FATES OF DUPLICATED GENES: SUB-LOCALIZATION, INTRACELLULAR GENE TRANSFER, AND CONCERTED DIVERGENCE

by

Yichun Qiu

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Fates of duplicated genes: sub-localization, intracellular gene transfer, and concerted divergence

submitted by Yichun Qiu in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Botany

Examining Committee:

Keith Adams
Supervisor

Sean Graham
Supervisory Committee Member

Quentin Cronk
Supervisory Committee Member

Patrick Keeling
University Examiner

Kota Mizumoto
University Examiner

Abstract

Gene duplication is a major contributor to genome evolution. There are several evolutionary fates of duplicated genes, such as retention with redundancy, subfunctionalization and neofunctionalization. I proposed and characterized examples of three new models here. The first is duplication of an alternatively spliced gene with dual-targeted products, followed by partitioning of the splice forms between the duplicates so that the products of each duplicate are sub-localized. I report the plastid ascorbate peroxidase (cpAPX) genes as an example of sub-localization. I show angiosperms typically have one cpAPX gene that generates both thylakoidal tAPX and stromal sAPX through alternative splicing. I then identified several independent, lineage-specific sub-localization events with paralogs of specialized tAPX and sAPX. I determined that the sub-localization happened through two types of sequence evolution patterns. Second, I show an unreported type of duplicative intracellular transfer: transfer of a nuclear gene to the mitochondrial genome and transcription of the gene. The transcribed orf164 gene in the mitochondrial genome of several Brassicaceae species is derived from a nuclear gene that codes for an auxin responsive protein. Third, I studied POLYCOMB REPRESSIVE COMPLEX2 (PRC2) in Brassicaceae to demonstrate concerted divergence of simultaneously duplicated genes whose products function in the same complex. The VERNALIZATION (VRN)-PRC2 complex contains VRN2 and SWINGER (SWN), and both genes were duplicated during a whole-genome duplication to generate FERTILIZATION INDEPENDENT SEED2 (FIS2) and MEDEA (MEA), which function in the Brassicaceae-specific FIS-PRC2 complex that regulates reproductive development. I found that FIS2 and MEA have correlated reproductive-specific expression patterns that are derived from the broadly expressed VRN2 and SWN. In vegetative tissues of *Arabidopsis thaliana*, repressive methylation marks are enriched in FIS2 and MEA, whereas active marks are associated with their paralogs. I detected comparable accelerated amino acid substitution rates in FIS2 and MEA but not in their paralogs. These lines of evidence indicate that FIS2 and MEA have diverged in concert, resulting in functional divergence of the PRC2 complexes in Brassicaceae. Overall, the three projects provide new insights into the retention and divergence of duplicated genes.

Lay Summary

Gene duplication is an ongoing process during genome evolution and the most common mechanism for the origin of new genes. There are several evolutionary fates of duplicated genes, including pseudogenization, retention with redundancy, subfunctionalization and neofunctionalization. I proposed and characterized three new models. I determined that the plastid Ascorbate Peroxidase (APX) genes in angiosperms experienced sub-localization, where an ancestral gene generating both stromal APX and thylakoidal APX by alternative splicing was duplicated, and subsequent partitioning of the splice forms between the duplicates caused the products of each duplicate locate to complementary subcellar compartments. I reported orf164 in several Brassicaceae species as an example of duplicative intracellular transfer of a nuclear gene to the mitochondrial genome. I studied POLYCOMB REPRESSIVE COMPLEX2 (PRC2) in Brassicaceae to demonstrate concerted divergence of simultaneously duplicated genes whose products function in the same complex. The three models provide new insights into the divergence of duplicated genes.

Preface

Chapter 3 has been published: Qiu, Y., Filipenko, S.J., Darracq, A, and Adams, K.L. (2014) Expression of a transferred nuclear gene in a mitochondrial genome. Current Plant Biology. 1: 68-72. I conceived this study, did most of the experiments and analyses, and wrote most of the manuscript. SJF helped with RT-PCR in Arabidopsis and made the first draft of Figure 3.1. AD performed the YASS alignment described in the methods and materials section. KLA helped design the study, supervised SJF, and took part in the manuscript writing and editing.

Chapter 4 has been published. Qiu, Y., Liu, S.-L., and Adams, K.L. (2017) Concerted divergence after gene duplication in Polycomb Repressive Complexes. Plant Physiology. 174:1192-1204. I designed and performed most of the experiments and analyses and wrote the manuscript. SLL participated in analyzing the microarray data. KLA helped conceive the study, took part in the experimental design, and edited the manuscript.

Chapter 2 is in preparation for a submission. Qiu, Y., Tay, Y.V., Ruan, Y., and Adams, K.L. (in prep) Repeated partitioning of splice forms and subcellular localization between duplicated genes. I designed and performed most of the analyses and prepared a manuscript. YVT provided some initial data. RY performed some RT-PCR assays. KLA conceived the study, helped with experimental design, and edited the manuscript.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Dr. Keith Adams, who has been supportive and helpful throughout my graduate programs, providing excellent opportunities and supervision for me to pursue my research goals.

I thank my committee members, Dr. Quentin Cronk and Dr. Sean Graham, for all the constructive suggestion and support. I also thank Dr. Mary Berbee, Dr. Naomi Fast and Dr. Patrick Martone for their help during my PhD study. I would like to thank Dr. Shao-Lun Liu, Yii Van Tay, Dr. Yuan Ruan, Dr. Aude Darracq for collaboration in my projects. I thank Dr. David Tack, Dr. Sishuo Wang and Alex Hammel for all the great discussion. Special thanks are owed to my parents.

My projects are not possible to be completed without the plant materials provided by many people and institutions. They are Daniel Mosquin and Eric La Fountaine from the UBC Botanical Garden for *Turritis glabra*, *Erysimum pulchellum*, *Armoracia rusticana*; Jamie Fenneman from UBC for *Capsella bursa-pastoris*, *Cardamine flexuosa*, *Lemna minor*; Shao-Lun Liu at Tunghai University for *Manihot esculenta*; Holly Forbes and Clare Loughran from University of California Botanical Garden at Berkeley for *Amborella trichopoda*, *Camellia sinensis*, *Mimulus guttatus*, *Musa acuminate*, *Nelumbo nucifera*, *Phoenix dactylifera*, *Theobroma cacao*; Cynthia Sayre and Samantha Sivertz from VanDusen Botanical Garden for *Aquilegia coerulea*, *Fraxinus excelsior*; Frank McDonough from the Los Angeles County Arboretum & Botanic Garden for *Ananas comosus*; Yangning Lu, Spencer Barrett, Ramesh Arunkumar at University of Toronto for *Eichhornia paniculata*; Gwendolyn Griffiths and Mary O'Connor at UBC for *Zostera marina*; Shona Ellis and Adams Wilkinson at UBC for *Petunia axillaris*, *Ricinus communis*; Lexuan Gao, Kaichi Huang and Joon Seon Lee at UBC for *Helianthus annuus*, *Linum usitatissimum*; Wayne Goodey at UBC for *Picea abies*; Yulin Sun at UBC for *Brachypodium dictachyon*; Yanli Hu and Jianwen Lu from Shanghai Chenshan Botanical Gardens for *Elaeis guineensis*. I owe my best

gratitude to them for their generous support. I also thank Jim Leebens-Mack from University of Georgia for providing the sequences of *Asparagus officinalis*.

My research was funded by a grant from the Natural Science and Engineering Research Council of Canada to Dr. Keith Adams. I was also supported by an NSERC CGS D fellowship and a UBC four-year fellowship.

献给陈德琴女士。

1 Introduction

The origin of new genes is a major process in genome evolution. The most common mechanism by which new genes originate is by gene duplication. Ongoing gene duplication events in most eukaryotes provide a large amount of genetic material for gene evolution and divergence. Gene duplication has been widely studied, and plenty of evidence has shown its major contribution to the evolutionary history of modern genomes.

1.1 Types of gene duplication

There are many types of gene duplication events giving rise to a large amount of ancient and recent duplicated genes in modern genomes. A gene family is a group of genes formed by rounds of gene duplication events. Considering the large total number of gene families, and the fact that many gene families comprise multiple gene copies, it is well recognized that the majority of genes have a detectible duplication history, especially in plants where it is thought to be a common phenomenon (Flagel and Wendel, 2009; Panchy et al., 2016). Below I will briefly discuss different types of gene duplications.

1.1.1 Whole genome duplication

Whole genome duplication, also referred to as polyploidy, is the largest scale of duplication events. Many common crop plants are recent polyploids (Chen, 2007). There are triploid

species such as banana and seedless watermelon, tetraploid species such as cotton and coffee, hexaploid species such as wheat, and octoploids such as strawberry. Based on the origin of the two merged genomes in the new polyploid genomes, polyploids can be categorized as autopolyploid when the merged subgenomes are from the same species and basically generate a multiplied genome, or allopolyploid when the merged subgenomes are from different species, likely as a result of the hybridization of related species, making up a heterogenous genome (Prince and Pickett, 2002; Taylor and Raes, 2004; Van de Peer et al., 2009).

After polyploidization, genomes typically undergo diploidization over time (Van de Peer et al., 2009). As a result, most current angiosperm diploid species have one or more rounds of whole genome duplication in their evolutionary history, often referred to as paleopolypoidy events. For example, *Arabidopsis thaliana* has experienced at least five rounds of WGDs in its ancestry during seed plant evolution: the zeta WGD at the base of seed plants, the epsilon WGD in the ancestral angiosperm lineage, the gamma WGD shared by all eudicots, the beta WGD shared by some Brassicales families including the sister family Cleomaceae, and the Brassicaceae-specific alpha WGD after the divergence of the family from Cleomaceae (Schranz and Mitchell-Olds, 2006; Jiao et al., 2011; Li et al., 2015). *Brassica* and *Camelina*, also in Brassicaceae, both have genus-specific genome triplications. WGDs have occurred in many other angiosperm lineages, such as those specific to Fabaceae (Bertioli et al., 2009; Schmutz et al., 2010), Salicaceae (Tuskan et al., 2007), Solanaceae (Tomato Genome Consortium, 2012), Asteraceae (Barker et al., 2016), and Poaceae (Paterson et al., 2004). Although plants in these families returned to diploid status after the WGDs, there are many features in their genomes, such as large blocks of

syntenic genes of the same age, that are remnants of their polyploid history. Rounds of WGD have provided a large number of paralogous genes to evolve and gain new functions or expression patterns. WGDs have been shown in association with many novel traits and may underlie several events of rate upshift in speciation or species diversification (Van de Peer et al., 2009; Schranz et al., 2012; Soltis et al., 2015).

1.1.2 Small scale duplication

Segmental duplication, where blocks of a chromosomal region are repeated, often tandemly, in the same chromosome or transferred into non-homologous chromosomes, is also commonly seen in plant genomes. This can be facilitated by the recombination of repetitive sequences especially those located in transposable elements, as segmental duplications have a significant enrichment of repeats on the transition boundaries (She et al., 2008).

Another kind of small scale gene duplication commonly seen is tandem duplication. Tandem duplicates are usually generated by unequal crossing over during recombination or by replication slippage during DNA replication (Achaz et al., 2000). Tandem duplicates are adjacent along a chromosome, sometimes separated by a few neighbouring genes as the duplicative unit may not concisely correspond to an intact gene. The exact duplicated chromosomal region might not be an entire gene or could span several genes. Tandem duplicates are sometimes found in large clusters as the sequence similarity between tandem duplicates facilitates

additional rounds of non-homologous crossing over (Achaz et al., 2000; Flagel and Wendel, 2009).

Some duplicated genes are dispersed in genomes. These may be initially in syntenic duplicative blocks, or local duplicative arrays, but subsequently become dispersed due to genome rearrangements. They can also be generated by retroposition or by duplicative transposition (Panchy et al., 2016). Retroposition occurs when a mRNA is reverse transcribed and the cDNA is integrated into a random chromosomal region or recombined with homologous sequences. As the gene duplication and reinsertion is mediated by mature mRNA, the newly formed gene is typically intron-less and resembles the entire parental gene coding sequences. For example, in the *Arabidopsis thaliana* genome, about 250 retrogenes have been detected through a genome wide survey, many of which have no introns, and a few have up to two introns which are formed after retroposition and located at new exonic regions (Abdelsamad and Pecinka, 2014). Duplicative transposition is another mechanism by which dispersed duplicates are formed (Panchy et al., 2016). Transposable elements typically contain the machinery for DNA replication and integration, and genes near a transposon or between two transposons of the same type can be captured and move along with the transposons (Freeling et al., 2008).

1.2 Fates of duplicated genes

1.2.1 Loss and pseudogenization

After formation by gene duplication, the duplicated genes can have different evolutionary fates. One copy is eventually lost in many cases, and in others, there could be a non-functioning copy retained in the genome as a pseudogene (Xiao et al., 2016). Thus, the once duplicated gene pair returns to single copy status. Because one ancestral gene might be adequate for protein function, an extra copy would be non-essential, and gene loss or pseudogenization can occur. There also could be selection for a return to single copy status. The dominant negative hypothesis suggests that the duplicated genes provide an extra target for the accumulation of deleterious mutations, and the malfunctioning copy may negatively interfere with the function of the normal copy by blocking proper protein interaction (Veitia, 2010).

1.2.2 Retention and gene balance

There are also many duplicated genes that are retained as pairs. There are several factors underlying the retention of both paralogs. Some duplicates may retain similar functions and compensate for each other in a single gene mutant. This type of redundancy or functional conservation is common for newly duplicated genes, but also may be retained during longer-term genome evolution, though less likely due to mutations that could either cause the knock-out of deleterious loci or the gradual divergence between duplicates (Li et al., 2016). There are

observations that genes of different functions can vary in their likelihood to be retained as duplicates, and paralogs that arise from different types of gene duplication events show biases in functional categories (De Smet et al., 2013; Li et al., 2016). The effect of gene dosage balance could explain this pattern (Birchler et al., 2005). Some gene products have many interactions, including physical protein-protein interaction and the formation of a regulatory hierarchy or network, and the stoichiometric balance between the interacting genes is important to maintain proper gene function (Birchler et al., 2005). According to this model, duplicate genes that are dosage sensitive tend to be retained after WGD since a single gene duplication may lead to imbalance in protein interactions, and thus the connected genes formed simultaneously through WGD tend to be retained together. In contrast, the retained genes derived from small scale duplications tend to have fewer connections (Freeling, 2009; Tasdighian et al., 2017).

1.2.3 Functional divergence through neofunctionalization and subfunctionalization

Duplicated genes may also diverge in functions. After gene duplication, an extra copy may be free from constraints and evolve a new function while the conserved paralog maintains the ancestral function. This is referred to as neofunctionalization, which plays an important role in the rise of key innovations (Panchy et al., 2016). Duplicates can instead diverge through subfunctionalization, where the paralogs each take on a proportion of ancestral functions reciprocally, and complement each other to play the full role of the ancestral gene (Panchy et al., 2016). This partitioning of ancestral function could be asymmetric, so that the two paralogs may experience different selective pressures (Panchy et al., 2016). Because the

subfunctionalized paralogs have non-overlapping functions, the deletion of either may be deleterious, and the retention of both copies would then be necessary (Zhang, 2003).

1.3 Additional models for the fates of duplicated genes

There are many theories or models describing the fates of duplicated genes (Flagel and Wendel, 2009; Panchy et al., 2016), some of which are discussed below.

1.3.1 Regulatory divergence

It has been well established that paralogs can diverge in expression patterns (Blanc and Wolfe, 2004; Liu et al., 2011). It is possible that one or both of the paralogs gains a novel expression pattern compared with the ancestral tissues and/or developmental stage spectrum. This phenomenon is referred to as regulatory neofunctionalization (Liu et al., 2011). Accordingly, regulatory subfunctionalization describes a scenario where each paralog reciprocally loses parts of the ancestral expression pattern, and two paralogs together make up the full ancestral expression profile (Liu et al., 2011). Many studies have shown regulatory neofunctionalization and subfunctionalization of anciently duplicated genes (Blanc and Wolfe, 2004; Haberer et al., 2004; Casneuf et al., 2006; Ganko et al., 2007; Liu et al., 2011), and the divergence of cis-regulatory elements between duplicated genes can lead to changes in regulatory networks (Arsovski et al., 2016).

