps by gaining regulatory elements from another gene for a mitochondrial protein, as have several mitochondrial genes that have been transferred to the nucleus (Adams and Palmer 2003). Alternatively there may have been mutations in the regulatory elements of numit *rps13* soon after gene duplication that produced an expression pattern similar to that of other genes for mitochondrial ribosomal proteins. Considering that numit *rps13* was formed in a common ancestor of most rosids, distinguishing between the above possibilities about the origin of its regulatory elements is impossible. Overall numit *rps13* and nucp *rps13* add to the growing number of duplicated genes reported in *Arabidopsis thaliana* that have experienced divergence in expression patterns.

4.4.3 Co-expression of Numit *rps13* and Mt *rps13* in *Malus*

I have shown that numit *rps13* and mt *rps13* in *Malus* are both transcribed in 14 different organ types and under five abiotic stress conditions, mt *rps13* is RNA edited at sites to make the transcripts contain an evolutionarily conserved sequence, and abundant steady-state transcripts from both genes are present in a variety of organs. Both genes are experiencing purifying selection. Taken together, the data obtained in this study indicate that it is likely that both numit *rps13* and mt *rps13* in *Malus* are functional genes and not pseudogenes. However there is the possibility, albeit unlikely, that transcripts from one gene (particularly mt *rps13*) might not be translated. Even if both proteins are present, there is the possibility that only one is assembled into mitochondrial ribosomes, and that process could vary among organs and environmental conditions. Showing that both numit RPS13 and mt RPS13 proteins are assembled into mitochondrial ribosomes, especially in a variety of organ types and under various environmental conditions, is beyond the scope of the current evolutionary study.

Although I have provided evidence for the functionality of the gene products from numit *rps13* and mt *rps13* in *Malus*, might numit *rps13* have experienced functional reversion to encode a chloroplast protein? Several lines of evidence do not support that possibility. The mitochondrial targeting presequence of numit RPS13 is similar to those of numit RPS13 in *Arabidopsis*, *Gossypium*, and *Glycine* (Figure 4.1), each of which was experimentally determined to be imported into mitochondria but not chloroplasts (Adams et al. 2002a; Mollier et al. 2002). Numit *rps13* in Malus does not have a greatly accelerated rate of sequence evolution or show more sites under positive selection, compared with numit *rps13* in other rosids, as might be expected if its product now functions in the chloroplast. Instead purifying selection is operating on numit *rps13*

in *Malus*. Finally, there is no evidence to suggest that nucp *rps13* in *Malus* is a pseudogene and indeed the gene is experiencing strong purifying selection.

Co-expression of numit *rps13* and mt *rps13* genes in *Malus* contrasts with the *atp9* genes in *Neurospora crassa* where the nuclear and mitochondrial copies are expressed during different stages of the life cycle (Bittner-Eddy et al. 1994) and expression has been partitioned between the two genes. Presumably the nuclear copy of *atp9* was derived from transfer of the mitochondrial gene to the nucleus in an ancestor of *Neurospora*, and the current availability of several ascomycete genome sequences could shed light on the evolutionary timing of the gene transfer. Only three other cases of co-expression of nuclear and mitochondrial genes have been reported, to my knowledge, including *cox2* genes in multiple legume species within the Phaseoleae tribe (Adams et al. 1999), *rpl5* genes in wheat (*Triticum aestivum*) (Sandoval et al. 2004), and *sdh4* genes in *Populus* (Choi et al. 2006). Those genes contrast to *rps13* in rosids because the nuclear copies were derived by transfer of the mitochondrial gene to the nucleus.

Considering that numit *rps13* was created by gene duplication in the common ancestor of most rosids, co-expression of mt *rps13* and numit *rps13* in *Malus* has presumably occurred for a long period of evolutionary time. Other cases of co-expression of nuclear and mitochondrial genes in plants represent evolutionarily recent gene transfers to the nucleus. Indeed following transfer of a mitochondrial gene to the nucleus the mitochondrial copy is often lost in a relatively short amount of time (Adams et al. 2002b). Thus it was unexpected that numit *rps13* and mt *rps13* have been preserved without partitioning of expression patterns. Why would both copies continue to be retained and expressed? One possibility is that there is partitioning of expression in organs, tissues, or cell types, or under environmental conditions that were not examined in this study. Another possibility is that, despite the considerable level of sequence divergence between

numit *rps13* and mt *rps13* (about 42% identity), the product of either gene functions well in the mitochondrial ribosomes of *Malus*. That possibility is supported by the fact that some rosids (such as *Arabidopsis*, *Gossypium*, *Glycine*, and *Citrus*) have only the numit *rps13* and non-rosid angiosperms have only mt *rps13*. Another possibility is that mutations have not occurred in numit *rps13* from *Malus* that cause it to be selected for over mt *rps13*, as presumably occurred in other lineages of rosids that have lost mt *rps13*. In this regard it is notable that there are no amino acid changes in numit RPS13 to the residue present in mt RPS13 that occurred along the terminal branch leading to *Malus* (Figure 4.5B), unlike all of the other terminal branches on the tree leading to species that have lost mt *rps13*.

Table 4.1. Gene-specific primers.

Primer ID	Gene	Sequence (5'→3')	Orientation
RT-PCR			
MalusRps13NuF1	Numt *rps13*	GTTGGGGTTACGCGGTTCAATCG	Forward
MalusRps13NuF2	Numt *rps13*	CACGAGGGGCAAAATCTAAGTATCCA	Forward
MalusRps13NuR1	Numt *rps13*	AATAAGTCCAAGACATCTAACAACC	Reverse
MalusRps13NuR2	Numt *rps13*	CACCTTGTTTGATGATGCAACCGCAATC	Reverse
MalusRps13MtF1	Mt *rps13*	GATCATCAGAGAGGAGACAG	Forward
MalusRps13MtF2	Mt *rps13*	TCAGGAGCTAGATCAGTTGCCGA	Forward
MalusRps13MtR1	Mt *rps13*	TTAGTATGAGTTCGTTGACCG	Reverse
MalusRps13MtR2	Mt *rps13*	GAAGCTATCAGTGTGATTCATCAGAC	Reverse
qRT-PCR			
MalusAct2F1	*ACT2*	AATGGTGAAGGCTGGATTTGCTGG	Forward
MalusAct2R1	*ACT2*	TGACCCATACCAACCATGACACCA	Reverse
MalusRps10F1	Numt *rps10*	CAAGGAAGATTGCACTGCCGGAAT	Forward
MalusRps10R1	Numt *rps10*	CGTATTGGGCTCCAAATATGCGCT	Reverse
qNumtRps13F1	Numt *rps13*	CGCGGTTCAATCGCAATCGTTTCT	Forward
qNumtRps13R1	Numt *rps13*	ATCCCACGTATGTACTGAAGCGC	Reverse
qMtRps13F1	Mt *rps13*	AGGAGCTAGATCAGTTGCCGATGA	Forward
qMtRps13R1	Mt *rps13*	CCTAATCGATAACGAACCTGAATGGC	Reverse

Table 4.2. Comparison of d_N/d_S ratios between nucp *rps13* and numit *rps13*.

Taxon	*Arabidopsis*	*Citrus*	*Glycine*	*Gossypium*	*Malus*	*Medicago*	*Populus*
nucp *rps13* (0.06 ± 0.02)[a]							
Arabidopsis	-						
Citrus	0.019761	-					
Glycine	0.031462	0.111523	-				
Gossypium	0.045554	0.051540	0.076516	-			
Malus	0.037126	0.049357	0.081350	0.059803	-		
Medicago	0.040972	0.084293	0.059997	0.057981	0.085442	-	
Populus	0.048739	0.112004	0.062117	0.049504	0.061190	0.067945	-
numit *rps13* (0.28 ± 0.08)							
Arabidopsis	-						
Citrus	0.276419	-					
Glycine	0.331700	0.253453	-				
Gossypium	0.257117	0.308765	0.299647	-			
Malus	0.299325	0.317365	0.238339	0.468253	-		
Medicago	0.267372	0.179926	0.032697	0.336763	0.273923	-	
Populus	0.382829	0.261000	0.241346	0.294584	0.301918	0.170266	-

[a]The mean and standard deviation are shown in parentheses.

Table 4.3. LRT statistics of site specific model for numit *rps13* and nucp *rps13*.

Gene	Number of sequences	Tree Length[a]	d_S[a]	d_N[a]	d_N/d_S[a]	11th class from M8		M7-M8 comparison		M8a-M8 comparison		Positively selected sites
						p, %	d_N/d_S	2δL	*P* Value	2δL	*P* Value	
numit *rps13*	7	6.93	5.35	1.30	0.24	2.7	9.7400	9.0670	**0.0107**[b]	7.5048	**0.0031**	**28R,** 114R ($P > 0.8$)
nucp *rps13*	7	4.08	5.16	0.31	0.06	1.6	1.6268	4.7952	0.0909	0.8448	0.1790	None

[a]Estimated using M0 model in PAML.
[b]Boldface indicates the statistically significant difference ($P < 0.05$).

Table 4.4. LRT statistics of branch specific model for *Malus* branch.