When tandem duplicates are formed, it is possible that the entire regulatory region is not completely duplicated, which can lead to considerable divergence in expression profile (Liu et al., 2011). It is also possible that tandem duplicates in a cluster can be expressed in a coordinated manner when the original shared regulatory elements are properly retained (Schmid et al., 2005). Out of all types of duplicated genes, retrogenes are hypothesized to show most divergent expression pattern compared to the parental genes (Abdelsamad and Pecinka, 2014), because the mRNA-mediated process does not copy the regulatory sequences of parental genes. Thus, usually retrogenes that are expressed have *de novo* cis elements or recruit nearby existing ones, which have little or no similarity with those in the parental duplicate. The duplicates from large syntenic blocks formed through segmental duplication or WGD can also have different level of co-expression or expression divergence (Liang and Schnable, 2018).

The expression pattern divergence is not only reflected by the spatial and/or temporal absence versus presence, but also at the quantitative level (Yoo et al., 2014). In autopolyploids, the expression patterns of paralogs are expected to be identical upon duplication, and subsequent divergence could happen (Garsmeur et al., 2014). Allopolyploids will usually show deviation from a hypothetical parental additivity, because hybridization is usually accompanied by extensive changes to patterns of parental gene expression due to the heterogeneity of the two merged subgenomes, a phenomenon described as transcriptome shock (Buggs et al., 2011; Hegarty et al., 2006). This nonadditive expression has two common features. One is about the total gene expression level. The total gene expression (expression of both copies) in a polyploid

sometimes is closer to that of one of its parents, which is regarded to have expression-level dominance, or the level can be even lower or higher than those of both parents, referred to as transgressive expression (Yoo et al., 2014). The other feature concerns homeolog expression shifts and bias: here, the relative abundance of a homeolog can be upregulated or downregulated compared to that in parental genomes, and the expression levels can be unequal between homeologs (Yoo et al., 2014). Because usually a more highly expressed gene experiences a higher level of selection pressure (Yang and Gaut, 2011), the divergence in regulation could have effects on the chances for retention, sequence divergence, and potentially subsequent functional changes. Sequence changes in non-coding regions can also affect the divergence of expression patterns. In addition, epigenetic factors are an important component in expression regulation (Chen, 2007). The gain and loss of a wide variation of epigenetic modifications, such as cytosine methylation and histone acetylation and methylation, could help define epigenetic conservation, neofunctionalization and subfunctionalization.

1.3.2 Duplication-Degenerative-Complementation

The duplication-degenerative-complementation (DDC) model is one the early proposed forms of subfunctionalization, with an emphasis on expression regulation (Force et al., 1999). In this model, duplicated genes each gradually accumulate different mutations in the cis-regulatory regions such that one paralog only retains a part of the ancestral expression domains, and the other evolves to complement and together they cover the full ancestral expression profile. The

DDC model could also be applied directly to protein function. After duplication of a multi-functional protein that contains different domains performing the difference functions, either paralog could have certain functional domains degenerate subsequently and loose the corresponding function. Thus, those paralogs may become specialized for fewer functions and the ancestral functions are maintained and allocated (Zhang, 2003).

1.3.3 Escape from adaptive conflict

The subfunctionalization of a multi-functional protein can be beneficial and thus selected for. It has been proposed that a dual-function protein can experience adaptive conflict, such that each function places constraints on the other, as any improvement of one function, typically associated with sequence or structure changes, may not happen without detrimentally affecting the other function (Hittinger and Carrol, 2007; Des Marais and Rausher, 2008). This dilemma can potentially be resolved after gene duplication under the escape from adaptive conflict (EAC) model (Hittinger and Carrol, 2007; Des Marais and Rausher, 2008). This model, when first proposed, described a scenario where a new function acquired by an existing gene eventually reduces the performance of the original function. The gain of a new function is harmful to the improvement of the original function, and the original function limits the potential for a better new function. This would apply to an ancestral multi-functional protein, which is similar to the “intermediate status” in the original hypothesis, where a new function is gained and the original function is retained. After gene duplication, one copy can be free to improve one function, even accompanying abolishment of the other functions, as long as the

other copy performs the original functions, while in return the other copy can improve the other function without constraints. When EAC happens, both paralogs experience adaptive changes and both ancestral functions are improved. This differs from the DDC model at the selection level: the degeneration of mutually exclusive functions is a neutral process, and thus there might not be elevation of protein performance for it (Flagel and Wendel, 2009).

1.3.4 Subneofunctionalization

Fitting in the definition of the general model of neofunctionalization, there could be least three scenarios for a neo-functionalized gene: while gaining a novel function, it may still perform the ancestral function either fully or partially, or it may lose the ancestral function completely (He and Zhang, 2005). For the first two scenarios, it is possible that both paralogs experience neofunctionalization as long as the ancestral essential functions are covered; and it is possible that the ancestral function is abolished if no longer necessary for survival (He and Zhang, 2005; Zhang, 2003). As with cases fitting in the general model of subfunctionalization, it is possible that the paralogs have shared ancestral functions and are thus redundant in some respects, or their partitioned functions become mutually exclusive (Force et al., 1999; He and Zhang, 2005; Zhang, 2003). The current observation of divergence pattern is not necessarily a one-step change, and it is possible that the two models interplay at different evolutionary stages, which is more complicated than simple explanations by neofunctionalization or subfunctionalization alone. This prompted a new model termed subneofunctionalization (He and Zhang, 2005), which suggests duplicated genes are more likely to be retained and diverging through rapid

subfunctionalization, probably through DDC as mentioned above, so that the two paralogs become non-redundant, and they continue to evolve through subsequent and substantial neofunctionalization or additional subfunctionalization (He and Zhang, 2005).

1.3.5 Coordinated functional divergence of genes

As mentioned above, many genes perform their function in different types of pathways or networks, such as regulation, metabolism and protein-protein interactions. They may give rise to divergent pathways or networks. For example, a change in the expression pattern of one gene could drive parallel changes of genetically or physically interacting genes in order to maintain the integrity of a gene interaction network, and this could involve a set of genes that diverge concertedly (Blanc and Wolfe, 2004). Duplicates that arise from WGDs are more likely to have this type of concerted evolution, as WGDs duplicate the entire pathways or networks, and such duplicated pathways or networks could evolve novel functions through the coordinated neofunctionalization and/or subfunctionalization of gene sets (De Smet et al., 2017).

1.4 Thesis project objectives

There are three parts to my thesis, corresponding to three new models for fates of duplicated genes.

1.4.1 Sub-localization

A single gene can potentially generate multiple types of transcripts through alternative splicing of introns (Reddy et al., 2013). When such an alternatively spliced gene is duplicated, the paralogs can diverge by gain or loss of certain splicing patterns (Zhang et al., 2010; Tack et al., 2014). It is possible that the paralogs could reciprocally retain one type of spliced form, which would be analogous to regulatory subfunctionalization. Alternative splicing is one mechanism by which the final protein products of a single gene can be localized to two subcellular compartments, as the peptides from alternatively spliced transcripts may vary in localization signals (Yogev and Pines, 2011). Thus, after duplication of an alternatively spliced gene with dual-targeted products, one potential outcome is partitioning of the alternatively spliced forms between the duplicates, such that the gene product of each duplicate is localized to one of the ancestral subcellular locations. An objective of Chapter 2 is to describe a case supporting this type of paralogous divergence. I studied the plastid ascorbate peroxidase (*cpAPX*) genes across angiosperms as an example. All flowering plants have two kinds of cpAPX peptides in plastids, thylakoidal APX (tAPX) and stromal APX (sAPX). I studied the evolutionary history of these two types of plastid APXs in various flowering plants. This study integrated alternative splicing and subcellular targeting into the classical subfunctionalization model to characterize a specific fate of duplicated genes.

1.4.2 Duplicative intracelluar gene transfer

There are three sets of independent genomes in a plant cell, the nuclear genome, and two organellar genomes in mitochondria and plastids, respectively. Gene duplication is not limited within a genome, and duplicative transfers have been observed, resulting in paralogs in different genomes of a plant cell (Liu and Adams, 2008). Transfer of mitochondrial genes to the nucleus, and subsequent gain of regulatory elements for expression, is an ongoing evolutionary process in plants. Many examples have been characterized, which in some cases have revealed sources of mitochondrial targeting sequences and cis-regulatory elements (Adams and Palmer, 2003; Bonen and Calixte, 2006; Liu et al., 2009). In contrast, there have been no reports of a nuclear gene that has undergone intracellular transfer from the nuclear genome to the mitochondrial genome and become expressed. An objective in Chapter 3 was to characterize a nuclear gene that has been transferred to the mitochondrial genome and become expressed, providing a new perspective on the movement of genes between plant genomes. The gene of interest is the mitochondrial *ORF164* in *Arabidopsis thaliana* which is a duplicated copy of the nuclear gene *AUXIN RESPONSIVE FACTOR* 17 (*ARF17*).

1.4.3 Concerted divergence of protein complexes

Many pairs of duplicated genes showing various types of divergence have been described individually; however, there are many protein complexes within the cell with members derived from different genes, and their evolutionary trajectory after the duplication of one or multiple

components has been understudied. Inspired by the idea of coordinated functional divergence in a network or pathway (Blanc and Wolfe, 2004), the potential co-evolution between the products of interacting genes was an intriguing possibility characterized in Chapter 4. An objective of Chapter 4 was to test the hypothesis that duplication of two genes whose products function together in a complex, followed by parallel evolution of each gene, can lead to divergence of the whole complex. I studied the POLYCOMB REPRESSIVE COMPLEX 2 (PRC2) in Brassicaceae to characterize the divergence pattern of protein complexes. I gained multiple lines of evidence to support this model of concerted divergence.

2 Repeated partitioning of splice forms and subcellular localization between duplicated genes

2.1 Introduction

During the evolutionary history of flowering plants there have been one or more rounds of polyploidy events in all lineages (Jiao et al., 2011; Li et al., 2015). These whole genome duplication events, along with other types of gene duplication events, including tandem duplication, segmental or chromosomal duplication, and retroposition, have supplied raw material for evolutionary innovations in morphology, biochemistry, and other phenotypic characters (reviewed in Zhang, 2003; Flagel and Wendel, 2009; Panchy et al., 2016). Duplicated genes have different fates over time. In some circumstances, one copy of the newly duplicated genes might lose its function or expression to become a pseudogene or be deleted from the genome. Among the retained duplicates, both copies may retain their original function, possibly to maintain the balance of dosage (Birchler et al., 2005; Panchy et al., 2016). Some pairs diverge in expression patterns and/or functions: either one copy undergoes neofunctionalization thus gaining a new function, or the pair undergo subfunctionalization where the original function is divided and reciprocally inherited by either of the duplicates (Zhang, 2003; Panchy et al., 2016). The potential functional or regulatory divergence could happen at multiple levels, pre- and post- transcription and translation, such as cytosine methylation and histone modifications, expression pattern, protein-protein interaction (such as dimerization) and enzymatic activity (Flagel and Wendel, 2009; Panchy et al., 2016).

Gene duplication can contribute to changes in subcellular localization of paralogous gene products, which is also known as protein subcellular relocalization (Byun-McKay and Geeta, 2007). Protein subcellular localization can be caused by a specific N-terminal peptide, C-terminal motifs, or the presence of internal localization signals (Byun-McKay and Geeta, 2007). Thus, protein subcellular relocalization after gene duplication could happen as a result of an initial imperfect duplication inclusive or exclusive of parts of the genes, or subsequent sequence divergence that leads to the gain or loss of these targeting sequences, and the peptides could remain in the cytosol or be targeted to a membrane-bound organelle like the endoplasmic reticulum, the nucleus, the peroxisome, the mitochondria, or the plastids (Byun-McKay and Geeta, 2007).

There are several known cases of a single gene that produces dual-targeted peptides, some of which are caused by alternative splicing (Yogev and Pines, 2011). Alternative splicing is a process where multiple types of mature mRNAs are generated from a single type of transcript. There are several types of alternative splicing, such as intron retention where a complete intron remains in the mature mRNA, exon skipping where an exon is excluded from the mature mRNA, alternative donor which depends on the use of a proximal or distal 5' splice site, and alternative acceptor which depends on the usage of a proximal or distal 3' splice site (reviewed in Reddy, 2007; Barbazuk et al., 2008). Thus, there can be different types of mature RNAs generated from a single precursor, which in some cases can be translated into peptides harboring different sequences and domains. This can have an impact on the function or protein-protein interaction of variant peptides.

Because protein subcellular localization can be caused by a specific transit peptide, the absence or presence of these sequences could be an outcome of the spliced mature mRNA. Some genes have an alternatively spliced exon that codes for a transit peptide; thus, alternative splicing results in localization of the gene products to different subcellular or sub-organellar compartments.

Considering the possibility that duplicated gene could experience protein relocalization, and the possibility for duplicated genes to diverge in their alternative splicing patterns (Zhang et al., 2010; Tack et al., 2014), I propose that duplicates derived from an ancestral gene encoding alternatively spliced, dual targeted peptides could show partitioning of alternative splicing patterns between the duplicates, such that each copy codes for one splice form, and its product is localized to a single subcellular or sub-organellar compartment. As an example of this proposed fate of duplicated genes, I studied plastid ascorbate peroxidase (*APX*). In plants, APX has a role in the scavenging of H_2O_2 by using ascorbate as an electron donor (Asada, 1999). There are three types of APX: the cytosolic, microsomal, and plastid. The plastid APX (cpAPX) includes the stromal APX (sAPX) and thylakoid APX (tAPX). In some studied species such as *Cucurbita* cv. (pumpkin), *Spinacia oleracea* (spinach) and *Nicotiana tabacum* (tobacco), the stromal APX and thylakoid APX are produced by a single copy gene using alternative splicing (Mano et al., 1997). The last exon of the plastid APX gene is alternatively spliced by two mutual splicing acceptors. The distal acceptor corresponds to an exon consisting of the corresponding codons of the hydrophobic anchor region and the entire 3' untranslated region of tAPX mRNA, but the proximal acceptor is followed by an early stop codon and a potential polyadenylation

signal of the sAPX mRNA (Fig. 2.1; Mano et al., 1997). Thus, in the tAPX, the early stop codon is spliced out, and the product of tAPX contains the hydrophobic anchor region which causes the tAPX to localize in the thylakoid in plastids. In contrast, in the sAPX, the longer exon which introduced the early stop codon can only be translated without this region, then the sAPX is localized in the stroma in plastids. However, in *Arabidopsis thaliana*, the two APX isoforms targeted to different sub-organellar locations are encoded by two genes instead of a single gene undergoing alternative splicing (Ishikawa and Shigeoka, 2008). The two genes have expression divergence and contribute at different levels in different tissues or in response to different stresses to the peroxidase activities (Kangasjarvi et al., 2008; Maruta et al., 2010).

The goals of this study were to identify gene duplication events and retention patterns among *cpAPX* in angiosperms, and to assay the transcripts of cpAPX genes to determine the splicing forms, to find cases of specialized *tAPX* and *sAPX* genes. The results provide evidence for multiple cases of gene duplication followed by partitioning of alternative splice forms between duplicates, and thus specialization of subcellular localization.

2.2 Methods and materials

2.2.1 Identification of plastid *APX* genes and phylogenetic analysis

Plastid *APX* genes were found from multiple species using different databases (Table 2.1) by reciprocal best BLAST hits starting with *cpAPXs* in *Spinacia oleracea* and *Arabidopsis thaliana*. I looked for cpAPX genes in more than 80 species with a sequenced genome, representing 45 families across flowering plants as well as some gymnosperm clades. The genomic sequences and coding sequences were both obtained to identify the presence or absence of alternative splicing variants, and to compare splicing forms between related species. The corresponding translated peptide sequences are suggestive of the sub-organellar location defining the thylakoid-bound APX and stromal APX, by the presence or absence of the conserved hydrophobic C-terminal peptides. In addition, I detected the presence of the N-terminal presequences that code for a plastid transit peptide (cTP) using TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/) to predict the plastid targeting of these cpAPX peptides (Emanuelsson et al., 2000).

Amino acid alignments were generated by MUSCLE, and reverse transcribed into codon alignments using the customized Perl script (Edgar, 2004). The N-terminal signal and C-terminal signal peptide regions were removed, and the remaining conserved enzymatic domains were used for phylogenetic analyses. The phylogenetic tree was generated in RAxML (version 7.2.6) using GTRGAMMA as the substitution model with 100 bootstrap replicates to determine the

support for each branch (Stamatakis, 2006). The phylogeny was compared with the species tree to identify the gene duplication events specific to certain taxonomic groups.

2.2.2 RT-PCR to detect the transcripts from selected genes

For the taxa that have two or more copies of paralogous cpAPXs, I designed gene-specific primers targeting the 3' end of the putative transcripts to confirm the splicing pattern and thus identify the sub-organellar localization of the encoded peptides (Table 2.2). I also designed primers for several species with single copy cpAPX genes, to confirm the expected alternative splicing. Total leaf RNA was extracted from each species with the Ambion RNAqueous Kit following the manufacturer's protocol. DNaseI (New England Biolabs) treatment was applied to remove any residual genomic DNA. The cDNA was generated with reverse-transcriptase (M-MLV from Invitrogen), and used as template in PCR reactions. The PCR cycling program for amplification was 94° for 3 min; 30-35 cycles of 94° for 30s, 50°- 55° annealing dependent on primers' Tm values for 30s, 72° for 30-60s (approximately 30s for 1kb products); and a final elongation of 72° for 7 min. The PCR product was checked on 1% agarose gels and the specific bands were purified for sequencing to confirm the sequences and spliced sites.

2.2.3 Sequence rate evolution analysis using PAML

To determine the rates of sequence divergence of duplicated paralogs and compare to their dual-targeted orthologs, Ka/Ks ratios were estimated with the branch model using Codeml in

PAML (Yang, 2007). A free-ratio test was applied to estimate the Ka/Ks ratios along each lineage. To determine if one or both of the duplicate genes of interest evolved faster than single copy genes, a one-ratio model, a two-ratio model and a three-ratio model were used. The first model assumes all sequences have the same Ka/Ks ratio. The second model assumes that the branches of the two duplicate genes of interest have one Ka/Ks ratio, while the orthologs in other species have a different ratio, which implies a hypothesis that the two duplicate genes of interest evolved at similar rates, that are different from the pre-duplication single copy genes. The third model assumes that the branches of the two duplicate genes of interest have different Ka/Ks ratios and the orthologous branch has a third Ka/Ks ratio. A likelihood ratio test was performed, where twice the difference of likelihood values ($2\delta L$) was calculated and compared against a chi-square distribution with the degree of freedom (df) equal to df_2-df_1 (difference of the number of branch-wise Ka/Ks ratios in the two models) to determine whether or not duplicated gene sequence evolution is significantly different. I also applied a branch-site model to detect positively selected sites in duplicates (Zhang et al., 2005).

2.3 Results

2.3.1 Duplicated and single copy plastid *APX* genes across flowering plant species

I identified cpAPX genes in all species that I investigated. Most of the species have a single copy *cpAPX* gene in their genomes, even though these lineages experienced multiple rounds of whole genome duplication (Fig. 2.2). These single copy genes all have the sequence features

suggestive of alternative splicing that give rise to both tAPX and sAPX: there are the pair of conserved alternative spliced acceptor sites (most likely TGCAG), of which the stop codon closely follows the proximal acceptor, and the coding exon after the distal acceptor corresponding to a hydrophobic tail as the anchorage to the thylakoid membrane (Fig. 2.1). I predicted the subcellular localization of these single copy cpAPX genes by analysing the N-terminal peptides (see Methods), as the plastid localization is the first step in sub-organellar localization in the stroma or thylakoid. All of them are highly likely to be targeted to plastids, except that cpAPX from *Amborella trichopoda* shows a similar likelihood between plastid vs. mitochondrion targeting.

I constructed a *cpAPX* gene tree to identify duplicated genes (Fig. 2.2). The gene tree indicated that some species have duplicated genes that are both retained (Fig. 2.2). In several species, the pair of paralogous *cpAPXs* share more than 95% identity in coding sequences that translate to nearly identical peptides, and both have the sequence features suggestive of alternative splicing. *Glycine max* is one of those species, and others include the agricultural crops *Linum usitatissimum*, *Nicotiana tabacum*, *Coffea arabicum* and *Chenopodium quinoa*, all of which have an evolutionarily recent polyploidy event in their genus (Clarkson et al., 2005; Cenci et al., 2010; Wang et al., 2012; Jarvis et al., 2017).

In contrast, I also found many pairs of paralogs in multiple lineages that have structural and/or sequence divergence around the 3'-end exons and introns that affects the translated C-terminal peptides of the gene products. I also predicted the subcellular localization of these duplicates.

Most of the cpAPXs I obtained are highly likely to be targeted to plastids, except that some Arecaceae sequences show similar possibilities between plastid vs. mitochondrion targeting, and one clade of Poaceae genes are highly likely to be targeted to mitochondria. As the products of most of those duplicates are still localized to plastids, I investigated if the divergence at 3'-end of the genes correspond to the divergence of sub-organellar localization in the stroma or thylakoid.

These pairs of divergent duplicates are likely to have been generated independently by 18 lineage-specific duplication events, as seen in Fig. 2.2: 1. *Manihot* and *Hevea* (in Euphorbiaceae) represented by *M. esculenta* and *H. brasiliensis*, 2. Salicaceae represented by *Populus trichocarpa* and *Salix purpurea*, 3. *Gossypium* (in Malvaceae) represented by *G. raimondii* and G*. arboreum*, 4. Cleomaceae and Brassicaceae represented by *Tarenaya hassleriana*, *Aethionema arabicum*, *Brassica rapa*, and *Arabidopsis thaliana*, 5. *Anacardium occidentale* (in Anacardiaceae), 6. *Sesamum indicum* (in Pedaliaceae), 7. Oleaceae shared by *Olea europaea* and *Fraxinus excelsior*, 8. *Solanum* and *Capsicum* (in Solanaceae) represented by *S. lycopersicum*, *S. tuberosum* and *C. annuum*, 9. *Nymphaea colorata* (in Nymphaeaceae), 10. Poaceae represented by *Zea mays*, *Sorghum bicolor* and *Brachypodium distachyon*, 11. *Ananas comosus* (in Bromeliaceae), 12. Arecaceae including *Phoenix dactylifera* and *Elaeis guineensis*, 13. Musaceae including *Musa acuminate* and *Ensete ventricosum*, 14. *Eichhornia paniculata* (in Pontederiaceae), 15. *Xerophyta viscosa* (in Velloziaceae), 16. *Zostera marina* (in Zosteraceae), 17. Lemnoideae (duckweeds, in Araceae) represented by *Lenma minor* and *Spiroldela polyrhiza*, 18. *Nelumbo nucifera* (in Nelumbonaceae).

2.3.2 Changes in AS patterns of paralogous cpAPXs

I investigated the transcripts from 18 species that have paralogs with divergent 3'-end sequences to determine if one or both paralogs show alternative splicing to produce both sAPX and tAPX transcripts. I found that for many species, only one RT-PCR band was amplified for either *cpAPX* paralog (Fig. 2.3). After sequencing the RT-PCR products, I could determine if it was the sAPX or tAPX. In most cases, the *cpAPX* paralogs have specialized to code for either tAPX or sAPX. They are *Nymphaea colorata*, *Zostera marina*, *Lemna minor*, *Xerophyta viscosa*, *Elaeis guineensis*, *Eichhornia paniculata*, *Musa acuminata*, *Brachypodium distachyon*, *Nelumbo nucifera*, *Solanum lycopersicum*, *Fraxinus excelsior*, *Populus trichocarpa*, *Manihot esculenta*, *Gossypium arboreum*, and *Tarenaya hassleriana*. In contrast, I found that in *Sesamum indica*, *Ananas comosus* and *Phoenix dactylifera*, one of the duplicated genes has two PCR bands, suggestive of alternative splicing corresponding to two types of cpAPXs, whereas the other paralog has only one band and its sequence turned out to be an sAPX after sequencing (Fig. 2.3).

I then assayed the splicing pattern of the duplicated *cpAPXs* in *Glycine max* and *Linum usitatissimum* which have similar structures at the 3' end of the duplicated genes. Due to the high similarity between the paralogs, I could only design a single pair of primers to amplify cDNAs of both genes in either species. In both cases, I obtained two PCR bands of the predicted sizes (Fig. 2.3). By sequencing the products, I saw polymorphic sites that distinguish each duplicate, indicating that both genes give rise to sAPX and tAPX transcripts.

The single copy *cpAPX* genes in pumpkin, spinach and tobacco have been shown to give rise to both tAPX and sAPX transcripts by alternative splicing (Mano et al., 1997; Ishikawa et al., 1997; Yoshimura et al., 2002). In addition, all the single copy *cpAPX* genes in the seed plants that I found have similar sequence features. Thus, I hypothesized they are also alternatively spliced into two forms of cpAPXs. To confirm the alternative splicing pattern in the single copy genes, I chose 18 *cpAPX* sequences from a diverse group of angiosperms (plus two non-angiosperms), with a focus on species closely related to those with duplicated *cpAPX* genes. I performed RT-PCR on leaf RNA using primers to amplify the 3'-end of cpAPX transcripts. All the cDNAs of those single copy genes were amplified into two RT-PCR bands of the expected sizes, indicative of alternative splicing (Fig. 2.3). The sequences obtained by Sanger sequencing of the RT-PCR products indicated the splicing events at the predicted splice acceptor sites; thus, the two bands corresponded to sAPX and tAPX from the same gene. These results indicate that the alternatively spliced *cpAPXs* are broadly present across flowering plants, as well as being present in *Ginkgo* and *Picea*.

Because I chose the single copy *cpAPX* genes to study with a focus including the close relatives of those taxa with duplicated *cpAPX* genes, the closely related outgroup orthologous serve as evidence that the change in splicing pattern and *cpAPX* specialization took place after lineage-specific duplication. The protein products from the ancestral *cpAPX* gene in the most recent common ancestor of flowering plants once were likely dual-targeted to stroma and thylakoid in plastids, as well as many pre-duplicated ancestral genes. This alternative splicing pattern is inherited by those extant single copy genes. The products of the extant paralogs are localized

reciprocally to either sub-organellar compartment after gene duplication and retention, together making up the ancestral localization pattern. Here I refer to this type of paralogous diversification after duplication as sub-localization.

2.3.3 A shared pattern of tAPX specialization and two ways of sAPX specialization

By comparing the aligned genomic, transcript, and conceptually translated protein sequences, I found that among the paralogs that code for a single tAPX or sAPX, I could infer the sequence or structural changes that led to the specialization of tAPXs and sAPXs in each case. Below I present a few cases that were examined:

The duplication event shared by Brassicaceae and Cleomaceae in the order Brassicales gave rise to paralogs *cpAPXs*, which are still retained in *Tarenaya hassleriana* and several Brassicaceae species I surveyed. Compared to cpAPX sequences in *Carica papaya*, *Theobroma cacao*, *Citrus sinensis* and *Vitis vinifera*, the specialized tAPX has either an expansion (seen in *T. hassleriana,* and *Aethionema arabicum*) and/or shrinkage (other Brassicaceaes) of the last intron which has lost the sAPX associated proximal acceptor and subsequent stop codon. As a result, only one acceptor homologous to the tAPX associated distal acceptor is utilized and only one spliced product is generated coding for tAPX (Fig. 2.4). As for the specialized sAPX, a novel premature stop codon is present in the homologous penultimate exon, which truncates the peptide sequence to be even shorter than the orthologous sAPXs. Thus, these specialized sAPXs do not have the thylakoid membrane anchorage chain. There is no sequence similarity to orthologous

cpAPXs after the novel stop codon. It is likely that the novel point mutation gave rise to the novel stop codon, and the downstream sequences were relaxed from selection and became divergent and unrecognizable. With the loss of sequence similarity after the penultimate exon, the homologous last intron and the homologous last exon are missing. Thus, the ancestral alternative splicing was abolished. As a result, the two retained *cpAPX* paralogs each gave rise to one transcript, tAPX or sAPX.

In Solanaceae, I inferred that the duplication event is shared by the family through phylogenetic analysis (Fig. 2.5), then the duplicated genes were reciprocally lost in the ancestral lineages of *Nicotiana* and *Petunia*, but presumably retained as duplicates with AS patterns in an ancestral lineage giving rise to *Solanum* and *Capsicum*. The duplicates underwent sub-localization after the common ancestor of *Solanum* and *Capsicum* diverged from the *Nicotiana* and *Petunia* lineages. The specialized *tAPX*, sister to the *Nicotiana cpAPX*, has an insertion that interrupts the proximal acceptor in *Capsicum annuum*. The specialized *sAPX*, sister to the *Petunia cpAPX*, has point mutations (GAC to TAA) to generate a premature stop codon in the penultimate exon. Subsequently there are several indels in the exon after the stop codon causing a frame shift of the exon coding for the hydrophobic chain in *Solanum lycopersicum*, *Solanum tuberosum* and *C. annuum*. In addition, the sAPX in *S. tuberosum* has a mutated distal acceptor. This case demostrates that there can be a lag time between gene duplication and sub-localization.

Two cases of cpAPX duplication are present in Malpighiales. The whole genome duplication event in Euphorbiaceae, which predated the divergence of *Hevea* and *Manihot*, gave rise to

paralogs with separate *tAPX* and *sAPX* genes present in syntenic blocks. The specialized *tAPX* still has the homologous sAPX splicing acceptor with the stop codon, but based on our RT-PCR results, it was not utilized by the splicing factors; thus, only the homologous tAPX splicing acceptor remains functional and recognized by the splicing machinery. The specialized *sAPX* has a point mutation (TAC to TAA in *Hevea brasiliensis* and TAC to TAG in *Manihot esculenta*) in the penultimate exon to create the stop codon of *sAPX*; in addition, there are stop codons in the conceptual homologous chain region of the ORF in *M. esculenta* specialized *sAPX*, and the homologous distal acceptor is gone in *H. brasiliensis* due to loss of sequence similarity, as demonstrated by the presence of the novel stop codon that prevents the coding potential for tAPX. This *sAPX* specialization is similar as described in Brassicaceae and Solanaceae. In another lineage of the Malpighiales, Salicaceae, the scenario of a duplicated pair from the salicoid duplication event is different. The specialized *tAPX* in *Salix purpurea* has a mutation in the proximal acceptor (AG to TG) to cause the loss of function. As for sAPX in the Salicaceae, unlike the above mentioned specialized *sAPX* genes, it has the same length as orthologous sAPX that contains the full penultimate exon and is spliced at the homologous proximal acceptor. Instead of an interruption in the coding sequence, *sAPX* in Salicaceae lost the conserved distal acceptor; moreover, there is a deletion in the tandem repeats, $(AG)_4$ to $(AG)_3$, to cause a frame shift in *Populus trichocarpa*. To summarize, the two *cpAPX* paralogs in Salicaceae reciprocally retain either of the ancestral alternatively spliced acceptors, to generate specialized sAPX and tAPX. The two lineage-specific WGDs are well documented, and the difference in the patterns of cpAPX specialization suggests that these sub-localization events are independent.

In *Gossypium*, the specialized *tAPX* has shrinkage of the last intron and loss of the proximal acceptor. The specialized *sAPX* has a point mutation generating a pre-mature stop codon (likely CAA to TAA) in the penultimate exon; after the novel stop codon, the homologous sequence has frameshift mutations, making the hypothetic peptide non-conserved, and contains an in-frame stop codon. Though the alternative acceptors are both present, only one type of transcript utilizing the proximal acceptor is generated. I observed that there is a putative poly(A) signal (AATAAA) before the distal acceptor, and I failed to amplify any cDNA using a primer located after the distal acceptor, which suggests that the distal acceptor is abolished. As with *Arabidopsis thaliana* and its relatives, *Gossypium tAPXs* and *sAPXs* reciprocally lost one of the alternative splicing patterns, such that each paralog codes for tAPX or sAPX. The Gossypium sAPXs are longer than the Brassicaceae sAPXs due to the novel stop codon present at a downstream site.

In a final case from eudicots, *Nelumbo nucifera* has a specialized *tAPX* which lost the proximal acceptor and the following stop codon due to expansion of the intron. The specialized *sAPX* has a deletion of two nucleotides in the penultimate exon resulting in a frameshift and a novel stop codon; it also has a mutation in the distal acceptor and pre-mature stop codon within the chain exon that would abolish the coding potential for tAPX.