Gene	Parameters from MA_{test1}	M1-MA_{test1} comparison		MA_{test2}-MA_{test1} comparison		Selected positive sites
		2δL	*P* Value	2δL	*P* Value	
numit *rps13*	$p_0 = 0.6485$, $p_1 = 0.2817$, ($p_2 + p_3 = 0.0698$) $\omega_0 = 0.12$, $\omega_1 = 1$, $\omega_2 = 112.98$	1.7396	0.4190	1.2140	0.2705	31H, 32Q, 117R ($0.5 < P < 0.8$)
nucp *rps13*	$p_0 = 0.9005$, $p_1 = 0.0733$, ($p_2 + p_3 = 0.0262$) $\omega_0 = 0.04$, $\omega_1 = 1$, $\omega_2 = 96.7$	9.7274	**0.0077**[a]	3.5026	0.0613	**32Q, 105C**
mt *rps13*	$p_0 = 0.6646$, $p_1 = 0.3354$, (p2 + p3 = 0) $\omega_0 = 0$, $\omega_1 = 1$, $\omega_2 = 1$	0	1	0	1	None

[a]Boldface indicates the statistically significant difference ($P < 0.05$).

Table 4.5. Selection on numit RPS13 in seven rosid species.

Taxon	Purifying Selection (%, $d_N/d_S < 1$)[a]	Neutral Evolution (%, $d_N/d_S = 1$)[a]	Positive Selection (%, $d_N/d_S > 1$)[a]	M1-MA_{test1} comparison		MA_{test2}-MA_{test1} comparison		Selected positive sites
				2δL	*P* Value	2δL	*P* Value	
Malus	65	28	7 (120.33)[b]	1.7396	0.4190	1.2140	0.2705	31H, 32Q, 117R
Citrus	65	30	5 (154.19)	4.3230	0.1152	3.9182	**0.0478**[c]	50K, 92H
Glycine	67	33	0	0	1	0	1	35C
Medicago	67	33	0	0	1	0	1	None
Gossypium	64	36	0	0.6314	0.7293	0.0024	0.9609	108K
Populus	66	30	4 (18.05)	3.6624	0.1603	3.4746	0.0623	39I, 88S, 102S
Arabidopsis	52	48	0	6.1770	**0.0456**	0	1	18G, 36H, 60G, 83D

[a]Estimated using MA_{test1}.
[b]Bracket indicates d_N/d_S ratio. If d_N/d_S ratio is close to one for the third and fourth class of MA_{test1} without statistical support to be greater than one, it will be counted into the category of neutral evolution.
[c]Boldface indicates a statistically significant difference ($P < 0.05$).

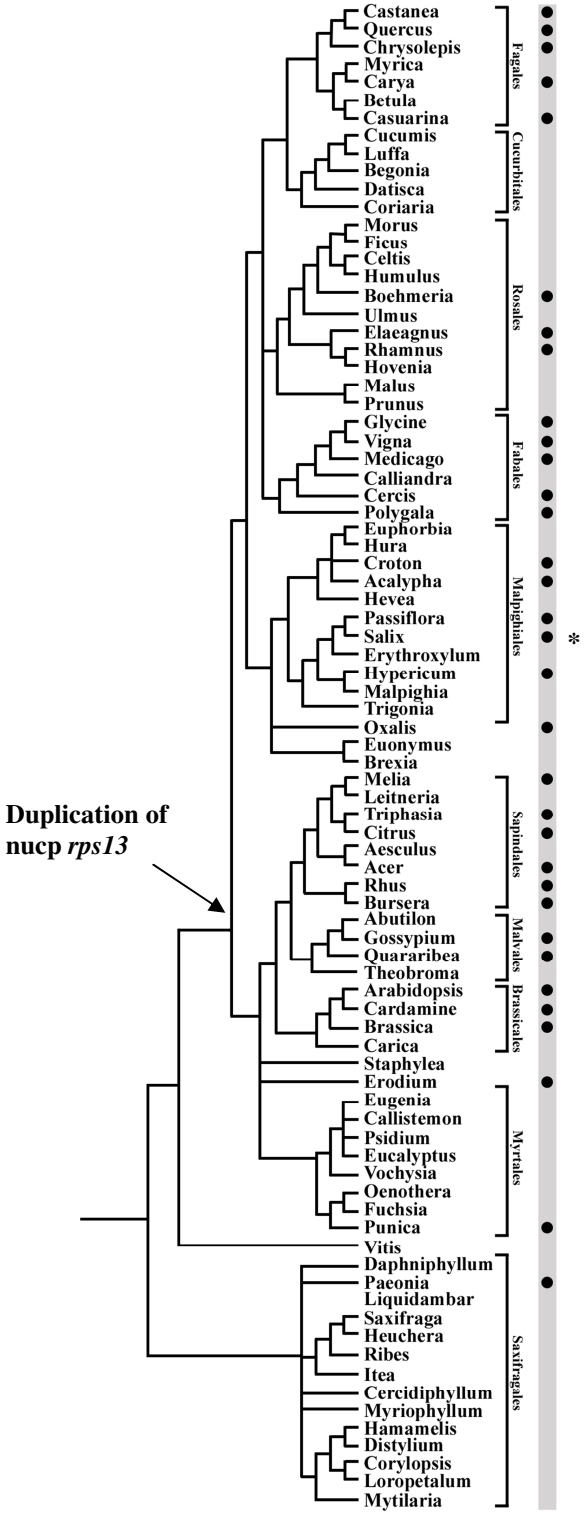


Figure 4.1. Phylogenetic tree showing losses of mt *rps13* among rosids. Bullets indicate loss of mt *rps13* as judged by Southern blot hybridizations. Data are from Adams et al. (2002b) and this figure was modified and redrawn from a figure in that paper. Rosid species included in this study are shown by rectangular boxes. Note that *Populus* is closely related to *Salix* (Asterisk).

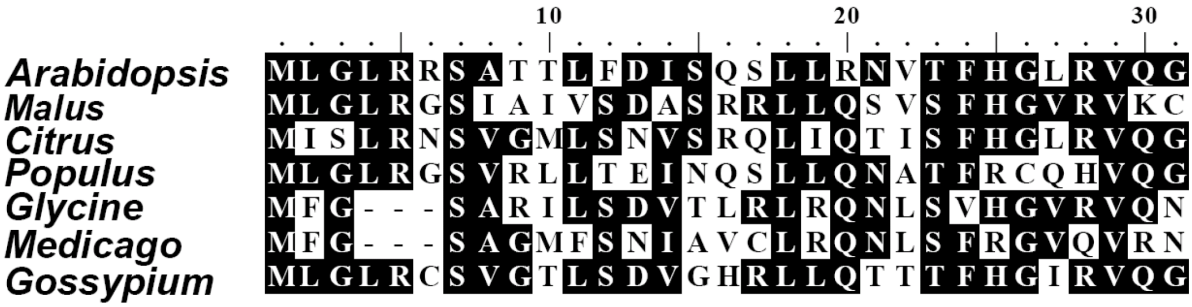


Figure 4.2. Mitochondrial targeting sequence alignment of numit RPS13. Predicted targeting presequences of numit RPS13 from seven rosid species are aligned. Identical amino acid residues are marked in white on the black background. Dots indicate gaps inserted to improve the alignment.

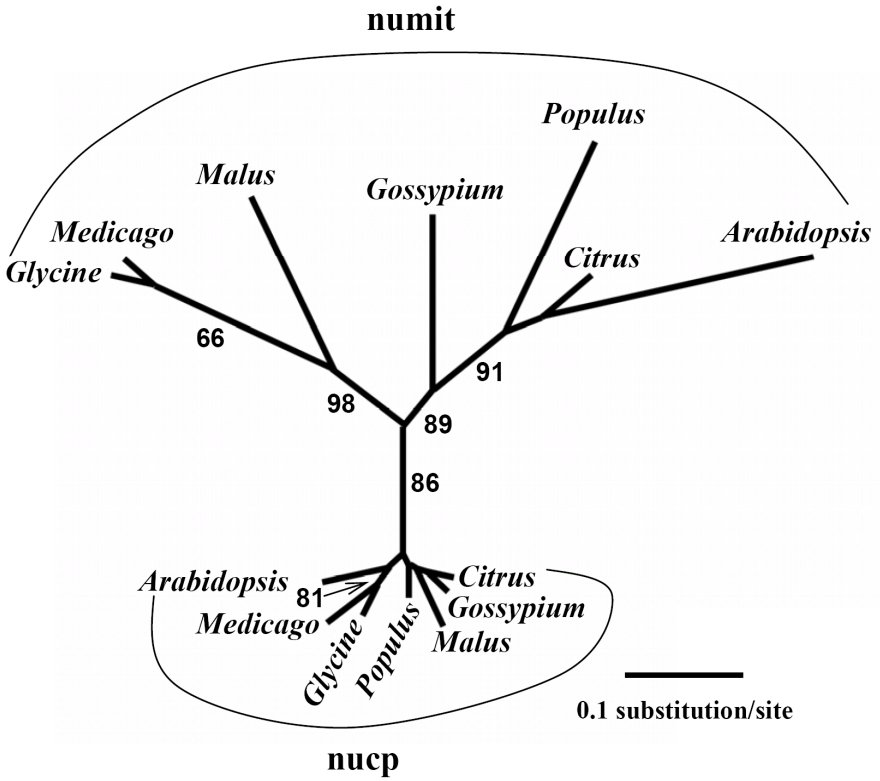


Figure 4.3. Phylogenetic analysis of numit *rps13* and nucp *rps13* from rosid species. Shown is an unrooted phylogram derived from maximum likelihood analyses of the first and second nucleotide positions. Bootstrap values from 100 replicates of ML analyses are labeled on the internodes.