In monocots there are several cases of *cAPX* duplication. For *Zostera marina*, *tAPX* has an expansion of the intron and loss of the proximal acceptor, while *sAPX* seems to have a rearrangement at the penultimate exon, generating a pre-mature stop codon followed by a

highly divergent sequence after the penultimate exon. In the duckweeds, *Spirodela polyrhiza* and *Lemna minor*, *tAPX* has lost the proximal acceptor and the following stop codon, and *sAPX* has a novel pre-mature stop codon at the penultimate exon as well. Similarly, in *Eichhornia paniculata*, *tAPX* has lost the proximal acceptor due to expansion of the intron; *sAPX* experienced a single nucleotide mutation (AAG to TAG) at the penultimate exon, and as a result only sAPX is encoded despite the loss of a conserved stop codon after the proximal acceptor, and there is a mutation at the distal acceptor causing a frame shift of the chain exon.

Unlike other monocots, Musaceae has a sub-localization pattern similar to Salicaceae. The specialized *tAPX* in *Musa accuminata* contains the proximal acceptor but it is not followed by a stop codon; the specialized *tAPX* in *Ensete ventricosum* lost the proximal acceptor along with the loss of the following stop codon; they both have novel distal acceptor (TCACAG). The specialized *sAPX* in *Musa accuminata* lost sequence similarlity after the distal acceptor with a premature stop codon in the chain exon, while in *Ensete ventricosum* the loss of similarity started after the proximal acceptor including the distal acceptor. As a result, the specialized sAPX is of the same length as orthologous sAPXs.

The Poaceae family has two specialized sAPXs, both of which are due to truncation with a premature stop codon. However, one of them (two as tandem duplicates in *Oryza sativa*) has been relocated to mitochondria as predicted by TargetP (Emanuelsson et al., 2000) and also shown by previous GFP experiments, and only one sAPX is still targeted to plastid stroma (Xu et

al., 2013). The tAPX is specialized by the shrinkage of the intron and loss of the proximal acceptor.

In summary, all the specialized tAPX peptides have the same length of C-terminal anchorage domain as those tAPX peptides generated by alternatively spliced single copy *cpAPX* genes. Looking closely at the genomic sequences of those specialized *tAPX* genes, the last two exons are spliced in the same pattern as the tAPX transcript spliced from single copy *cpAPX* genes. In terms of specialized sAPX, some of them have the same length as the sAPX spliced from single copy *cpAPX* genes, and the last exon only encodes one amino acid, usually Asp (D) or Ala (A) before the stop codon. The others have a shorter peptide, as the stop codon is located on the orthologous penultimate exon, which truncates the peptide at the C-terminus, and thus the coding potential of the orthologous 3' mRNA is abolished. In the sub-localization cases I identified, most sAPXs are specialized by truncation at the penultimate exon. These include *Nymphaea colorata*, *Zostera marina*, Duckweeds, *Xerophyta viscosa*, *Eichhornia paniculata*, Poaceae, *Nelumbo nucifera*, *Solanum* and *Capsicum*, *Manihot* and *Hevea*, *Gossypium*, Cleomaceae and Brassicaceae. I name the truncated sAPXs Type I specialized sAPX, and along with the specialized tAPX, I refer to as Type I sub-localization of cpAPX. Only sAPXs in Salicaceae, Oleaceae and Musaceae have the full-length peptides, and I name them Type II (Fig. 2.4).

2.3.4 Slight acceleration of sequence evolution after sub-localization

I assessed whether the sequence evolution is affected by sub-localization. The single copy genes with alternative splicing all have an extremely low value of Ka/Ks, typically close to 0, and none were over 0.1 (Table 2.3). For many of those sub-localized duplicates, I found a statistically significant elevation of Ka/Ks ratios in one or both paralogs, although the increased values are still only around 0.1 to 0.2. These findings suggest the overall functional conservation of single copy or duplicated *cpAPX* genes (Table 2.3). A similar elevation of Ka/Ks is observed in duplicated paralogs, both of which retain alternative splicing, such as *Glycine max* (Table 2.3). The slight acceleration of sequence evolution I estimated might be simply be a feature seen in duplicated genes as a potential release from constraints initially upon duplication (Lynch and Conery, 2000; Kondrashov et al., 2002). Purifying selection is still acting strongly with minimal relaxation, likely related to the important function cpAPXs play in photosynthesis. This is consistent with the results from the other selection analyses I performed attempting to detect positively selected codons. I did not detect any positively selected sites in any of these duplicated *cpAPXs* according to the branch-site model in PAML.

2.4 Discussion

2.4.1 Duplication of *APX* and loss of alternative splicing

In several lineages, an ancestral alternatively spliced and dual-targeted APX underwent duplication and sub-localization: the paralogous APX genes each encode either the stromal APX or the thylakoid APX to complement the dual localization. In the type I sub-localization, as seen in *Arabidopsis*, *Gossypium*, *Solanum,* for example. (Fig. 2.4), the sAPX was specialized by a point mutation and/or chromosomal rearrangement which introduced a novel stop codon in the penultimate exon, truncating the peptide to be slightly shorter than the sAPX generated by splicing in single copy *cpAPX* genes, while the enzymatic domain stays intact. For some of these specialized *sAPX* genes, I still found the conserved alternative acceptors for the last exon, and some even still harbor detectible homologous exon features such as the stop codon after the proximal acceptor, or homologous sequences after the distal acceptor; however, due to the truncation at the penultimate exon only sAPX peptides will be encoded by those genes. Indeed, as detected by our RT-PCR assays, these *sAPX* genes usually have only one splice form each. Within these specialized *sAPX* genes, I found some cases where the gene sequence is different after the novel stop codon, perhaps due to initial incomplete duplication or subsequent chromosomal rearrangement, and the homologous last exon cannot be detected. In the type II sub-localization, as seen in *Populus*, *Musa*, for example. (Fig. 2.4), the specialized sAPX is generated by the loss of the distal acceptor. Thus, only the proximal acceptor is utilized,

followed closely by the stop codon, and the peptide is the same length as the spliced sAPX in other species that have a single gene with alternative splicing.

The specialized tAPX could only be generated by the degeneration of the proximal acceptor to ensure the skipping of any early stop codon and presence of the C-terminal hydrophobic tail. For all the specialized *tAPX* genes, only the distal acceptor is utilized without alternative splicing. Some specialized *tAPX* genes still have the proximal acceptor, but it is probably not utilized or recognized by spliceosomes as only one transcript is detected by RT-PCR. For many other specialized *tAPX* genes, the proximal acceptor has degenerated due to a point mutation or shrinkage or expansion of the intronic sequences. In summary, the type II sub-localization is accompanied by the reciprocal loss of alternative splicing patterns, and for the type I sub-localization the loss of alternative splicing pattern is necessary for tAPX specialization, but not the deciding contributor for sAPX specialization.

2.4.2 Multiple independent sub-localization events of plastid *APX* genes

I identified many cases of duplicated *cpAPX* genes in angiosperms. Most of them show loss of alternative splicing and specialization of each duplicate to code for sAPX or tAPX. The inferred phylogeny of the *cpAPX* gene tree (Fig. 2.2) largely resembles the expected species tree (The Angiosperm Phylogeny Group, 2016), specifically, the nodes where the sub-localized paralogs arose by duplication are well supported, and many of them have several well supported single copy orthologous genes from sister groups. This provides very strong evidence that those

duplication events are lineage-specific and relatively recent instead of an ancient shared duplication early during angiosperm evolution. The lineage-specific duplication events provide one line of evidence that there have been multiple independent origins of specialized sAPX and tAPX in different lineages. Secondly, there are two types of sub-localization patterns (Type I and II; Fig. 2.4) and they are interspersed across the phylogeny. Third, pairs of paralogs have unique sequence changes causing the localization specialization; for example, the truncated sAPXs from different lineages do not share the same point mutation and/or the novel stop codon. Fourth, I have shown that species with a single copy of *cpAPX* that are relatively closely related to many of those species with duplicated cpAPX show alternative splicing of the single *cpAPX* gene to produce the sAPX and tAPX. Taken together, these lines of evidence indicate that there have been multiple independent cases of duplication of *cpAPX* and loss of ancestral alternative splicing in the duplicates to produce specialized sAPX and tAPX. I propose that this is a type of convergent evolution of gene structures and expression patterns.

2.4.3 Frequent return to single copy status after ancient rounds of whole genome duplication

I found that many flowering plant species that I studied have only have one copy of *cpAPX* despite one or more rounds of WGDs in each lineage as well as other types of expected gene duplication events. The frequent loss of duplicates and reversion to single copy status could be explained by paralogs increasing the targets for malfunctional mutations, which are strongly selected against considering the functional importance of cpAPXs. The product of the deleterious paralog could even play a dominant negative role causing blocking of the proper

functioning paralog, and thus deletion of the mutated paralog occurs (Fig. 2.6A). That might also be a reason why I only found duplicated genes of nearly identical sequences in some polyploid species when there has not been enough evolutionary time for substantial sequence mutation. The rare exceptions, where both duplicates continue doing alternative splicing, are *Glycine max*, and some other quite recent meso-/neo- polyploids that experienced polyploidy after the genus diversification, such as in *Linum usitatissimum* and *Chenopodium quinoa*, and some agricultural polyploids such as *Coffea arabicum* and *Nicotiana tabacum* where the paralogs are over 95% identical in nucleotide sequences (Clarkson et al., 2005; Cenci et al., 2010; Wang et al., 2012; Jarvis et al., 2017).

It has been reported that many organelle associated genes are under-represented among duplicated genes, perhaps due to the above discussed functional importance, as well as a dosage effect: the interacting genes in the organellar genomes are frequently "single copy" regardless of nuclear gene duplication (De Smet et al., 2013; Li et al., 2016).

Pseudogene formation is a possible process that a degrading gene undergoes before a complete deletion, through either transcriptional silencing by loss of regulatory elements or disruption by pre-mature stop codons in the open reading frame. I found two paralogs of *cpAPX* in *Malus domestica* and two *Pyrus* species that are 90% identical at the nucleotide level, but are missing the 5' region and thus represent a pseudogene. I also found a paralogous sequence of *cpAPX* in the *Asparagus officinalis* genome, but there are several in-frame stop codons in it.

These genes might be an intermediate status in the loss of *APX* paralogs in these paleopolyploid genomes.

2.4.4 Retention of paralogous *cpAPX* by a stepwise process of sub-localization

Sub-localization is a mechanism facilitating the long-term retention of functionally important organellar-associated duplicated genes. When paralogs of *cpAPX* diverged and became non-redundant, they were selected to be preserved given their unique and important roles. I present evidence from multiple cases that the specialization of tAPX and sAPX could be stepwise. For example, there are cases of partial sub-localization seen in *Sesamum indicum*, *Ananas comosus* and *Phoenix dactylifera*. In these species, the sAPX specialization took place in one paralog, while the other paralog generates both types of cpAPXs by alternative splicing. They may be an intermediate stage of sub-localization (Fig. 2.6A). When one duplicate experienced sequence changes and became specialized, sAPX in the cases I identified, the other paralog was no longer redundant in terms of tAPX coding capacity. At this stage, the specialized sAPX could get lost, as long as the other paralog performs dual-functions. While the specialized sAPX remains functioning, it is possible that, subsequently, mutations that abolish sAPX coding capacity would occur in the second paralog to make it only encode tAPX. Alternatively, this paralog may primarily produce tAPX to maintain a balanced ratio between two types of cpAPXs. This could be why I observed some specialized *tAPX* genes that gave rise to one type of transcript while two sets of splicing acceptors were still present. By the stepwise process, or in a hypothetical case where it is possible that the specialization of sAPX and tAPX

could happen simultaneously, the sub-localized paralogous *cpAPX* genes must be preserved in a genome to realize the functions in the photosystem properly.

The fates of angiosperm *cpAPX* genes after duplication can be summarized as: 1. gene loss after duplication and thus single copy; 2. pseudogenization of one paralog caused by in-frame stop codons or loss of expression; 3. short-term redundancy with both paralogs produces sAPX and tAPX by alternative splicing, that would be inevitably lost; 4. partial sub-localization with specialized sAPX, but the paralog produces both sAPX and tAPX by alternative splicing; 5. sub-localization with specialized sAPX and tAPX genes (Fig. 2.6B). In the models of the evolutionary trajectories of angiosperm *cpAPX* genes after duplication (Fig. 2.6A), the fates 1 and 5 would be two ultimate states, and the fate 2 is an intermediate stage towards fate 1, while the fate 4 is an intermediate stage towards fate 5 (also possible going back to fate 2 or fate 1 only through pseudogenization of the specialized paralog). Fate 3 is the initial status upon duplication before subsequent divergence (Fig. 2.6B).

2.4.5 Consideration of escape from adaptive conflict

I speculate that, after sub-localization, the specialized sAPX and tAPX gene products can escape from adaptive conflict. An early model for escape from adaptive conflict describes a scenario where a gene with two functions may have either function impeding the improvement of the other function due to constraints in sequences of functional domains (Des Marais and Rausher, 2008). In the context of cpAPXs, the dual function could be considered as the sAPX and tAPX.

Although both cpAPXs have the same enzymatic function, to catalyze the reduction of peroxide produced during photosynthesis, their roles are differentiated because of their sub-organellar location. tAPX is tightly associated with the photosystem 1 complex in the thylakoidal scavenging systems on the membrane, which serves as the primary mechanism of reduction of peroxide, whereas sAPX, which is soluble in the stroma, makes up the stromal scavenging systems and captures peroxide that escaped from the membrane. The substrates, peroxide and ascorbate, are the same, but the local concentration of tAPX and sAPX is different (Asada, 1999). Thus, it is possible that the sub-localized, and thus specialized, cpAPXs evolved to perform better in their respective sub-organellar environments. Evidence to support that hypothesis would be a greatly elevated sequence evolutionary rate, or some positively selected amino-acid residues, in both paralogs. However, I did not find either of these features of sequence evolution in our analyses of selection on the paralogs. A possible explanation is that the shared sequence of tAPX and sAPX in a single gene does not cause strong adaptive conflict. Or perhaps the functions are close enough that sequence differences are not needed. It is also possible the enzymatic activity is close to optimal after long term evolution, considering the cpAPX were already present in the common ancestor of all green lineages, and there is little room for the improvement of kinetics of enzyme-catalyzed reactions. As I observed, after gene duplication, sub-localized genes are still highly conserved in their catalytic domain.

2.4.6 Gene duplication followed by protein subcellular relocalization

A large-scale study of localization of the products of duplicated genes in *Arabidopsis thaliana* showed that changes in subcellular localization are relatively common after gene duplication, affecting at least 15% (19/128) of gene pairs, which is possibly an underestimation as only duplicates with GFP tagging data from both paralogs were analyzed (Liu et al., 2014). An even larger proportion of yeast duplicated pairs, 37% (88/238), experienced extensive protein subcellular relocalization using GFP data (Marques et al., 2008). Protein subcellular relocalization can be represented by neolocalization and sub-localization. Neolocalization happens after gene duplication, where one of the duplicate genes encountered changes in its transit peptide region which caused its products to be directed to a novel location. This is likely a more common type of protein subcellular relocalization in *Arabidopsis thaliana* and yeasts (Marques et al., 2008; Liu et al., 2014). Several examples of neolocalization are also accompanied by neofunctionalization (Liu et al., 2014). The yeast study suggested that sub-localization is relatively rare (Marques et al., 2008). This might be due to the definition of a true sub-localization, that the ancestral single copy gene must be multi-targeting, which is not a necessary feature for most genes.

One likely example of incomplete sub-localization after gene duplication is seen in the paralogous genes encoding the small subunits (SSU) of ADP-glucose pyrophosphorylase (AGPase), a starch-biosynthetic enzyme, in *Zea mays* (Rosti and Denyer, 2007). This incomplete case of sub-localization is accompanied by regulatory subfunctionalization. In most grasses, one

gene encodes two SSU proteins by an alternative first exon, which targeted to the cytosol in the endosperm and the plastids in the leaf. *Zea mays* has a genus-specific whole genome duplication after its divergence with *Sorghum*, which gave rise to *Bt2* and *L2*, a pair of syntenic paralogous SSU genes. These each took a complementary part of the ancestral functions: *Bt2* majorly encodes the endosperm cytosolic SSU although a likely non-functional alternative transcript encoding plastid SSU is still generated, while *L2* has lost one alternative exon, no longer encoding the endosperm cytosolic SSU, and the only transcript form is responsible as the major leaf plastid SSU (Rosti and Denyer, 2007). In this example, L2 experienced contraction of localization pattern due to loss of one of alternative splicing; however, Bt2 remained both spliced transcripts and thus hypothetically are still dual-targeted. This is an incomplete case of sub-localization of duplicated genes, though a perfect example of regulatory subfunctionalization.

2.4.7 Concluding remarks

Protein subcellular relocalization after gene duplication has made an extensive contribution to the divergence of duplicated genes and the process gives rise to potential novel gene functions (Byun-Mckay and Geeta, 2007), leading to retention of relocalized duplicates, and signals for positive selection (Ka/Ks >1.5; Byun and Singh, 2013). Those relocalized genes in most studies are cases of neolocalization, and these neolocalized genes are more likely to coincide with neofunctionalization, because the novel subcellular environments (Marques et al., 2008). However, in the paleo-polyploid yeast *Saccharomyces cerevisiae* there are eight gene families

displaying patterns of sub-localization, where paralogs are subjected to more restricted cellular environments and experience functional specialization accompanied by subfunctionalization (Marques et al., 2008). There are several mechanisms of multi-targeting of peptides encoded by the same gene, such as different post-translational modification, translation from alternative start codon, or different transcripts generated by alternative splicing (reviewed in Silva-Filho, 2003). These mechanisms are the onsets of sub-localization, and they can also be features of a gene that subfunctionalization after gene duplication could act on. The interplay between different cellular processes and evolutionary driving forces can lead to new model of paralogous divergence, which should not be ignored.

Many *cpAPX* genes in angiosperms utilize the alternative splicing mechanism at its transit peptide region to direct its products to different sub-organellar compartments, the plastid stroma or thylakoid respectively. I showed many independent cases of sub-localization of specialized sAPX and tAPX through partitioning of ancestral splicing patterns, which provides a new model on fates of duplicated genes.

Table 2.1. *CpAPX* genes examined in this study.

Species	Family	Gene	Gene identifier	Resourses
Picea abies	Pinaceae	cAPX	PAB00020180	Gymno PLAZA1.0
Pinus taeda	Pinaceae	cAPX	PTA00047602	Gymno PLAZA1.0
Ginkgo biloba	Ginkgoaceae	cAPX	GBI00018595	Gymno PLAZA1.0
Amborella trichopoda	Amborellaceae	cAPX	evm_27.model.AmTr_v1.0_scaffold00017.247; ATR_00017G02440	phytozome v.11; Dicots PLAZA3.0
Nymphaea colorata	Nymphaeaceae	tAPX	Nym0715560	angiosperms.org
		sAPX	Nym0479520	angiosperms.org
Spirodela polyrhiza	Araceae	tAPX	Spipo0G0091000	phytozome v.11
		sAPX	Spipo9G0044500	phytozome v.11
Lemna minor	Araceae	tAPX	Lminor_002279-mRNA-0	CoGe
		sAPX	Lminor_012190-mRNA-0	CoGe
Zostera marina	Zosteraceae	tAPX	Zosma373g00060.1	phytozome v.11
		sAPX	Zosma182g00510.1	phytozome v.11
Dioscorea alata	Dioscoreaceae	cAPX	CZHE02000024.1	GenBank
Xerophyta viscosa	Velloziaceae	tAPX	MJHO01000070	GenBank
		sAPX	MJHO01000119	GenBank
Asparagus officinalis	Asparagaceae	cAPX	lcl|evm.model.AsparagusV1_02.662; MPDI01005603.1; DRX051357; SRX750379	GenBank
Phalaenopsis equestris	Orchidaceae	cAPX	PEQU_10453	CoGe
Dendrobium catenatum	Orchidaceae	cAPX	XM_020824005.1 (LOC110097560)	GenBank
Eichhornia paniculata	Pontederiaceae	tAPX	LTAE01000284.1	GenBank
		sAPX	LTAE01002286.1	GenBank
Elaeis guineensis	Arecaceae	tAPX	p5_sc00045.V1.gene378	MPOB
		sAPX	p5_sc00012.V1.gene1009	MPOB
Phoenix dactylifera	Arecaceae	cAPX	XM_017841389.1 (LOC103701344)	GenBank
		sAPX	XM_008809434.1 (LOC103719946)	GenBank
Musa acuminate	Musaceae	tAPX	MA11G12990; GSMUA_Achr11T12610_001	Monocots PLAZA3.0; phytozome v.11
		sAPX	MA10G16420; GSMUA_Achr10T16040_001	Monocots PLAZA3.0; phytozome v.11
Ensete ventricosum	Musaceae	tAPX	AMZH02015934.1	GenBank
		sAPX	AMZH02012293.1	GenBank
Ananas comosus	Bromeliaceae	cAPX	Aco007908	phytozome v.11

Species	Family	Gene	Gene identifier	Resourses
		sAPX	Aco003278	phytozome v.11
Aquilegia coerulea	Ranunculaceae	cAPX	Aqcoe6G024900	phytozome v.11
Macleaya cordata	Papaveraceae	cAPX	MVGT01002396	GenBank
Nelumbo nucifera	Nelumbonaceae	tAPX	NNU_02610-RA + NNU_02609-RA; XM_010270674.1 (LOC104605785)	LotusDB; GenBank
		sAPX	NNU_10113-RA; NM_001302845.1 (LOC104587545)	LotusDB; GenBank
Beta vulgaris	Amaranthaceae	cAPX	BV1G09500; XM_010676647.1 (LOC104891011)	Dicots PLAZA3.0; GenBank
Spinacia oleracea	Amaranthaceae	cAPX	D77997 and D83669	GenBank
Amaranthus hypochondriacus	Amaranthaceae	cAPX	AHYPO_015530	phytozome v.11
Camellia sinensis	Theaceae	cAPX	JQ011381 and JQ740734	GenBank
Helianthus annuus	Asteraceae	cAPX	Ha412v1r1_05g031810	INRA Sunflower Bioinformatics
Lactuca sativa	Asteraceae	cAPX	Lsat_1_v5_gn_6_95861.1	CoGe
Daucus carota	Apiaceae	cAPX	DCAR_010287	phytozome v.11
Coffea canephora	Rubiaceae	cAPX	Cc10_g12080	Solunam Genome Network
Mimulus guttatus	Phrymaceae	cAPX	Migut.C01118.1	phytozome v.11
Mentha longifolia	Lamiaceae	cAPX	TRINITY_DN54788_c1_g1_i2	Mint Genomics Resource
Mentha piperita	Lamiaceae	cAPX	c92_g1_i1 APXT_ARATH	Mint Genomics Resource
Utricularia gibba	Lentibulariaceae	cAPX	Scf00053.g5658	CoGe
Genlisea aurea	Lentibulariaceae	cAPX	AUSU01003137.1	GenBank
Fraxinus excelsior	Oleaceae	tAPX	FRAEX38873_v2_000093850; CBXU010012425.1; FTPI01002027.1; ERX1054761	ashgenome; Genbank
		sAPX	FRAEX38873_v2_000396400; CBXU010030218.1	ashgenome; Genbank
Olea europaea	Oleaceae	tAPX	FKYM01000953	GenBank
		sAPX	FKYM01042317	GenBank
Sesamum indicum	Pedaliaceae	cAPX	XM_011071197.1 (LOC105155321)	GenBank
		sAPX	XM_011076537.1 (LOC105159461)	GenBank
Ipomoea trifida	Conbolvulaceae	cAPX	Itr_sc002275.1_g00002.1	Sweetpotato GARDEN
Petunia axillaris	Solanaceae	cAPX	Peaxi162Scf00051g00126	Solunam Genome Network
Nicotiana sylvestris	Solanaceae	cAPX	gene_12480	Solunam Genome Network
Nicotiana tabacum	Solanaceae	cAPX1	gene_73315	Solunam Genome Network
		cAPX2	gene_60165	Solunam Genome Network
Solanum lycopersicum	Solanaceae	tAPX	SL11G018550	Dicots PLAZA3.0
		sAPX	SL06G060260	Dicots PLAZA3.0
Solanum tuberosum	Solanaceae	tAPX	ST11G014780	Dicots PLAZA3.0

Species	Family	Gene	Gene identifier	Resourses
		sAPX	ST06G019350	Dicots PLAZA3.0
Capsicum annuum	Solanaceae	tAPX	Capana04g002111	Solunam Genome Network
		sAPX	Capana06g001731	Solunam Genome Network
Kalanchoe laxiflora	Crassulaceae	cAPX	Kalax.0246s0004	phytozome v.11
Vitis vinifera	Vitaceae	cAPX	VV18G01950	Dicots PLAZA3.0
Cephalotus follicularis	Cephalotaceae	cAPX	BDDD01000953	GenBank
Ricinus communis	Euphorbiaceae	cAPX	RC29648G01100	Dicots PLAZA3.0
Jatropha curcas	Euphorbiaceae	cAPX	Jcr4S00512.40; NM_001308719.1 (LOC105638985)	Jatropha database; GenBank
Manihot esculenta	Euphorbiaceae	tAPX	ME02904G00010; ME01393G00720	Dicots PLAZA3.0; PLAZA2.0
		sAPX	ME04314G00120; ME07504G00190	Dicots PLAZA3.0; PLAZA2.0
Hevea brasiliensis	Euphorbiaceae	tAPX	XM_021793603.1	GenBank
		sAPX	XM_021808162.1	GenBank
Linum usitatissimum	Linaceae	cAPX1	Lus10025680	phytozome v.11
		cAPX2	Lus10018155; AFSQ01023381.1	phytozome v.11; GenBank
Populus trichocarpa	Salicaceae	tAPX	PT05G17920	Dicots PLAZA3.0
		sAPX	PT02G08190	Dicots PLAZA3.0
Salix purpurea	Salicaceae	tAPX	SapurV1A.0740s0120	phytozome v.11
		sAPX	SapurV1A.0010s1260 (tandem duplicate SapurV1A.0365s0280)	phytozome v.11
Medicago truncatula	Fabaceae	cAPX	MT3G088160	Dicots PLAZA3.0
Phaseolus vulgaris	Fabaceae	cAPX	Phvul.009G126500	phytozome v.11
Glycine max	Fabaceae	cAPX1	GM04G42720	Dicots PLAZA3.0
		cAPX2	GM06G12020	Dicots PLAZA3.0
Cucumis sativus	Cucurbitaceae	cAPX	Cucsa.060660	phytozome v.11
Citrullus lanatus	Cucurbitaceae	cAPX	CL08G05780	Dicots PLAZA3.0
Ziziphus jujuba	Rhamnaceae	cAPX	XM_016030238.1 (LOC107421087)	GenBank
Humulus lupulus	Cannabaceae	cAPX	HL.Tea.v1.0.G002914.1	Hop Base
Malus domestica	Rosaceae	cAPX	MD00G171270; XM_008366976.2	Dicots PLAZA3.0; GenBank
Fragaria vesca	Rosaceae	cAPX	FV2G29780	Dicots PLAZA3.0
Rubus occidentalis	Rosaceae	cAPX	Bras_T01465_v1.0.a1	Genome Database for Rosaceae
Prunus persica	Rosaceae	cAPX	PPE_001G49870	Dicots PLAZA3.0
Punica granatum	Lythraceae	cAPX	MTKT01000553	GenBank
Eucalyptus grandis	Myrtaceae	cAPX	EG0006G37170	Dicots PLAZA3.0

Species	Family	Gene	Gene identifier	Resourses
Anacardium occidentale	Anacardiaceae	cAPX	Anaoc.0012s0345	phytozome v.12
		sAPX	Anaoc.0009s0964	phytozome v.12
Citrus sinensis	Rutaceae	cAPX	CS00004G00600	Dicots PLAZA3.0
Theobroma cacao	Malvaceae	cAPX	TC0008G12390	Dicots PLAZA3.0
Corchorus capsularis	Malvaceae	cAPX	AWWV01007742	GenBank
Gossypium raimondii	Malvaceae	tAPX	GR10G05140	Dicots PLAZA3.0
		sAPX	GR09G24690	Dicots PLAZA3.0
Gossypium arboreum	Malvaceae	tAPX	Cotton_A_16300 + Cotton_A_16299	CottonGen
		sAPX	Cotton_A_16087	CottonGen
Carica papaya	Caricaceae	cAPX	CP00064G01000 + CP00064G00990; EX262979.1	Dicots PLAZA3.0; GenBank
Tarenaya hassleriana	Cleomaceae	tAPX	Th2v20787	CoGe
		sAPX	Th2v10988	CoGe
Aethionema arabicum	Brassicaceae	tAPX	AA_scaffold1843_141	BRAD
		sAPX	AA_scaffold6394_33	BRAD
Brassica rapa	Brassicaceae	tAPX	Bra015668	BRAD
		sAPX	Bra037859	BRAD
Schrenkiella parvula	Brassicaceae	tAPX	c0013_00407	BRAD
		sAPX	c0018_00115	BRAD
Capsella rubella	Brassicaceae	tAPX	Carubv10020327m	BRAD
		sAPX	Carubv10003730m	BRAD
Arabidopsis thaliana	Brassicaceae	tAPX	AT1G77490	TAIR10
		sAPX	AT4G08390	TAIR10
Oryza sativa	Poaceae	tAPX	OS02G34810	Monocots PLAZA3.0
		sAPX1	OS04G35520	Monocots PLAZA3.0
		sAPX2	OS12G07820 (tandem duplicate OS12G07830)	Monocots PLAZA3.0
Brachypodium distachyon	Poaceae	tAPX	BD3G45700	Monocots PLAZA3.0
		sAPX1	BD5G10490	Monocots PLAZA3.0
		sAPX2	BD4G41180	Monocots PLAZA3.0
Sorghum bicolor	Poaceae	tAPX	SB04G022560	Monocots PLAZA3.0
		sAPX1	SB06G017080	Monocots PLAZA3.0
		sAPX2	SB08G004880	Monocots PLAZA3.0
Zea mays	Poaceae	tAPX	ZM05G30510	Monocots PLAZA3.0

Species	Family	Gene	Gene identifier	Resourses
		sAPX1	ZM02G14900	Monocots PLAZA3.0
		sAPX2	ZM10G03060; XM_008663211 or BT063223.1	Monocots PLAZA3.0; GenBank

Table 2.2 Primers designed for this study.