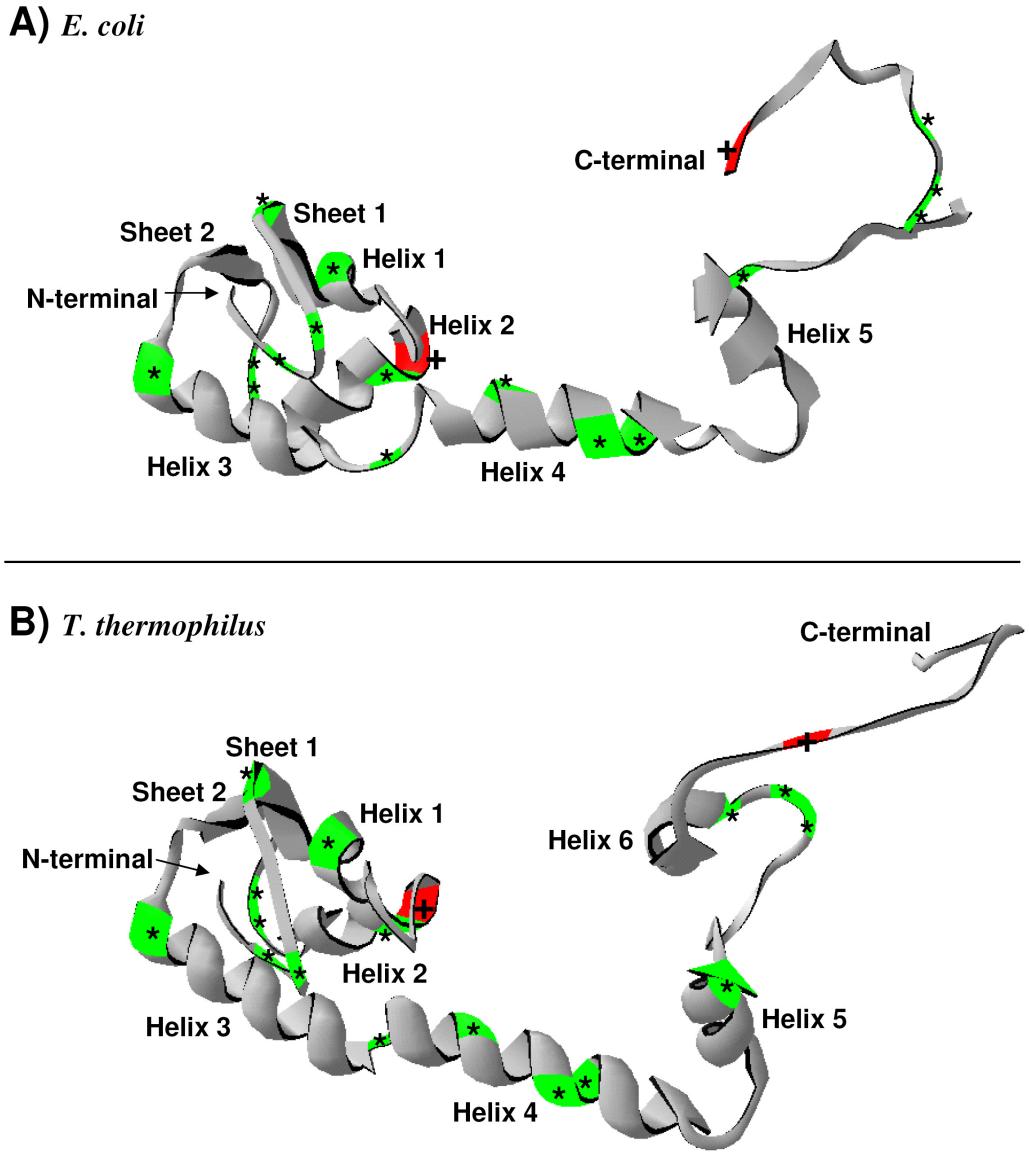


Figure 4.4. Tertiary structure of numit RPS13. The structures of RPS13 from *E. coli* (A) and *T. thermophilus* (B) are shown with amino acids from numit RPS13 plotted, as in the alignment (Figure 4.5). Red regions and plus signs indicate relative positions of positively selected sites. Green regions and asterisks indicate relative positions of amino acids in numit *Rps13* in at least one rosid species that have mutated to the amino acid present in mt *rps13* genes (Figure 4.5). Among them, sites 4 and 10 (Figure 4.5) are located in the coil region of the N-terminal end, site 15 is in the beginning of the first β-sheet, site 21 is in the first α-helix, site 32 is in the second α-helix, sites 42 and 43 are in the coil region between the second α-helix and the second β-sheet, site 53 are in the third α-helix, site 67 is in the coil region between the third α-helix and the fourth α-helix, sites 74, 79, and 80 are in the fourth α-helix, and sites 97, 107, 108, and 110 are in the C-terminal coil region.

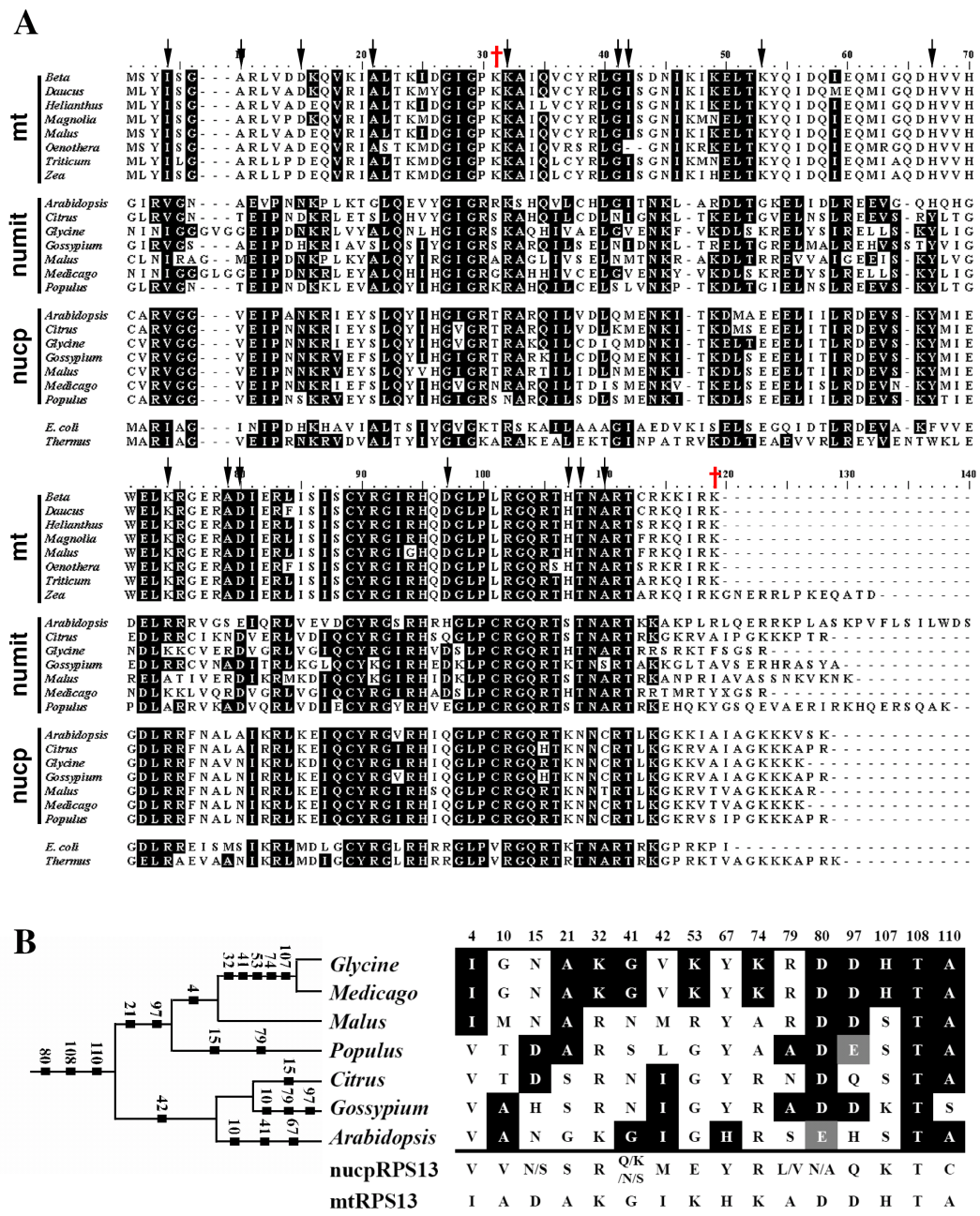


Figure 4.5. Amino acid evolution in numit RPS13. (A) Alignment of mt RPS13, numit RPS13, nucp RPS13 sequences, and RPS13 from *E. coli* and *Thermus thermophilis*. Amino acids in numit RPS13 that are identical to those in mt RPS13 or nucp RPS13 are shown on a black background. Dashes indicate gaps inserted to improve alignment. Red plus signs indicate positions of positively selected amino acids. Black arrows indicate amino acids in numit *rps13* in at least one rosid species that have mutated to the amino acid present in all eight mt *rps13* genes, summarized in panel B. Panel B also includes a phylogeny of the rosid species and my hypothesized evolutionary timing of each amino acid change. Numbers in panel B indicate positions in the alignment shown in panel A. Abbreviations: mt, mitochondrial protein; nucp, nuclear-encoded chloroplast protein; numt, nuclear-encoded mitochondrial protein.