Single copy genes		
Amborella trichopoda	F	TCAAGGAGAAGAGAGACGAAGA
	R	CTCCCAAGCAGAGATGTCAAA
Aquilegia coerulea	F	GCCAACTGATGCTGTTCTTT
	R	CTGCCTCCAAATCCTTCATATTC
Asparagus officinalis	F	TGCCTACTGATGCTGCTTTAT
	R	CCCTCCAAGTGCTTCATACTC
Beta vulgaris	F	GGTTTCTCCTTAGACGGAAGTC
	R	TGCAAGATATGATAGAACAGCCA
Camellia sinensis	F	CGCAATGGCTGAAGTTTGATAA
	R	AGGAAATAGTTTGACTGGAGAGG
Carica papaya	F	GGAGCACCAGGAGGACAATC
	R	GCAGAGGCTTATCTGGGCTTC
Daucus carota	F	TACGCTGAAGACCAGAAAGC
	R	CTACCACCAACAGCTTGATACT
Fragaria vesca	F	GAAGACGCATCATTCAAGGTATTT
	R	TTATCTGGGCTTCCACCAATAG
Ginkgo biloba	F	CATGGACAGTGGAATGGCTAAA
	R	TCCAAACAGTGCTGCCAAA
Helianthus annuus	F	CTACAGATGCTGCTCTCTTTGA
	R	GAGAGGTTTATCAGGGCTTCC
Mimulus guttatus	F	GGGCAAACCTGAAACCAAATAC
	R	TGGAGAGGCTTATCTGGACTT
Nicotiana sylvestris	F	CTGTGCAGTGGTTGAAGTTTG
	R	GTTTGTTGGGAGAGGCTTATCT
Petunia axillaris	F	TGAGCAACCTTGGAGCTAAAT
	R	CCTCCAAGCAGAGATGTCAAA
Phalaenopsis equestris	F	GCACATTCCAAGCTCAGTAATC
	R	GCAAGCAGAGCAACAACTATC
Picea abies	F	GGACAGTAGAGTGGCTGAAAT
	R	GTTGAGAAAGTAATTAGACTGAAGAGG
Ricinus communis	F	GACAATCCTGGACAGCAGAGTG
	R	CCAGAACAGCAATCACAATCATG
Theobroma cacao	F	GGATGAAGATCTGCTCGTGTT
	R	CCCTCATATTCTGCTCGAATCTT
Vitis vinifera	F	GTGTTGCCAACTGATGCTATTC
	R	TCCACCAACTGCTTCATACTC

Duplicated genes (co-amplify)

Glycine max	F	CTCCAGAGGGCATTGTGATAG
	R	GTTTCCAAGCAGTGATGTCAAA
Linum usitatissimum	F	GACCTACTTGTACTGCCAACTG
	R	CATACTCTGCACGCATCTTCT

Duplicated genes

Ananas comosus	cAPX	F	TTACAGAATGGGGCTTGATGATAAGG
		R	CTGCCTCCTATGGCTTCATACT
	sAPX	F	CTACAGAATGGGCCTAACAGACAAGG
		R	TGAAGCGGTAGGAGGATACACA
Brachypodium dictachyon	tAPX	F	GAAGTACGCAGAAGACCAAGAG
		R	GTAGTTAGACTGCAGAGCCTTC
	sAPX1	F	TTGATGTCACAGGACCTGAGC
		R	GGGAGTGGAAAGGAGTAAACAC
	sAPX2	F	TGTTAAAGAGCGACGAGATGAG
		R	GAACCAAGCTCCATGACCATA
Eichhornia paniculata	tAPX	F	ACGTAGAGATGCAGATCTTT
		R	GCTTCATACTCTGCCCTAATC
	sAPX	F	GCAGAGGGATGCAGATCTGC
		R	GGATGCCACCAATTGCTTCATATTC
Elaeis guineensis	tAPX	F	CCTTCACCTGCTGGTCATTTG
		R	CTGCCTCCAAGTGCTTCATATTC
	sAPX	F	CCATCACCGGGTGCTCATTTA
		R	CCTCCAATTGCTTCATGCTCTG
Fraxinus excelsior	tAPX	F	GCACAACCAGAAAAGTTTGTGG
		R	AGTTCATAACCAAAGATGTCACAAG
	sAPX	F	AAGTTCAACCAGAAAAATTCGTGA
		R	TCTTAGTTCATAACCAAGGATGTCA
Gossypium arboreum	tAPX	F	TGGATGTCTCCGGTCCTAAT
		R	CCCAAATGATTCATATTCTGCTCG
	sAPX	F	TATCAAAGCTAAAAGAGATGAAGATC
		R	CCCACTTCAATAACCAAGAAAC
Lemna minor	tAPX	F	GAAGTATGCAGACGATCAGGAG
		R	GTTGAGGAAGTAGTTGGACTGG
	sAPX	F	CGCTGAGGATGAGAGAGCATTT
		R	AATAACGTCCAACAAGTCCTCGA
Manihot esculenta	tAPX	F	AGGATGAAGATCTACTTGTGTTG
		R	TAATGCCAAAACAGCAATCAC
	sAPX	F	GGGATGAGGATCTACTTGTATTA
		R	AAAAATGCTAAAACAGCAATCATA

Duplicated genes

Musa acuminata	tAPX	F	TCATGGAACGGAAAGATGAAGAG
		R	TGCCAAACCAGCGATCAA
	sAPX	F	AAACAAGGAAAAGATGAAGATCTG
		R	CTCTCTGCAGTTCAAGCATATAA
Nelumbo nucifera	tAPX	F	GATGAAGATCTACTGGTTTTGC
		R	CACGAATCTTCTGCTTCATAGA
	sAPX	F	GATCTAGATCTTTTAGTTCTGC
		R	TGGGATTCATGTTTGGAAATAG
Nymphaea colorata	tAPX	F	GGGATGAAGATTTGCTGGTTTTA
		R	CTTCGTACTCTGCTCGGATTT
	sAPX	F	GACAATTCTTACTTCAAGGAAGTT
		R	GAGTCTCTTCGTCTCTTGGT
Phoenix dactylifera	cAPX	F	CTTCAAGGACATCAAAGAACGAAGG
		R	GCCTCCAACCGCTTCATATTC
	sAPX	F	TTTCGAGGATATCAAACAACGAAGA
		R	AAGCAGAGACGCCAGAAATG
Populus trichocarpa	tAPX	F	CAGCAGAATGGCTGAAGTTTGA
		R	AAGTGCTAGAACAGCAATCACA
	sAPX	F	CAGCAGAATGGCTGAAGTTTGA
		R	GAATGCAAGAACAGCAATCGTG
Sesamum indicum	cAPX	F	GGATTTGCTAGTTTTACCCACA
		R	ACAGCCTGATATTCTGCTCTAA
	sAPX	F	AGATCTATTGGTTTTACCTACC
		R	TGAATACCGAGTGTGTCTATTT
Solanum lycopersium	tAPX	F	ACAAAGAGATGAAGATCTACTAG
		R	CTCCCAAGCCTTCGTATTC
	sAPX	F	ACGAGACAATGATCTGCTAGTTT
		R	CTTGGTTCCGAAAGGCTTCT
Tarenaya hassleriana	tAPX	F	AGTGAAATGGCTAAGATTCGAT
		R	CAGAGAAATTACCGGAGAGATAAG
	sAPX	F	ACAGTGGTTGAAGTTTGACAATTCG
		R	ATGCCCAACACAAACGACATAG
Xerophyta viscosa	tAPX	F	CGGTAGATCAGGAGGCATTT
		R	TTGCTAAGCCGGCTACTAAG
	sAPX	F	AGGGATGAGGATCTGCTAGTTT
		R	TGTGTTCACCTTGTTGGATCAG
Zostera marina	tAPX	F	CTTGCCAACTGATGCAGTTATT
		R	AGCAATGTGAATGCAATCTGTC
	sAPX	F	CCTTCTCCTGCTGAACATCTAC
		R	GAGGCAAGTTTACCACAAATCC

Table 2.3 The selection analyses of *cpAPX* genes

Taxa with paralogs	One-ratio likelihood	One-ratio Ka/Ks	Two-ratio likelihood	Two-ratio Ka/Ks		Three-ratio likelihood	Three-ratio Ka/Ks		
				orthologs	paralogs		orthologs	sAPX	tAPX
Ananas comosus	-4078.295084	0.04475	-4078.287697	0.045	0.0436	-4074.060773	0.0449	0.0202	0.0894
Anacardium occidentale	-4023.474084	0.05139	-4015.984267	0.0453	0.1427	-4015.442916	0.0453	0.1075	0.2119
Arecaceae	-4118.966597	0.05514	-4111.41021	0.0482	0.1348	-4110.543777	0.0482	0.0939	0.181
Brassicaceae	-4699.185047	0.04919	-4696.762375	0.0424	0.063	-4696.495691	0.0424	0.0698	0.0571
Eichhornia paniculata	-3947.079399	0.04908	-3942.38093	0.0449	0.1192	-3942.373585	0.0449	0.1262	0.1146
Gossypium	-3656.19513	0.0476	-3653.066018	0.0433	0.0937	-3652.458144	0.0433	0.1231	0.0665
Lemnoideae	-4944.298141	0.05207	-4943.573546	0.0485	0.0601	-4942.490798	0.0487	0.0462	0.0718
Manihot and *Hevea*	-4303.874537	0.06058	-4296.583988	0.0519	0.1253	-4296.337435	0.052	0.1454	0.1093
Musaceae	-4340.295228	0.05682	-4328.99248	0.046	0.1298	-4328.927261	0.046	0.1202	0.1401
Nelumbo nucifera	-3592.115596	0.05395	-3591.438747	0.0515	0.0725	-3591.362592	0.0515	0.0652	0.0821
Nymphaea colorata	-3754.810248	0.0495	-3754.455087	0.0476	0.0595	-3754.376678	0.0476	0.0532	0.0658
Oleaceae	-4177.342643	0.05177	-4171.629652	0.0477	0.1473	-4170.920702	0.0477	0.2357	0.0932
Poaceae	-5646.59829	0.06965	-5630.77401	0.0485	0.1125	-5630.361996	0.0485	0.1214	0.0984
Salicaceae	-3955.505901	0.05398	-3944.216098	0.0451	0.1609	-3944.095066	0.0451	0.1868	0.146
Sesamum indicum	-3858.574407	0.04511	-3858.232842	0.0464	0.0355	-3858.135576	0.0464	0.0308	0.0419
Solanum and *Capsicum*	-4125.569954	0.05269	-4118.62481	0.0433	0.0968	-4118.56556	0.0433	0.1025	0.09
Xerophyta viscosa	-4092.747737	0.05353	-4086.862871	0.0475	0.1175	-4084.567036	0.0476	0.0649	0.278
Zostera marina	-4250.603628	0.05222	-4244.154369	0.0443	0.1007	-4243.776549	0.0443	0.0831	0.1287
Glycine max	-3844.142231	0.04725	-3842.58622	0.0458	0.118	-3842.556677	0.0458	0.1014	0.1343
Linum usitatissimum	-3671.703391	0.04514	-3670.463927	0.0443	0.1511	-3670.299087	0.0443	0.0706	0.2586

Fig. 2.1 Structure of single copy *cpAPX* gene that can generate dual-targeted peptides by alternative splicing. The last three exons are shown, and the dotted line represents omitted exons and introns. The last exon is alternatively spliced using a pair of acceptors, shown as squares. Arrows indicate transcription, splicing, translation and localization. The blue color is associated with sAPX and the green color is assigned to tAPX. The red bar is the stop codon. The striped bar represents the hydrophobic tail that anchors to the thylakoidal membrane.

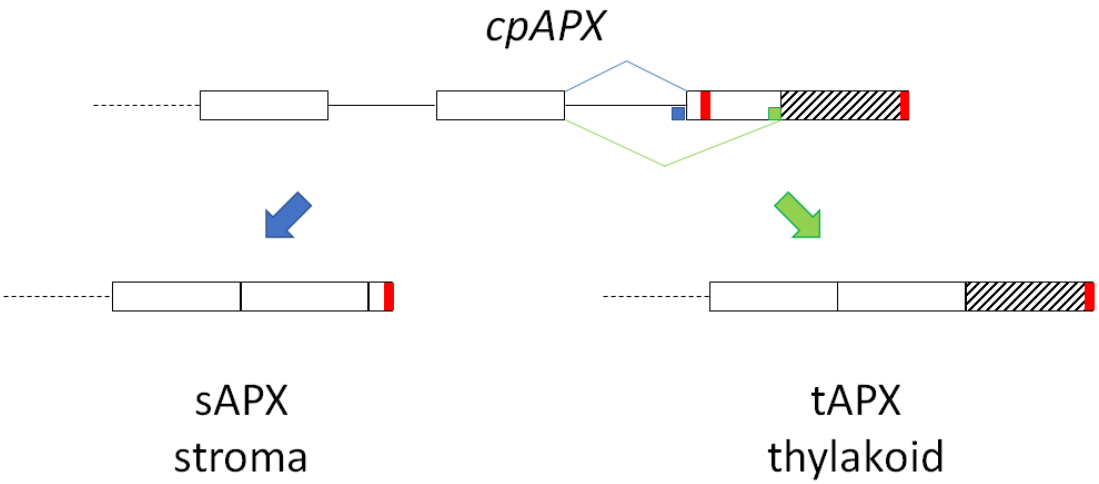


Fig. 2.2. The inferred phylogeny of angiosperm *cpAPX* genes using gymnosperm *cpAPX* genes as the outgroups. Branch width are decided by the bootstrap support for each branch, and a value greater than 50 was labelled. Green terminal branches are *tAPX* genes, and blue terminal branches are *sAPX* genes. Brown terminal branches are duplicates with alternative splicing. One grey terminal is undetermined. The numbered taxa with duplication events are in the same order in the main text.

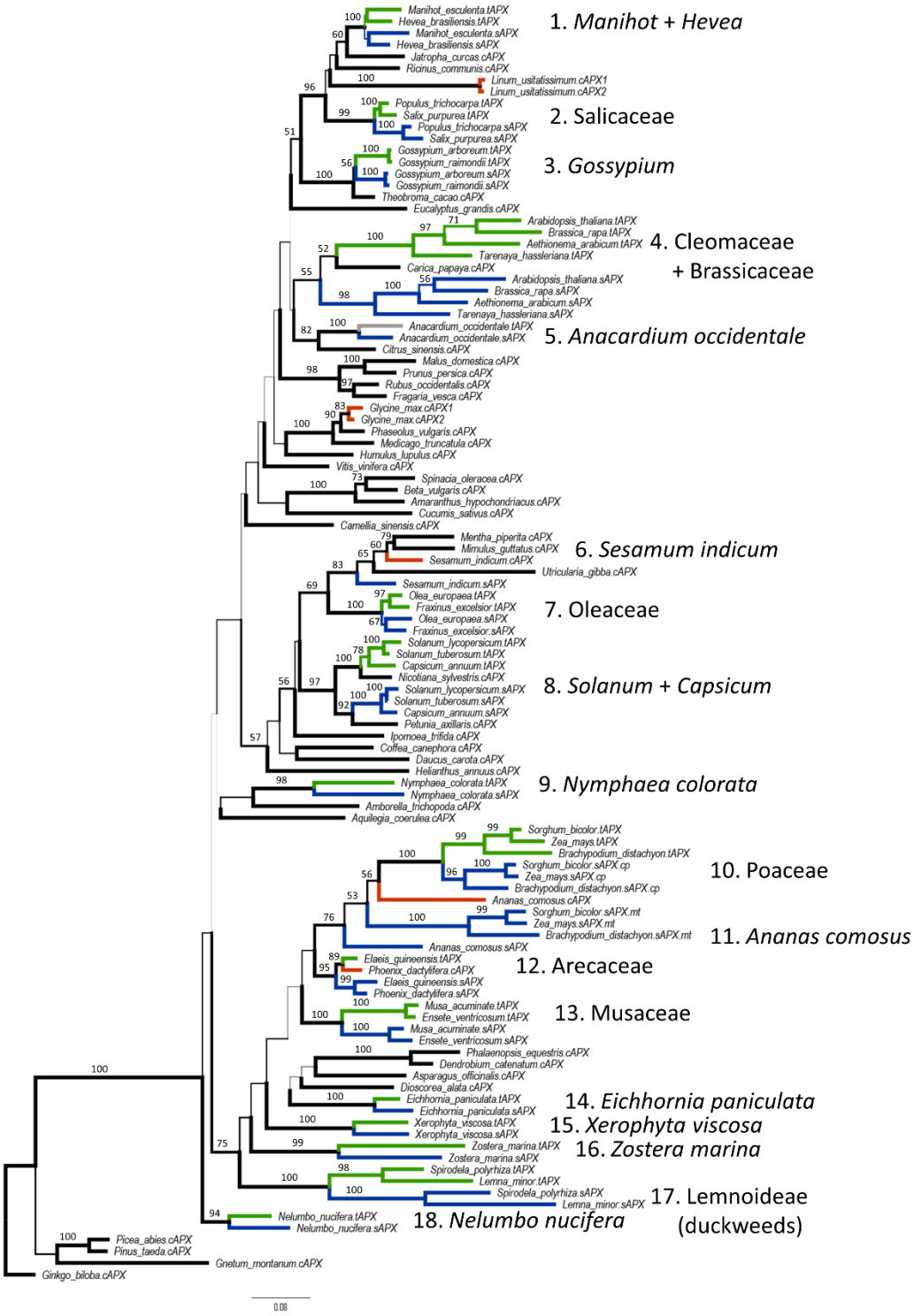


Fig. 2.3. Splicing pattern of surveyed seed plant plastid APXs as detected by RT-PCR. C, single copy cAPX, which are spliced into sAPX (upper band) and tAPX (lower band); T, specialized tAPX and paralogous S, specialized sAPX both give rise to one type of transcripts. Black indicates species with a single copy cpAPX gene. Red indicates species with pairs of specialized tAPX and sAPX. Blue indicates species with pairs of a specialized sAPX and an alternatively spliced paralog. * *Glycine max* and *Linum usitatissimum* each has a pair of cpAPX genes co-amplified. § *Brachypodium dictachyon* has two sAPX forms.