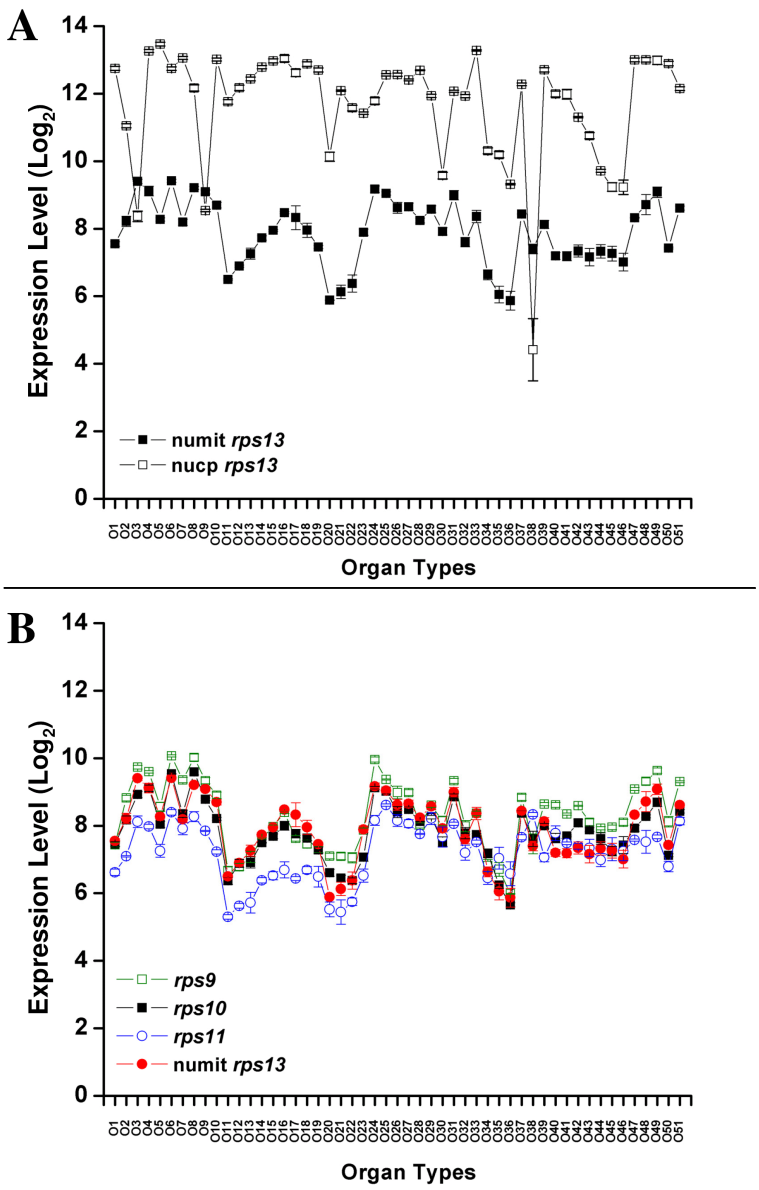
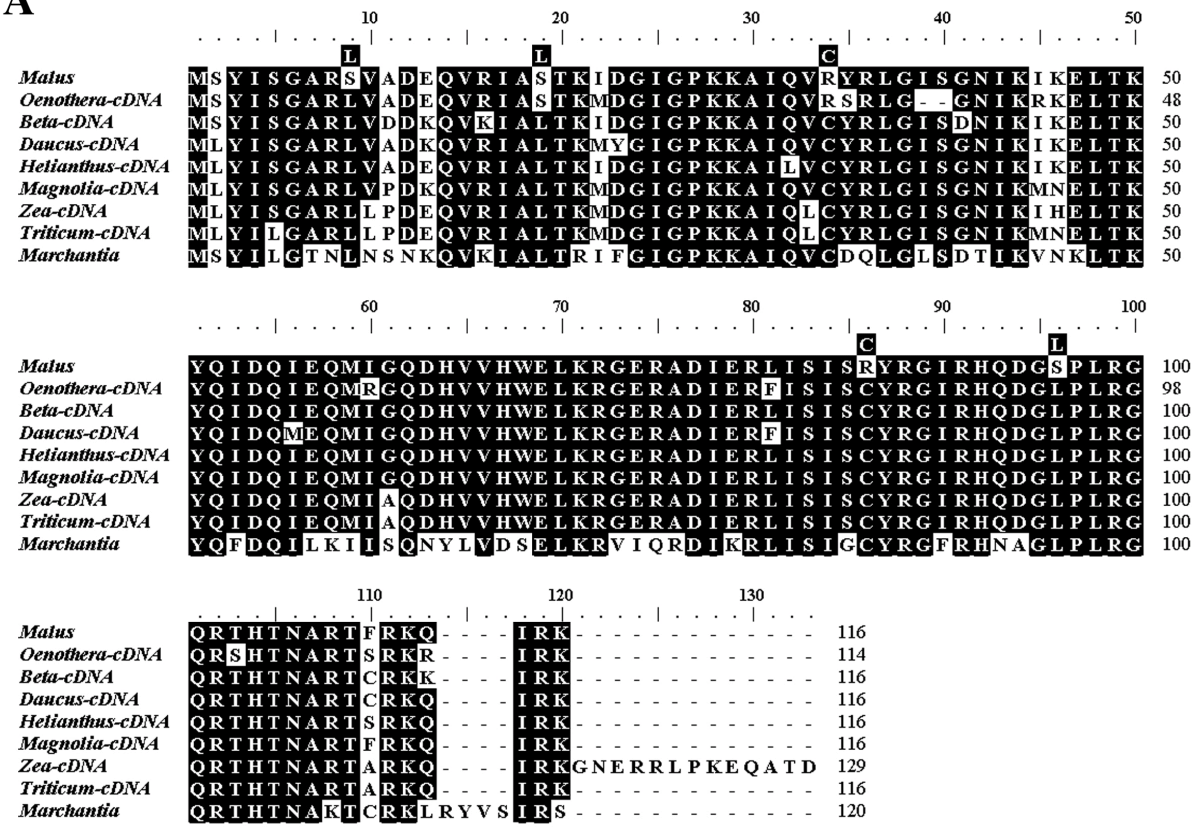


Figure 4.6. Expression patterns of numit *rps13* in *Arabidopsis thaliana*. Shown are graphs from ANOVA analysis of microarray data from 51 organs and developmental stages (see Methods). Organ types and developmental stages are listed in Additional File 3. Error bars show 3 biological replicates. The Y-axis indicates the expression level normalized by log2. (A) Numit *rps13* compared with nucp *rps13*. (B) Numit *rps13* compared with three other nuclear-encoded mitochondrial ribosomal protein genes, *rps9, rps10*, and *rps11*.

A



B

RNA editing of mt *rps13* among different plant species.

Codon	*Malus*	*Oenothera*	*Beta*	*Daucus*	*Helianthus*	*Magnolia*	*Triticum*	*Zea*
2	TCA (S)	TCA (S)	TCA (S)	TTA (L)	TTA (L)	TCA (S→L)	TCA (S→L)	TCA (S→L)
5	TCA (S)	TCA (S)	TCA (S)	TCA (S)	TCA (S)	TCA (S)	TCA (S→L)	TCA (S)
9	TCA (S→L)	TCA (S→L)	TTA (L)	TCA (S→L)	TCA (S→L)	TCA (S→L)	TCA (S→L)	TCA (S→L)
19	TCA (S→L)	TCA (S)	TCA (S→L)	TCA (S→L)	TCA (S→L)	TCA (S→L)	TCA (S→L)	TCA (S→L)
34	CGT (R→C)	CGT (R)	TGT (C)	CGT (R→C)	CGT (R→C)	CGT (R→C)	CGT (R→C)	CGT (R→C)
78	ATC (I)	ATC (I)	ATC (I→I)	ATC (I)	ATC (I)	ATC (I)	ATC (I)	ATC (I)
81	TTA (L)	TTC (F→F)	TTA (L)	TTC (F→F)	TTA (L)	TTA (L)	TTA (L)	TTA (L)
86	CGT (R→C)	TGT (C)	TGT (C)	TGT (C)	TGT (C)	CGT (R→C)	CGT (R→C)	CGT (R→C)
96	TCG (S→L)	TCG (S→L)	TTG (L)	TTG (L)	TCA (S→L)	TCG (S→L)	TCG (S→L)	TCG (S→L)

Figure 4.7. RNA editing of mitochondrial *rps13* in *Malus domestica* compared with other plants. (A) RPS13 sequences of *Malus domestica* deduced from genomic DNA sequence are aligned with deduced cDNA sequences of seven rosid species and the genomic sequence of *Marchantia polymorpha*. Amino acids changed by RNA editing for *Malus domestica* are given above the genomic DNA sequence. Identical amino acids are shaded in black. Dots refer to gaps inserted to improve the alignment or missing amino acids. Numbers of amino acids are indicated on the right side of each sequence. (B) Genomic triplets are shown for an RNA editing site in at least one plant species. Underlined nucleotide indicates the site where a C-to-U transition occurs. The corresponding amino acid change is indicated and signified by the single letter code in parenthesis.

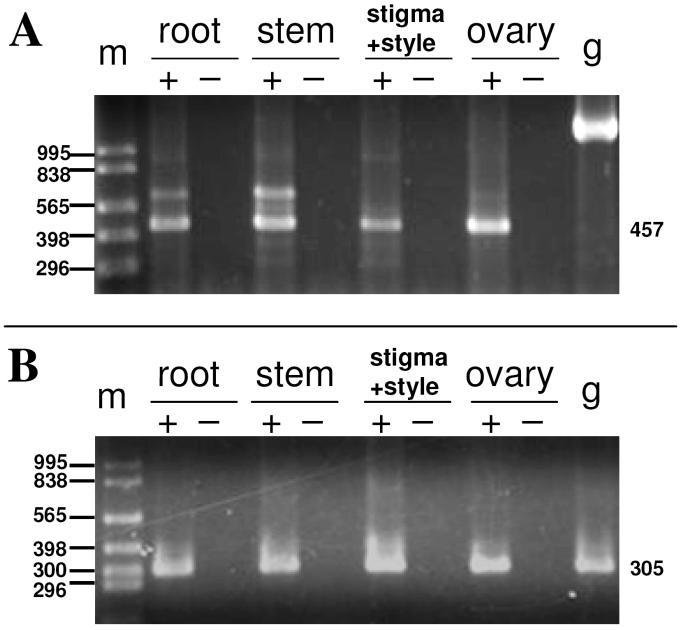


Figure 4.8. Expression of numit *rps13* and mt *rps13* from in *Malus* in different organ types. Plus signs indicate reactions containing reverse transcriptase (RT) and minus signs indicate reactions without RT. Abbreviations: g, genomic; m, marker. (A) A gel showing a subset of the RT-PCR products of numit *rps13* (457 bp). Sequencing of a larger band in RT-PCR assays of numit *rps13* revealed that it was due to non-specific primer binding, whereas the smaller band was the numit *rps13*. (B) A gel showing a subset of the RT-PCR products of mt *rps13* (305 bp).

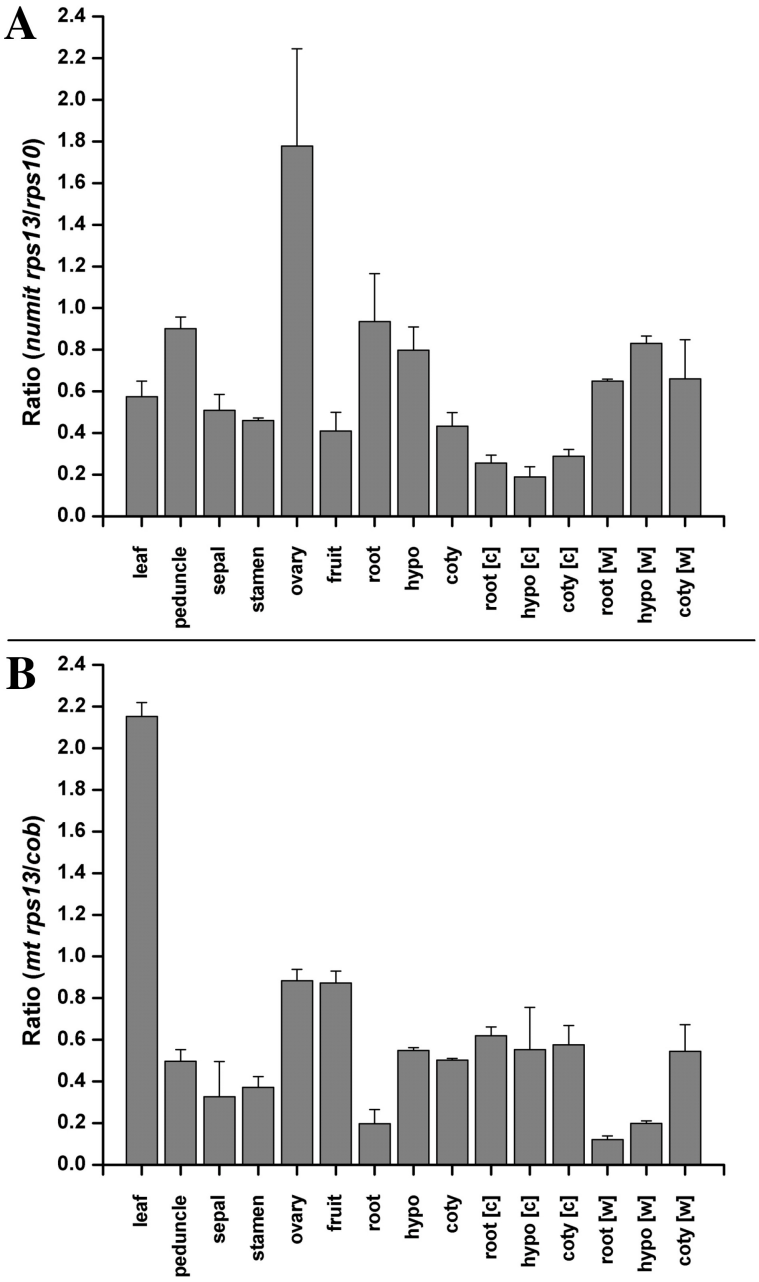


Figure 4.9. qRT-PCR of numit *rps13* and mt *rps13* in *Malus*. (A) Ratios of numit *rps13* to *rps10* transcripts. (B) Ratios of mt *rps13* to *cob* transcripts. Error bars show variation from two biological replicates.

5 Concluding Chapter

Results from my dissertation provide new insights into how duplicated genes diverge in expression patterns, functions, and protein subcellular localization. Expression divergence was the focus of chapter 2, and the topic also was considered in chapters 3 and 4. I studied functional divergence in chapter 3. I characterized examples of protein subcellular localization in chapters 2 and 4. Another aspect of duplicate gene evolution, sequence divergence, was a component of all three chapters. Below I briefly explain how my research has advanced our knowledge in each of those areas. Additional discussion of how each study relates to, and advances, the field is provided in the Discussion sections of chapter 2, 3, and 4.

5.1 Reciprocal Expression Patterns Between Duplicated Genes

In chapter 2, I assessed the frequency of reciprocal expression pattern between duplicated genes in *Arabidopsis thaliana* in different cell types, organ types, tissue types, and developmental stages on a large scale. Reciprocal expression is considered to be an important type of expression divergence that contributes to the retention of duplicated genes (e.g., Adams et al. 2003; Bottley et al. 2006; Chaudhary et al. 2009; Buggs et al. 2010). My results showed that a large number of duplicated genes (30-38%) were reciprocally expressed among different cell types, organ types, tissue types, and developmental stages, indicating that reciprocal expression is a common phenomenon that plays an important role in the retention of duplicated genes. My study is the first attempt to investigate the frequency of reciprocal expression on a large scale in plants across a large number of cell types, organ types, tissues, and developmental stages. Instead of regulatory subfunctionalization, my results from chapter 2 suggested that reciprocal expression

patterns in the majority of duplicated genes result from regulatory neofunctionalization. This observation might support the point of view, “subfunctionalization downgraded”, by Freeling (2008), who indicated that subfunctionalization is not a major factor in duplicate gene retention.

5.2 Functional Divergence of Duplicated Genes

To better understand the effects of reciprocal expression and regulatory neofunctionalization on phenotypic innovation, I conducted a detailed evolutionary analysis for a pair of protein kinase genes, *SHORT SUSPENSOR* (*SSP*) and *BRASSINOSTEROID KINASE 1* (*BSK1)*, that showed reciprocal expression (chapter 3). In this case study, I discovered a dramatic change of expression pattern and function for *SSP*, which was formed by a whole genome duplication event at the base of the Brassicaceae family. After duplication, *SSP* acquired a new expression pattern in pollen compared to its duplicated partner, *BSK1*, which is expressed in most organ types but not in pollen. In addition to the acquisition of a new expression pattern, *SSP* lost its ancestral function of involvement in the brassinosteroid signaling pathway and gained a new function n regulation of the first cell division during embryo development.

This study of the *SSP* gene provided an excellent example for the classic neofunctionalization process proposed by Ohno (1970). Examples that link the expression divergence between duplicated genes with its impact on phenotypic changes remains rare in the literature. An example from animals that has similarities to the *SSP* case was reported for a pair of tandemly duplicated genes in *Drosophila* (Loppin et al. 2005). A duplicated gene, *K81*, derived from an evolutionarily recent retroposition event within the *melanogaster* subgroup showed a paternal effect on the zygote viability. After gene duplication, *K81* lost its ancestral ubiquitous expression pattern and showed a male germline-specific expression. In addition, *K81* underwent an

accelerated sequence evolution and the knock-out of this gene caused the zygote inviability, thereby showing a paternal effect on the zygote viability.