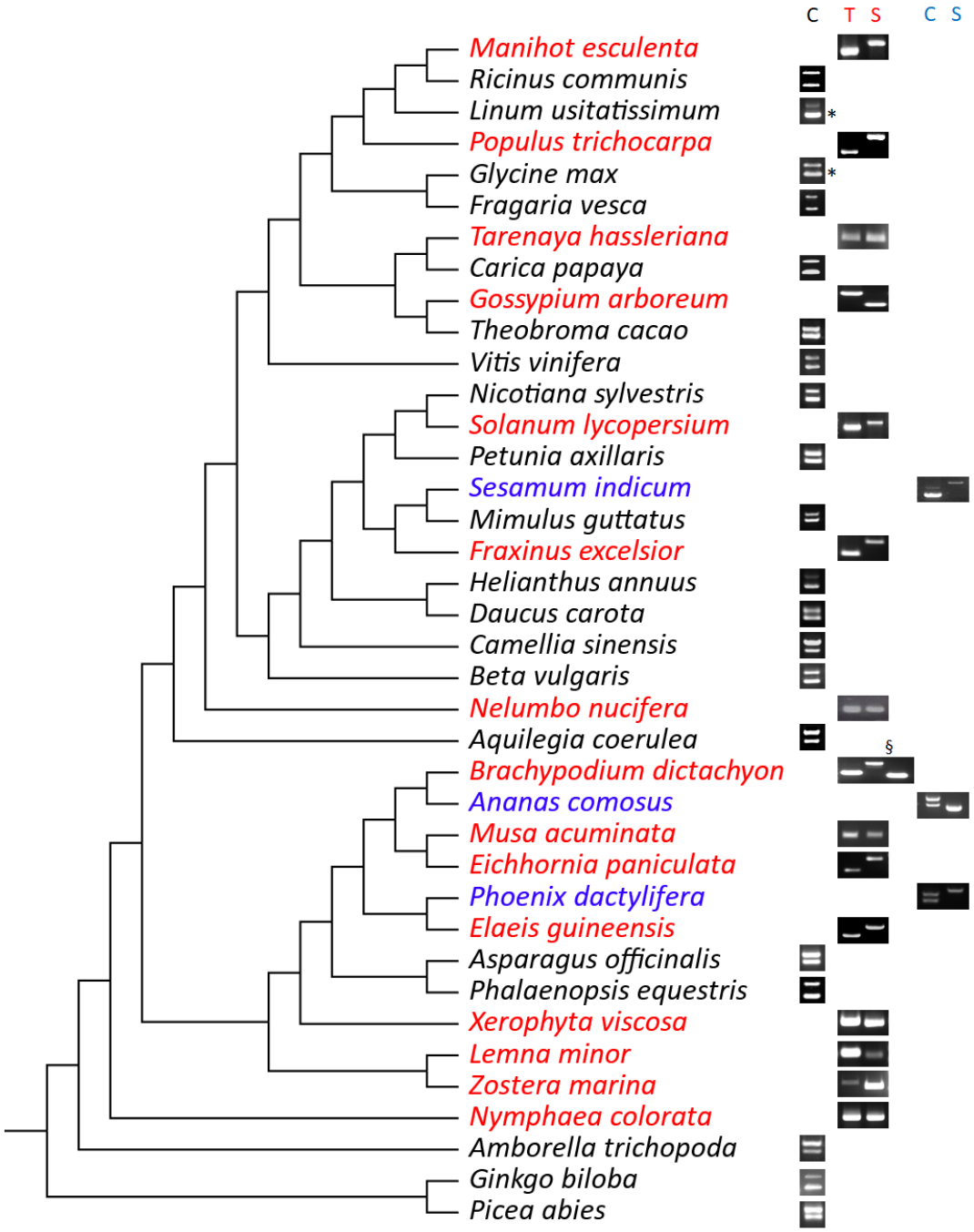


Fig. 2.4. Two types of sub-localization after gene duplication of a single copy alternatively spliced *cpAPX* gene, Type I andType II. The last three exons are shown, and the dotted line represents omitted exons and introns. The horizontal arrows indicate transcription, splicing, translation and localization. The last exon of *cpAPX* is alternatively spliced using a pair of acceptors, shown as squares. The blue color is associated with sAPX and the green color is assigned to tAPX. The red bar is the stop codon. The striped bar represents the hydrophobic tail which anchors APX to the thylakoidal membrane. Vertical black arrows indicate gene duplication and functional specialization. Type I features a truncated sAPX; Type II features a specialized sAPX with the same length as pre-duplicate.

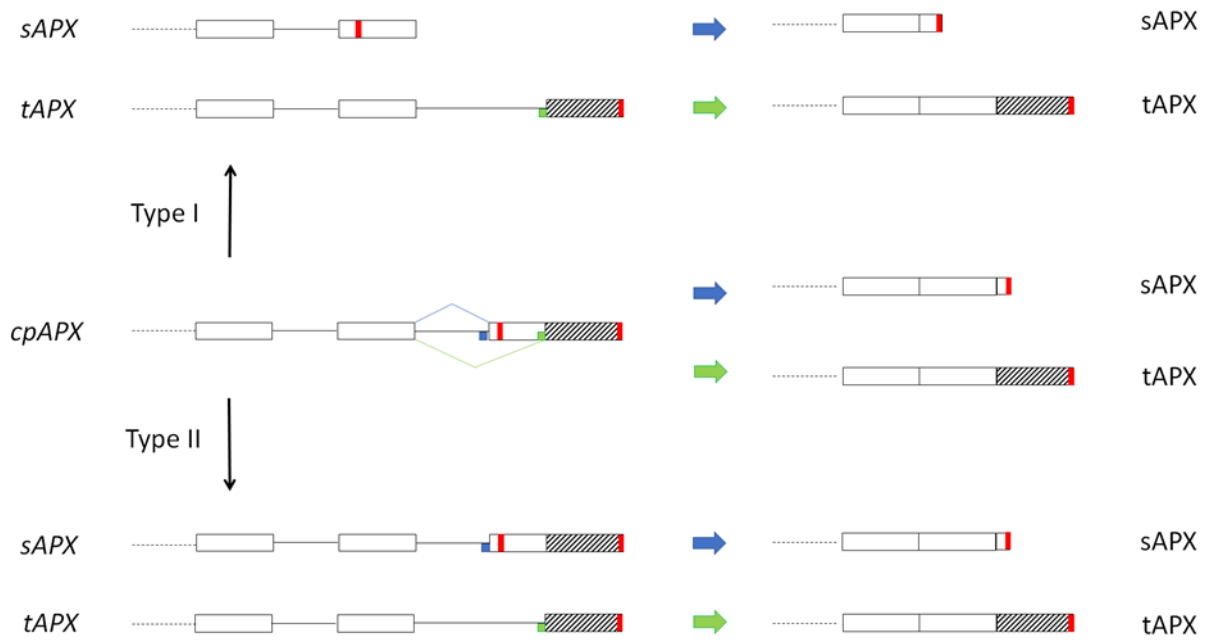


Fig. 2.5. The evolution of *cpAPX* genes in Solanaceae. The genes in black are in the extant genomes. The genes in grey were presumably present in the ancestral genome and were lost. Brown bar indicates gene duplication. Brown circles indicate ancestral alternatively spliced genes. Green and blue boxes indicated possible timing when the inferred gene function specialization to *tAPX* and *sAPX* took place. Green and blue diamonds indicated the ancestral *tAPX* and *sAPX* gene before *Solanum* and *Capsicum* speciation.

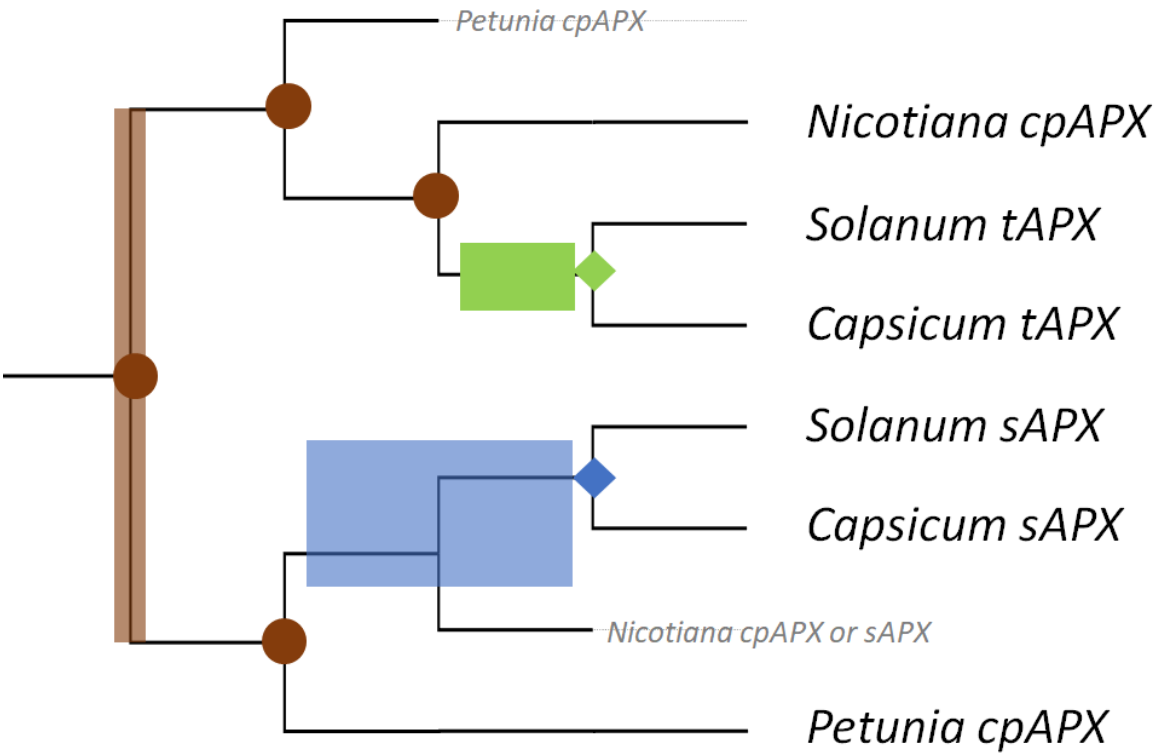
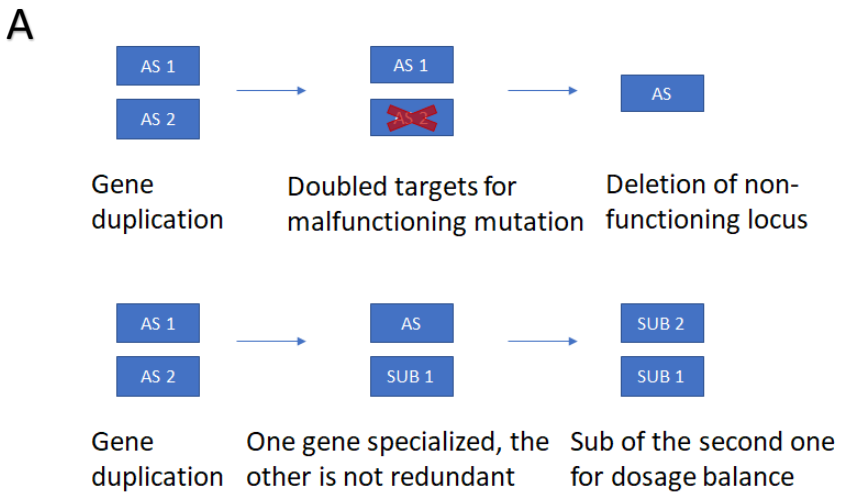
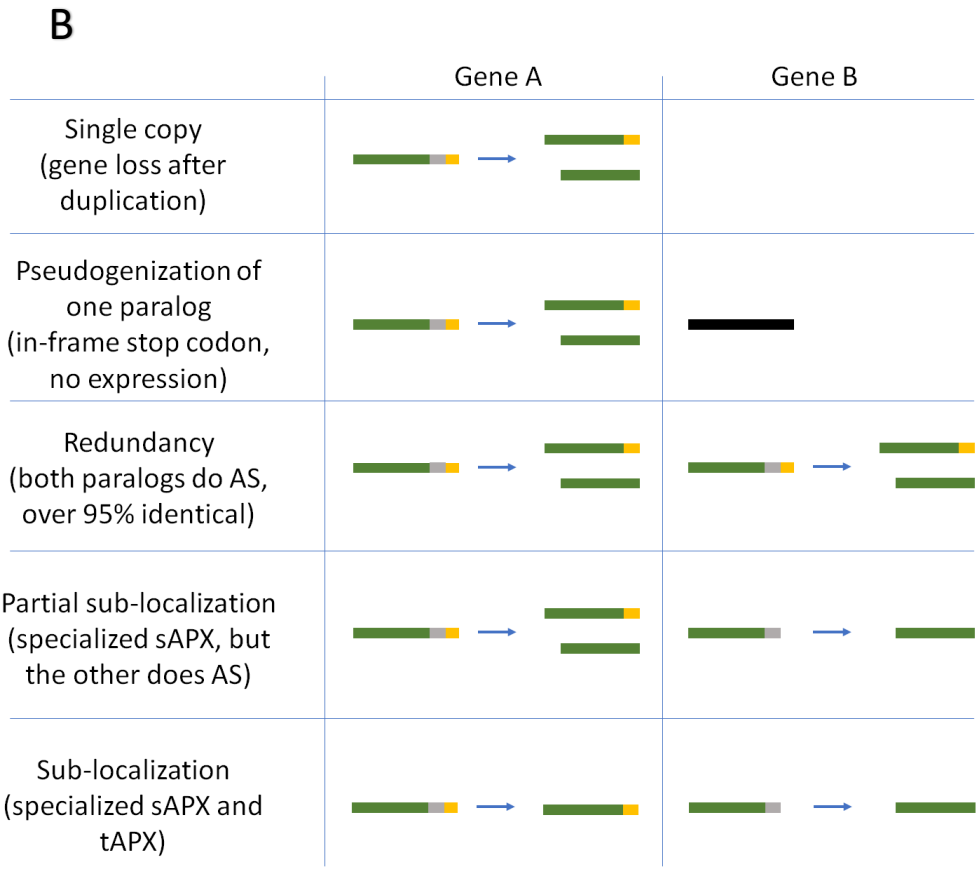


Fig. 2.6. Fates of angiosperm plastid *APXs* after gene duplication. **A.** Models demonstrating the two major mechanisms driving *cpAPX* gene loss or retention. AS indicates alternatively spliced genes, SUB indicates sub-localized genes. **B.** The *cpAPX* genes in extant genomes. The green bar represents constitutive exons, the gray bar and yellow bar represent alternatively spliced intron and exon, and the yellow bar corresponds to the thylakoid anchoring peptide. A black bar is a pseudogene. The arrows indicate transcription plus splicing direction.





3 Expression of a transferred nuclear gene in a mitochondrial genome

3.1 Introduction

Since the origins of mitochondria and plastids by endosymbiosis, three genomes have been coexisting in plant cells. There has been a tendency for DNA from the organelle genomes to be transferred to the nuclear genome, creating many nuclear mitochondrial (numt) sequences and nuclear plastid (nupt) sequences. Numerous pseudogenes of mitochondrial or chloroplast origin are present in nuclear genomes of a wide variety of eukaryotes (reviewed in Bensasson et al., 2001; Kleine et al., 2009). In some cases, large regions of mitochondrial and chloroplast DNA have been transferred to the nuclear genome (e.g., Lin et al., 1999; Lough et al., 2008; Roark et al., 2010). Some mitochondrial and plastid genes were transferred to nuclear genome and then became expressed by acquiring existing nuclear cis-regulatory elements, as well as mitochondrial or chloroplast targeting sequences, then often replacing the functions of their counterparts in the organellar genomes (reviewed in Adams and Palmer, 2003; Bonen and Calixte, 2006; Liu et al., 2009).