SSP adds to the small number of cases of gain of a new function after gene duplication in literature. Together with other exmples of neofunctionalization (see chapter 1 for details), gene duplication can indeed provide raw materials for physiological and morphological innovations. Some of these new traits from neofunctionalized duplicated genes might be beneficial for the adaptation of organisms over evolution.

5.3 Protein Subcellular Relocalization After Gene Duplication

Divergence in protein subcellular relocalization (PSR) can be an outcome of gene duplication. From large-scale studies of duplicated genes in yeasts and fish, PSR after gene duplication was shown to play an important role in the retention of duplicated genes in yeasts (at least 25%) and fish (ca. 14%) (Marques et al. 2008; Kassahn et al. 2009). It has been shown that gene duplication can promote the PSR because of the relaxation of localization constraint compared to single copy genes (e.g., Szklarczyk and Huynen 2009). In chapter 2 and chapter 4, I examined three examples of PSR after gene duplication. The first two examples are duplicated genes derived from a whole genome duplication event at the base of the Brassicaceae family: a pair of class III peroxidase genes and a pair of calcium-dependent protein kinase genes (chapter 2). The third example is a pair of nuclear-encoded genes for organellar ribosomal protein S13, derived from a duplication event shared by most rosids (chapter 4). In each case there was a change in subcellular localization that was accompanied by expression pattern changes, including reciprocal expression patterns in two of the cases, an accelerated rate of amino acid sequence

evolution, and critical sequence changes at the N-terminus of the encoded proteins. In each case there is evidence for neofunctionalization. The third example, numit *Rps13,* showed a signature of positive selection, indicative of molecular adaptation. The two positively selected amino acids were shown to be involved in the interaction with the messenger RNA and ribosomal RNA, suggesting that strong selection favored the replacement of these two amino acids to potentially improve the ribosomal protein to interact better with messenger RNA and ribosomal RNA in the new mitochondrial environment. Interestingly, I showed that expression and protein sequence of this copy underwent convergent evolution and evolved to have a higher similarity of expression and protein sequence with the mitochondrial-encoded ribosomal S13 gene (mt *rps13*). My thesis provided some interesting examples where PSR after gene duplication can play an important role on the retention of duplicated genes. After PSR, gene might undergo adaptive evolution to function better in the new subcellular environment.

5.4 Possible Future Directions

My thesis research has raised additional questions about reciprocal expression, neofunctionalization, and protein subcellular relocalization after gene duplication. Below I describe three possible projects that could be pursued in the future. First, I found that pollen is the most common structure, out of those examined, for expression gain after gene duplication. Since pollen is the male gametophyte and belongs to a different stage of life history compared with the sporophytic organ types, my finding suggests that gametophytes might serve as an important location for evolutionary innovations (i.e., gain of a new function) or arm race (i.e., sexual conflict) after gene duplication. To test this hypothesis, one could assay reciprocal expression patterns between duplicated genes by incorporating transcriptomic data from the

female gametophyte (i.e., egg cells, central cells, and synergids). To determine if genes become expressed in the female gametophyte after gene duplication, i.e. regulatory neofunctionalization, one could reconstruct the ancestral expression pattern of duplicated genes with reciprocal expression that are involved in the female gametophyte. In addition, one could perform detailed sequence and selection analyses for some cases. These analyses might shed light on the evolutionary importance of gene duplication on the development of the gametophyte in plants.

Second, I found that a pair of class IV homeodomain-leucine zipper genes, AT4G25330 (*HDG6*; *FWA*) and AT5G52170 (*HDG7*), show a reciprocal expression pattern between siliques and roots. *HDG6* (or *FWA*) is a maternally imprinted gene with a siliques-specific expression pattern (Kinoshita et al. 2004), whereas *HDG7* is not an imprinted gene with a predominant expression in roots, but no expression in siliques (Nakamura et al. 2006). These two duplicated genes arose from a whole genome duplication event at the base of the Brassicaceae family from a non-imprinted ancestral gene (Nakamura et al. 2006). Such an observation raises an interesting question: *HDG6* might acquire a new expression pattern in siliques and gain a new function. To determine if silique-specific expression in the imprinted copy, *HDG6*, is a derived state, the ancestral, pre-duplication, expression pattern of this duplicated gene pair could be determined by assaying expression of orthologs in outgroup species. In addition, to see if *HDG6* underwent accelerated and asymmetric sequence rate evolution after gene duplication, sequence analyses including asymmetric sequence rate analysis and positive selection analysis could be conducted. A detailed analysis for this case may provide some insights into the association between regulatory diversification and genomic imprinting after gene duplication in plants.

In contrast to expression divergence, protein subcellular relocalization (PSR) can also play an important role in the retention of duplicated genes in plants. Although I provided some examples

of potential neofunctionalization by neolocalization after gene duplication, it is still unknown the contribution of PSR on the retention of duplicated genes in a large-scale study in plants. A large-scale or genome wide study will shed light on the importance of PSR on the retention of duplicated genes in plants. To pursue this question, one could assess the frequency of PSR after gene duplication using the subcellular localization database in *Arabidopsis thaliana* that contains data from GFP and mass-spectrometry analyses. One could also perform detailed sequence and selection analysis for some cases to examine how PSR contributes to the retention of duplicated genes in plants.

References

Achaz G, Coissac E, Viari A, Netter P. 2000. Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. *Mol Biol Evol*. 17: 1268-1275.

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci USA*. 100: 4649-4654.

Adams KL, Daley DO, Whelan J, Palmer JD. 2002a. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. *Plant Cell*. 14: 931-943.

Adams KL, Palmer JD. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol Phylogenet Evol*. 29: 380-395.

Adams KL, Qiu YL, Stoutemyer M, Palmer JD. 2002b. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene Loss and transfer to the nucleus during angiosperm evolution. *Proc Natl Acad Sci USA*. 99: 9905-9912.

Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, Palmer JD. 1999. Intracellular gene transfer in action: Dual transcription and multiple silencings of nuclear and mitochondrial cox2 genes in legumes. *Proc Natl Acad Sci USA*. 96: 13863-13868.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opinion Plant Biol*. 8: 135-141.

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol*. 24: 1219-1228.

Backues SK, Korasick DA, Heese A, Bednarek SY. 2010. The *Arabidopsis* dynamin-related protein2 family is essential for gametophyte development. *Plant Cell*. 22: 3218-3231.

Bailey CD, Koch MA, Mayer M, Mummenhoff K, O'Kane SL Jr, Warwick SI, Windham MD, Al-Shehbaz IA. 2006. Toward a global phylogeny of the Brassicaceae. *Mol Biol Evol*. 23: 2142-2160.

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. 2002. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*. 18: 298-305.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidization during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol*. 25: 2445-2455.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. *Genome Biol Evol*. 1: 391-399.

Bayer M, Nawy T, Giglione C, Galli M, Meinnel T, Lukowitz W. 2009. Paternal control of embryonic patterning in *Arabidopsis thaliana*. *Science*. 323: 1485-1488.

Becker JD, Boavida LC, Carneiro J, Haury M, Feijo JA. 2003. Transcriptional profiling of *Arabidopsis* tissues reveals the unique characteristics of the pollen transcriptome. *Plant Physiol*. 133: 713-725

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J. 2006. Positive selection driving diversification in plant secondary metabolism. *Proc Natl Acad Sci USA*. 103: 9118-9123.

Bernasconi G, Ashman TL, Birkhead TR, Bishop JD, Grossniklaus U, Kubli E, Marshall DL, Schmid B, Skogsmyr I, Snook RR, Taylor D, Till-Bottraud I, Ward PI, Zeh DW, Hellriegel B. 2004. Evolutionary ecology of the prezygotic stage. *Science*. 303: 971-975.

Bikard D, Patel D, Metté CL, Giorgi V, Camilleri C, Bennett MJ, Loudet O. 2009. Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*. 323:

623-626.

Bininda-Emonds OR. 2005. transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*. 6: 156.

Birchler JA, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: Implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol*. 234: 275-288.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell*. 19: 395-402.

Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN. 2003. A gene expression map of the *Arabidopsis* root. *Science*. 302: 1956-1960.

Bittner-Eddy P, Monroy AF, Brambl R. 1994. Expression of mitochondrial genes in the germinating conidia of Neurospora crassa. *J Mol Biol*. 235: 881-897.

Blackman BK, Strasburg JL, Raduski AR, Michaels SD, Rieseberg LH. 2010. The role of recently derived *FT* paralogs in sunflower domestication. *Curr Biol.* 20: 629-635.

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

Blanc G, Wolfe KH. 2004a. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*. 16: 1667-1678.

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

Bonen L, Calixte S. 2006. Comparative analysis of bacterial-origin genes for plant mitochondrial ribosomal proteins. *Mol Biol Evol*. 23: 701-712.

Bottley A, Xia GM, Koebner RMD. 2006. Homoeologous gene silencing in hexaploid wheat. *Plant J*. 47: 897-906.

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

Bradley D., Carpenter R., Sommer H., Hatley N., Coen E. 1993. Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of Antirrhinum. *Cell*. 72: 85-95.

Brady SM, et al. 2007. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*. 318: 801-806.

Broderson DE, Clemons Jr. WM, Carter AP, Wimberly BT, Ramakrishnan V. 2002. Crystal structure of the 30S ribosomal subunit from *Thermus thermophilus*: structure of the proteins and their interactions with 16S RNA. *J Mol Biol*. 316: 725-768.

Buggs RJ, Elliott NM, Zhang L, Koh J, Viccini LF, Soltis DE, Soltis PS. 2010. Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *New Phytol.* 186: 175-183.

Byrne KP, Wolfe KH. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*. 175: 1341-1350.

Byun-McKay SA, Geeta R. 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. *Trends Ecol Evol*. 22: 338-344.

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

Carter AP, Clemons WM, Broderson DE, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V.

2000. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*. 407: 340-348.

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

Chaudhary B, et al. 2009. Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). *Genetics*. 182: 503-517.

Choi C, Liu Z, Adams KL. 2006. Evolutionary transfers of mitochondrial genes to the nucleus in the Populus lineage and coexpression of nuclear and mitochondrial Sdh4 genes. *New Phytol*. 172: 429-439.

Chung WY, Albert R, Albert I, Nekrutenko A, Makova KD. 2006. Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC Bioinformatics*. 27: 7: 46.

Claros MG, Vincens P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem*. 241: 779-786.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J*. 16: 735-743.

Coca M, San Segundo B. 2010. *AtCPK1* calcium-dependent protein kinase mediates pathogen resistance in *Arabidopsis*. *Plant J*. 63: 526-540.

Coen ES, Meyerowitz EM. 1991. The war of the whorls: genetic interactions controlling flower development. *Nature*. 353: 31-37.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* 13: 2052-2058.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*. 9: 938-950.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res*. 16: 738-749.

Dammann C, Ichida A, Hong B, Romanowsky SM, Hrabak EM, Harmon AC, Pickard BG, Harper JF. 2003. Subcellular targeting of nine calcium-dependent protein kinase isoforms from *Arabidopsis*. *Plant Physiol*. 132: 1840-1848.

Demuth JP, Hahn MW. 2009. The life and death of gene families. *BioEssays*. 31: 29-39.

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*. 454: 762-765.

Drea SC, Lao NT, Wolfe KH, Kavanagh TA. 2006. Gene duplication, exon gain and neofunctionalization of *OEP16*-related genes in land plants. *Plant J*. 46: 723-735.

Duarte J.M., Cui L., Wall P.K., Zhang Q., Zhang X., Leebens-Mack J., Ma H., Altman N., dePamphili C.W. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol Biol Evol*. 23: 467-478.

Dykhuizen D, Hartl DL. 1980. Selective neutrality of *6PGD* allozymes in *E. coli* and the effects of genetic background. Genetics. 96: 801-817.

Edgar, Robert C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32: 1792-1797.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res*. 17: 699-717.

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

Farré D, Albà MM. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Mol Biol Evol*. 27: 325-335.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783-791.

Felsenstein J. 2008. PHYLIP (phylogeny inference package), version 3.6.8. Distributed by the author. Department of Genetics, University of Washington, Seattle, USA.

Felsenstein J. 2009. PHYLIP (Phylogeny Inference Package) version 3.6.9. Distributed by the author. Department of Genetics, University of Washington, Seattle, USA.

Fiebig A, Kimport R, Preuss D. 2004. Comparisons of pollen coat genes across Brassicaceae species reveal rapid evolution by repeat expansion and diversification. *Proc Natl Acad Sci USA*. 101: 3286-3291.

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

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16: 805-814.

Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn*. 4: 28-40.

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

Friedman R, Hughes AL. 2006. Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal nonconservative and anomalous properties of widely used methods. *Mol*

Phylogenet Evol. 42: 388-393.

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

Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. *Nature*. 403: 886-889.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci USA*. 102: 707-712.

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

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

Ha M, Li WH, Chen ZJ 2007. External factors accelerated expression divergence between duplicate genes. *Trends Genet*. 23: 162-166.

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

Haerizadeh F, Wong CE, Bhalla PL, Gresshoff PM, Singh MB. 2009. Genomic expression profiling of mature soybean (*Glycine max*) pollen. *BMC Plant Biol*. 9: 25.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*. 100: 605-617.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp Ser*. 41: 95-98

Hartung F, Suer S, Puchta H. 2007. Two closely related RecQ helicases have antagonistic roles in homologous recombination and DNA repair in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*. 104: 18836-18841.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet*. 10: 551-564.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*. 169: 1157-1164.

Hernandez-Hernandez T, Martinez-Castilla LP, Alvarez-Buylla ER. 2007. Functional diversification of B MADS-box homeotic regulators of flower development: adaptive evolution in protein-protein interaction domains after major gene duplication events. *Mol Biol Evol*. 24: 465-481.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*. 449: 677-681.

Honys D, Twell D. 2003. Comparative analysis of the *Arabidopsis* pollen transcriptome. *Plant Physiol*. 132: 640-652.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 17: 754-755.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*. 256: 119-124.

Hulsen T, Huynen MA, de Vlieg J, Groenen PM. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*. 7: R31.

Innan H., Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11: 97-108.

Ishimizu T, et al. 1998. Identification of regions in which positive selection may operate in *S-RNase* of Rosaceae: implication for S-allele-specific recognition sites in *S-RNase*. *FEBS Lett*. 440: 337-342.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Fabbro CD, Alaux M, Gaspero GD, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Clainche IL, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 449: 463-467.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H. 2006. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol*. 6: 32.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, Depamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*. [Ahead of print]

Johnson DA, Thomas MA. 2007. The monosaccharide transporter gene family in *Arabidopsis* and rice: A history of duplications, adaptive Evolution, andfunctional divergence. *Mol Biol Evol*. 24: 2412-2423.

Johnson-Brousseau SA, McCormick S. 2004. A compendium of methods useful for characterizing *Arabidopsis* pollen mutants and gametophytically-expressed genes. *Plant J*. 39: 761-775.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite

the initial rate increase. *BMC Evol Biol*. 4: 22.

Kapralov MV, Filatov DA. 2007. Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol Biol*. 7: 73.

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

Keane TM, Naughton TJ, McInerney JO. 2007. MultiPhyl: A high-throughput phylogenomics webserver using distributed computing. *Nucleic Acids Res*. 35: W33-W37.

Kim TW, Guan S, Sun Y, Deng Z, Tang W, Shang JX, Sun Y, Burlingame AL, Wang ZY. 2009. Brassinosteroid signal transduction from cell-surface receptor kinases to nuclear transcription factors. *Nat Cell Biol*. 11: 1254-1260.

Kimura. M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kinoshita T, Miura A, Choi Y, Kinoshita Y, Cao X, Jacobsen SE, Fischer RL, Kakutani T. 2004. One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. *Science*. 303: 521-523.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*. 7: 474.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol*. 3: research0008.

Konopka CA, Backues SK, Bednarek SY. 2008. Dynamics of *Arabidopsis* dynamin-related protein 1C and a clathrin light chain at the plasma membrane. *Plant Cell*. 20: 1363-1380.

Kumar R, Drouaud J, Raynal M, Small I. 1995. Characterization of the nuclear gene encoding chloroplast ribosomal protein S13 from *Arabidopsis thaliana*. *Curr Genet*. 28: 346-352.

Kurihara-Yonemoto S, Handa H. 2001. Low temperature affects the processing pattern and RNA editing status of the mitochondrial cox2 transcripts in wheat. *Curr Genet*. 40: 203-208.

Lawton-Rauh A. 2003. Evolutionary dynamics of duplicated genes in plants. *Mol Phylogenet Evol*. 29: 396-409.

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombauts S. 2002. PlantCARE, a database of plant cisacting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*. 30: 325-327.

Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet*. 21: 602-607.

Liu SL, Adams KL. 2008. Molecular adaptation and expression evolution following duplication of genes for organellar ribosomal protein S13 in rosids. *BMC Evol Biol*. 8: 25.

Liu SL, Adams KL. 2010. Dramatic change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the Brassicaceae. *Mol Biol Evol*. 27: 2817-2828.

Liu Z, Adams KL. 2007. Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. *Curr Biol*. 17: 1669-1674.

Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability. *Curr Biol*. 5: 87-93.

Lu SX, Hrabak EM. 2002. An *Arabidopsis* calcium-dependent protein kinase is associated with the endoplasmic reticulum. *Plant Physiol*. 128: 1008-1021.

Lupski JR. 2007. Genomic rearrangements and sporadic disease. *Nat Genet*. 39: S43-S47.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science*.

290: 1151-1155.

Lynch M, Force AG. 2000a. The origin of interspecific genomic incompatibility via gene duplication. *Am Nat*. 156: 590-605.

Lynch M, Force AG. 2000b. The probability of duplicate gene preservation by subfunctionalization. *Genetics*. 154: 459-473.

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol*. 148: 1772-1781.

MacCarthy T, Bergman A. 2007. The limits of subfunctionalization. *BMC Evol Biol*. 7: 213.

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

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. 2007. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res*. 35: D237-D240.

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

Matsuno M., Compagnon V., Schoch G.A., Schmitt M., Debayle D., Bassard J.-E., Pollet B., Hehn A., Heintz D., Ullmann P., Lapierre C., Bernier F., Ehlting J., Werck-Reichhart D. 2009. Evolution of a novel phenolic pathway for pollen development. *Science*. 325: 1688-1692.

Mena M., Ambrose B.A., Meeley R.B., Briggs S.P., Yanofsky M.F., Schmidt R.J. 1996. Diversification of C function activity in maize flower development. *Science.* 274: 1537-1540.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Pérez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*. 452: 991-996.

Mitra K, Schaffitzel C, Fabiola F, Chapman MS, Ban N, Frank J. 2006. Elongation arrest by SecM via a cascade of ribosomal RNA rearrangements. *Mol Cell*. 22: 533-543.

Mizuta Y., Harushima Y., Kurata N. 2010. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. *Proc Natl Acad Sci USA*. 107: 20417-20422.

Mollier P, Holffmann B, Debast C, Small I. 2002. The gene encoding Arabidopsis thaliana mitochondrial ribosomal protein S13 is a recent duplication of the gene encoding plastid S13. *Curr Genet.* 40: 405-409.

Monson RK. 2003. Gene duplication, neofunctionalization, and the evolution of C^4 photosynthesis. *Int J Plant Sci*. 164: S43-S54.

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. *Proc Natl Acad Sci USA*. 100: 15682-15687.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol*. 8: 122-128.

Moutinho A, Trewavas AJ, Malhó R. 1998. Relocation of a Ca^{2+}-dependent protein kinase

activity during pollen tube reorientation. *Plant Cell*. 10: 1499-1509.

Nakamura M, Katsumata H, Abe M, Yabe N, Komeda Y, Yamamoto KT, Takahashi T. 2006. Characterization of the class IV homeodomain-Leucine Zipper gene family in *Arabidopsis*. *Plant Physiol*. 141: 1363-1375.

Nakamura T, Yamaguchi Y, Sano H. 2000. Plant mercaptopyruvate sulfurtransferases: molecular cloning, subcellular localization and enzymatic activities. *Eur J Biochem*. 267: 5621-5630.

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EH, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ, Laing WA, McArtney S, Nain B, Ross GS, Snowden KC, Souleyre EJ, Walton EF, Yauk YK. 2006. Analyses of expressed sequence tags from apple. *Plant Physiol*. 141: 147-166.

Oakley TH, Ostman B, Wilson AC. 2006. Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. *Proc Natl Acad Sci USA*. 103: 11637-11641.

Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer-Verlag.

Ohta T. 1987. Simulating evolution by gene duplication. *Genetics.* 115: 207-213.

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

Pagel M, Meade A. 2009. BayesTraits, version 1.0. Distributed by the author. School of Biological Sciences, University of Reading, Reading, UK.

Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* 48: 612-622.

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

Peeters NM, Hanson MR. 2002. Transcript abundance supercedes editing efficiency as a factor

in developmental variation of chloroplast gene expression. *RNA*. 8: 497-511.

Pin PA, Benlloch R, Bonnet D, Wremerth-Weich E, Kraft T, Gielen JJ, Nilsson O. 2010. An antagonistic pair of *FT* homologs mediates the control of flowering time in sugar beet. *Science.* 330: 1397-1400.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*. 21: 3718-3731.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol.* 5: 28.

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

Sandoval P, Leon G, Gomez I, Carmona R, Figueroa P, Holuigue L, Araya A, Jordana X. 2004. Transfer of RPS14 and RPL5 from the mitochondrion to the nucleus in grasses. *Gene.* 324: 139-147.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*. 440: 341-345.

Schein M, Yang Z, Mitchell-Olds T, Schmid KJ. 2004. Rapid evolution of a pollen-specific oleosin-like gene family from *Arabidopsis thaliana* and closely related species. *Mol Biol Evol.* 21: 659-669.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of Arabidopsis thaliana development. *Nat Genet.* 37: 501-506.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu

D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. 2010. Genome sequence of the palaeopolyploid soybean. *Nature.* 463: 178-183.

Schwede T, Kopp J, Gues N, M.C.P. 2003 SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*31: 3381-3385.

Sémon M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proc Natl Acad Sci USA.* 105: 8333-8888.

Sémon M., Wolfe KH. 2007. Consequences of genome duplication. *Curr Opin Genet Dev.* 17: 505-512.

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

Seoighe C. 2003. Turning the clock back on ancient genome duplication. *Curr Opin Genet Dev.* 13: 636-643.

Shi T, Huang H, Barker MS. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Ann Bot.* 106: 497-504.

Shikanai T. 2006. RNA editing in plant organelles: machinery, physiological function and evolution. *Cell Mol Life Sci*. 63: 698-708.

Shiu S-H, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li W-H. 2004. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell*. 16:1220-1234.

Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*. 4: 1581-1590.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot.* 96: 336-348.

Soltis DE, Soltis PE, Endress PK, Chase MW. 2005 Phylogeny and Evolution of Angiosperms. Sunderland, Sinauer Associates, Inc., USA. p. 370.

Spillane C, Schmid KJ, Laoueillé-Duprat S, Pien S, Escobar-Restrepo JM, Baroux C, Gagliardini V, Page DR, Wolfe KH, Gossniklaus U. 2007. Positive Darwinian selection at the imprinted *MEDEA* locus in plants. *Nature.* 448: 349-352.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22: 2688-2690.

Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y. 2005. EST data suggest that poplar is an ancient polyploidy. *New Phytol.* 167: 165-170.

Storey J., Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA.* 100: 9440-9445.

Sun HZ, Ge S. 2010. Molecular evolution of the duplicated *TFIIAγ* genes in Oryzeae and its relatives. *BMC Evol Biol.* 10: 128.

Sun X, Cao Y, Wang S. 2006. Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice. *Plant Physiol.* 140: 998-1008.

Swanson WJ, Nielson R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization protein. *Mol Biol Evol.* 20: 18-20.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet.* 3: 137-144.

Swingley WD, Chen M, Cheung PC, Conrad AL, Dejesa LC, Hao J, Honchak BM, Karbach LE,

Kurdoglu A, Lahiri S, Mastrian SD, Miyashita H, Page L, Ramakrishna P, Satoh S, Sattley WM, Shimada Y, Taylor HL, Tomo T, Tsuchiya T, Wang ZT, Raymond J, Mimuro M, Blankenship RE, Touchman JW. 2008. Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci USA*. 105: 2005-2010.

Szklarczyk R, Huynen MA. 2009. Expansion of the human mitochondrial proteome by intra- and inter-compartmental protein duplication. *Genome Biol*. 10: R135.

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

Tang W, Kim TW, Oses-Prieto JA, Sun Y, Deng Z, Zhu S, Wang R, Burlingame AL, Wang ZY. 2008. BSKs mediate signal transduction from the receptor kinase *BRI1* in *Arabidopsis*. *Science.* 321: 557-560.

Taylor JS, Van de Peer Y, Meyer A. 2001. Genome duplication, divergent resolution and speciation. *Trends Genet.* 17: 299-301.

Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plant. *Genome Res.* 16: 1017-1030.

Throude M., Bolot S., Bosio M., Pont C., Sarda X., Quraishi U.M., Bourgis F., Lessard P., Rogowsky P., Ghesquiere A., Murigneux A., Charmet G., Perez P., Salse J. 2008. Structure and expression analysis of rice paleo duplications. *Nucleic Acids Res.* 37: 1248-1259.

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

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

Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D, Meinke D. 2004. Identification of genes required for embryo development in

Arabidopsis. *Plant Physiol*. 135: 1206-1220.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 10: 725-732.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays.* 24: 175-184.

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

Wagner A. 2002. Selection after gene duplication: a view from the genome. *Genome Biol*. 3: 1012.1-1012.3.

Walsh J.B. 1995. How often do duplicated genes evolve new functions? *Genetics*. 139: 421-428.

Wang D, Tyson MD, Jackson SS, Yadegari R. 2006. Partially redundant functions of two *SET*-domain polycomb-group proteins in controlling initiation of seed development in *Arabidopsis*. *Proc Natl Acad Sci USA.* 103: 13244-13249.

Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *Am Nat.* 137: 515-526.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet.* 2: 333-341.

Wood T.E., Takebayashi N., Barker M.S., Mayrose I., Philip B. Greenspoon, P.B., Rieseberg L.H. 2009. The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci USA* 106: 13875-13879.

Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F. 2004. A model based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc.* 99: 909-917.

Yamagata Y, Yamamoto E, Aya K, Win KT, Doi K, Sobrizai, Ito T., Kanamori H, Wu J,

Matsumoto T, Matsuoka M, Ashikari M, Yoshimura A. 2010. Mitochondrial gene in the nuclear genome induces reproductive barrier in rice. *Proc Natl Acad Sci USA* 107: 1494-1499.

Yang Z, Nielson R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics.* 155: 431-449.

Yang Z, Nielson R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol.* 19: 908-917.

Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site clesses. *Mol Biol Evol.* 19: 49-57.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107-1118.

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

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24: 1586-1591.

Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM. 1990. The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. *Nature* 346: 35-39.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF. 2001. Crystal structure of the ribosome at 5.5A resolusion. *Science.* 292: 883-896.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA.* 95: 3708-3713.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18: 292-298.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genetics* 5: e1000581.

Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Austin (TX): The University of Texas at Austin.