Angiosperm mitochondrial genomes contain DNA derived from the nuclear genome, although amounts vary among species. A large amount of the nuclear-derived DNA is from transposable elements, although sequences derived from exons of nuclear genes also are present in some mitochondrial genomes (Kubo et al., 2000; Notsu et al., 2002; Alverson et al., 2010; Alverson et al., 2011; Goremykin et al., 2012; Rice et al., 2013). It has been inferred that the sequences

derived from nuclear genes in mitochondrial genomes are pseudogenes. No nuclear-derived sequences have yet been reported as expressed. Here we show a case of a mitochondrial gene transferred from the nuclear genome that has become expressed.

3.2 Methods and materials

Sequences of *orf164* and *ARF17* from *Arabidopsis thaliana* were obtained from TAIR (v.10). Sequences of *orf164* and *ARF17* from *Arabidopsis lyrata* were obtained from the PLAZA v3.0 Dicots database (http://bioinformatics.psb.ugent.be/plaza/; Van Bel et al., 2012). BLAST searches of GenBank were used to search for sequences homologous to *orf164* and *ARF17* in other species. The nucleotide and amino acid alignments were generated by MUSCLE and followed by manual adjustments (Edgar, 2004).

To analyze sequence rate evolution of *orf164*, sequences of *ARF17* were obtained from several eurosid species including *Carica papaya*, *Citrus sinensis*, *Eucalyptus grandis*, *Populus trichocarpa* and *Prunus persica* from PLAZA v3.0 Dicots (Van Bel et al., 2012), and *Tarenaya hassleriana* from the GenBank wgs database (gb|AOUI01012032.1), and aligned with *orf164* and *ARF17* using MUSCLE with the default settings (Edgar, 2004). The dN/dS ratio along each branch was determined using a phylogeny-based free-ratio test using Codeml in PAML (Yang, 2007).

Total RNA was extracted from multiple organ types of *A. thaliana* (ecotype col-0) and from seedlings of *A. arenosa* using the Ambion RNAqueous Kit following the manufacturer's protocol.

Leaves from *Capsella bursa-pastoris*, *Turritis glabra*, *Erysimum pulchellum*, *Cardamine flexuosa* and *Armoracia rusticana* were used for RNA extraction as above. Nucleic acid concentration and purity were determined using a spectrophotometer and quality was visualized through gel electrophoresis on 2% agarose gel. RNA was treated with DNase-I (New England Biolabs) as outlined by the manufacturer's instructions. Reverse transcription was carried out using M-MLV reverse transcriptase (Invitrogen) following the manufacturer's instructions along with random hexamer primers (IDT). Then PCR reactions were performed with cDNAs as templates. Two pairs of orf164-specific primers were: forward-1, 5'-ATTGACGGCTGAAGCTGTCTCTGA-3'; reverse-1, 5'-ACGCCATGGACCAGTTTCCTGATA-3'; forward-2, 5'-TGTAGTTATTATCAGAGCAATGGAGGCG-3'; reverse-2, 5'-ATAGTGAAGGGGATCTTATACCTGAAGC-3'. Primers for other genes included: *orfX* forward (5'-TGGAGAACAAAGGACGAAATACA-3') and reverse (5'-TATCCGGAGGTGTGGAAAGA-3'); *ccb203* forward (5'-GACCACTACTTCGCCTCTTTG-3') and reverse (5'-CTATGAACGGGAGCTAGCAATC-3'); *matR* forward (5'-TTAAGGACAGGTCGTCGTATTG-3') and reverse (5'-GGTCTCTCATGGCCCAATTAT-3'); *cox2* forward (5'-CGATGAGCAGTCACTCACTTT-3') and reverse (5'-ATTGGATACCCGAGAACCATAATC-3'). The PCR cycling program for orf164 amplification was 94° for 3 min; 20–35 cycles of 94° for 30 s, 55° for 30 s, 72° for 30 s; and 72° for 7 min. PCR cycling conditions for the other genes were the same except that 52° was used as the annealing temperature. PCR products were visualized on 1.2% agarose gels, the bands were cut out of the gels, DNA was eluted and then sequenced to confirm that the amplified sequences were the correct targets.

To identify other mitochondrial open reading frames of nuclear origin, all sequences of nuclear genes in *A. thaliana* were obtained from TAIR (v.10) and aligned against the *A. thaliana* mitochondrial genome (GI:26556996) using YASS software (Noe and Kucherov, 2005) with default parameters. We identified genes having an e-value <1.0E−10 and not located in the chr2:3247243-3509307 region (corresponding to the mitochondrial genome insertion into chromosome 2). The resulting list was filtered to remove transposable element-related sequences, mitochondrial sequences transferred to the nucleus, short open reading frames (less than 300 bp), nuclear intron-derived sequences, and mitochondrial-nuclear sequence pairs with less than 60% identity (Table 3.1).

3.3 Results and discussion

3.3.1 *Orf164* in the mitochondrial genome of *A. thaliana*

Orf164 is a predicted gene in the *A. thaliana* mitochondrial genome (Marienfeld et al., 1999), located between the tRNA gene *trnQ* (tRNA-Gln) and a pseudo-tRNA gene *ψtrnW* for tRNA-Trp (Fig. 3.1). *Orf164*, which has a locus number ATMG00940, contains an intact open reading frame of 495 nucleotides corresponding to 164 amino acids, according to TAIR (v.10) database (http://www.arabidopsis.org/). However, when we analyzed the genomic DNA and cDNA of orf164 by Sanger sequencing following PCR, we detected a sequencing error close to the 3' end of the predicted coding sequence, where an additional A should be present after the 446th nucleotide A, causing subsequent frame shift, and introducing a new stop codon that ends the

coding region earlier. We also checked the recently sequenced and assembled complete mitochondrial genomes from three different *A. thaliana* ecotypes (C24, Ler and Col-0) from Davila et al. (2011) and we found the same additional A. The corrected orf164 open reading frame should be 462 nucleotides and 153 amino acids.

3.3.2 *Orf164* is similar to nuclear *ARF17* and derived from nuclear to mitochondrial intracellular gene transfer

Orf164 has high sequence similarity to a nuclear gene, *ARF17* (*AUXIN RESPONSE FACTOR 17*, AT1G77850). Comparing the sequences, orf164 and ARF17 share 79% nucleotide sequence identity and 81% amino acid identity. *ARF17* has two exons, and the first exon contains a DNA-binding domain and a domain regulating auxin-response gene expression (Fig. 3.2). *Orf164* starts at the position corresponding to the 206th codon within exon 1 of *ARF17*, using an ATG start codon that corresponds to an internal methionine codon in *ARF17*. At the 3'end of the *orf164* coding region, there are eight out of nine consecutive nucleotides that are identical to the intron at the exon–intron junction within *ARF17* (Fig. 3.3). The nucleotide in this region that is not identical to *ARF17* was a mutation that created the *orf164* stop codon. Eighty-four bp of the 5'UTR of *orf164* is derived from *ARF17* (Fig. 3.3). A mitochondrial sequence with similarity to *ARF17* was noticed by Hagen and Guilfoyle (2002) and Liscum and Reed (2002) in articles on *ARF* genes, but neither report identified the mitochondrial sequence as being *orf164* nor did any further characterization.

Using BLAST searches, we found many *ARF17* orthologous genes in a variety of angiosperm species. However, *orf164* has no homologous sequence in any sequenced mitochondrial genomes other than in *Arabidopsis*. We found a sequence from *A. lyrata*, AL3G32400, which is almost identical to *orf164* but annotated as a nuclear gene. However, a block of ten thousand base pairs surrounding AL3G32400 is about 99% identical to the *A. thaliana* mitochondrial genome, indicating that the sequence in *A. lyrata* is actually mitochondrial. *ARF17* in *A. thaliana* and *A. lyrata* are 90% identical, whereas *orf164* and AL3G32400 have an identity of over 99%, with only two nucleotides substituted out of 462 base pairs. We suspect this error regarding annotation of the *orf164* sequence in *A. lyrata* is due to the insertion of the whole mitochondrial genome into the centromere of chromosome 2 in *A. thaliana* (Lin et al., 1999; Stupar et al., 2001). When this region was used as a reference to assemble and annotate the *A. lyrata* genome (Hu et al., 2011), the mitochondrial genome of *A. lyrata* was annotated as being in the nucleus.

Collectively the comparative analyses presented above indicate that *orf164* is derived from *ARF17* through duplicative intracellular gene transfer, from the nuclear genome to the mitochondrial genome. We hypothesize that the transfer was DNA-mediated, and not RNA-mediated, because at the 3'end of *orf164* there are eight out of nine consecutive nucleotides that are identical to the first intron of *ARF17* (Fig. 3.2; Fig. 3.3).

To analyze *orf164* for possible purifying selection, we performed a branch-wise dN/dS test on *A. thaliana orf164* and *ARF17* in the phylogeny with several outgroup ARF17s across eurosids,

using a free-ratio model in PAML (Fig. 3.4). The dN/dS ratios of orf164 and ARF17 are 0.08 and 0.03, respectively, which are statistically not significantly different. Although the dN/dS ratio of *orf164* suggests purifying selection, it may due to the very low sequence evolution rate in plant mitochondria instead of evolutionary constraints on the sequence.

3.3.3 *Orf164* is expressed in several organ types and in five other genera within the Brassicaceae

We used RT-PCR to determine if *orf164* is transcribed. Our results show that *orf164* is transcribed in roots, rosette leaves, stems, cauline leaves, flowers and siliques of *A. thaliana*, indicating a broad expression pattern (Fig. 3.5A). We sequenced the orf164 RT-PCR products to confirm their identity. To verify that the expressed *orf164* is the mitochondrial copy and not the identical copy present in nuclear chromosome 2, derived from transfer of a mitochondrial genome to the nucleus (Lin et al., 1999; Stupar et al., 2001), we assayed *orf164* expression in *A. arenosa*. *A. thaliana* and *A. arenosa* are estimated to have diverged about 5 million years ago (Jakobsson et al., 2006), whereas the timing of the whole mitochondrial genome transfer to nuclear chromosome 2 in *A. thaliana* was estimated at 44,000–176,000 years ago (Huang et al., 2005). We detected expression of *orf164* in *A. arenosa* (Fig. 3.5B).

To determine if *orf164* is present and expressed in other Brassicaceae genera, we performed RT-PCR using RNAs from *Capsella bursa-pastoris*, and *Turritis glabra*, both of which are in the tribe Camelineae along with *Arabidopsis*, as well as *Erysimum pulchellum* in the tribe Erysimeae,

Cardamine flexuosa and *Armoracia rusticana* in the tribe Cardamineae which are close sister tribes to Camelineae (Couvreur et al., 2010). We focused on these tribes because *orf164* is not present in the published mitochondrial genomes of *Brassica* or *Raphanus* (Chang et al., 2011; Tanaka et al., 2012), which are in the tribe Brassiceae. We detected expression of *orf164* in all five species (Fig. 3.5B) and sequenced the RT-PCR products which confirmed their identity.

To compare expression levels of orf164 to other mitochondrial genes, we performed RT-PCR with *orf164* along with *cox2*, *matR*, *ccb203*, and *orfX* using varying numbers of PCR cycles (20, 25, and 30). Although not a quantitative assessment of transcript levels, the assay allows for rough comparisons of transcript levels among the different genes. *Cox2* transcripts were easily detectable at 20 cycles and were most abundant among the five genes, whereas *orf164* transcripts were detectable only with 30 cycles and appeared to be the least abundant (Fig. 3.6). These results suggest that *orf164* transcripts are less abundant than those of several other mitochondrial genes in *A. thaliana*.

How might *orf164* have acquired regulatory elements for expression? One possibility is that the transferred copy inserted near existing cis-regulatory elements, similar to the mechanism by which many mitochondrion-derived genes gained expression after being transferred to the nucleus. *Orf164* is located upstream of the tRNA-Trp pseudogene *ψtrnW* which was derived from the chloroplast *trnW* (Fig. 3.1; Duchene and Marechal-Drouard, 2001). The chloroplast-derived *trnW* genes are present and expressed in the mitochondrial genomes of several other angiosperm species, including potato (Marechal-Drouard et al., 1990), wheat (Joyce and Gray,

1989), sunflower (Ceci et al., 1996), maize (Leon et al., 1989), and beet (Kubo et al., 1995). Thus, it is possible that, after transfer from the nucleus, *orf164* in the Brassicaceae inserted upstream of *trnW* and seized its cis-regulatory elements, acquiring expression while abolishing expression of *trnW*. It is also possible that pseudogenization of *trnW* was not caused by the insertion of *orf164* and instead by mutations in its cis-regulatory elements that abolished transcription. Although *orf164* in *A. thaliana* is 1283 bp upstream of *trnW*, the actual insertion site of *orf164* in an ancestral Brassicaceae species could have been closer to *trnW*, followed by expansion of the intergenic region.

We have shown that *orf164* is transcribed in Arabidopsis and five other genera within the Brassicaceae, but it is not known if the transcripts are translated. Even if the transcripts are translated, the resulting proteins might not be functional in mitochondria. Orf164 contains the auxin responsive element, involved in regulating auxin-response gene expression, derived from ARF17. It is not clear what type of function such a protein would have in mitochondria. Many other transcribed open reading frames with no obvious functions in mitochondria, and typically not conserved among species nor derived from the nuclear genome, have been identified in the mitochondrial genomes of rice and tobacco (Fujii et al., 2011; Grimes et al., 2014). Thus, plant mitochondria may contain numerous transcribed ORFs that do not code for functional proteins, with the number and type varying by species.

3.3.4 Search for other nuclear-derived open reading frames in the *A. thaliana* mitochondrial genome

To search for other sequences in the mitochondrial genome of *A. thaliana* that are derived from a nuclear protein-coding gene, we aligned the sequences of all annotated nuclear genes to the mitochondrial genome (see Section 2). We found only one other gene of possible interest, *orf160* which is partly derived from *MMD1* (AT1G66170), but the sequence has many indels relative to *MMD1* that disrupt the reading frame; thus, we did not study it further.

3.3.5 Conclusions

This study shows a case of transfer of a nuclear gene to the mitochondrial genome and expression of the transferred gene, which is a phenomenon that has not been previously reported. The transfer appears to be DNA-mediated, rather than RNA-mediated. It is possible that *orf164* gained transcriptional regulatory elements from the *trnW* gene for tRNA-Trp. *Orf164* is present and expressed in several genera of the Brassicaceae, but not in *Brassica* or *Raphanus*, and thus the transfer may be a relatively recent evolutionary event. Other angiosperm mitochondrial genomes may contain genes that were transferred from the nucleus and gained regulatory elements to become expressed. This study provides a novel perspective on the movement of genes between the genomes of subcellular compartments.

Table 3.1 YASS alignment results from nuclear genes searched against the mitochondrial genome.

gene_name	% identity	alignment_length	gene_start	gene_end	mt_start	mt_end	E-value	filtering results
AT1G75930.1	100	67	297	363	177394	177328	7.90E-17	short alignment
AT1G66170.1	60.79	380	1082	1453	151832	151455	7.50E-28	orf160
AT1G31163.1	71.92	146	1033	1170	222882	223025	6.60E-12	short alignment
AT1G28135.1	95.77	71	89	159	9155	9225	7.90E-17	short alignment
AT1G28135.1	94.12	68	156	223	237239	237306	3.70E-15	short alignment
AT1G28135.1	91.18	68	22	89	230759	230692	1.10E-12	short alignment
AT1G65350.1	99.9	2883	1432	4314	122557	125439	0	mitochondrial sequence transferred to nucleus
AT1G65350.1	98.01	604	831	1431	289241	288638	5.20E-246	mitochondrial sequence transferred to nucleus
AT1G65350.1	97.43	389	4318	4706	86898	86510	3.20E-150	mitochondrial sequence transferred to nucleus
AT1G65350.1	72.61	230	1435	1663	90394	90598	4.80E-22	short alignment
AT1G65350.1	77.55	147	1436	1580	319760	319615	7.60E-22	short alignment
AT1G65350.1	73.94	142	1437	1576	359264	359124	1.30E-17	short alignment
AT1G65350.1	95.38	65	1620	1684	38146	38082	8.90E-14	short alignment
AT1G65350.1	86.67	75	1469	1538	305929	306003	5.30E-12	short alignment
AT1G77850.1	79.42	549	624	1169	251981	251434	1.50E-136	orf164
AT1G58602.1	60.97	948	5079	6012	228923	227983	1.90E-66	TE
AT1G58602.1	63.08	409	3133	3540	1484	1091	2.30E-33	completely intron, no ORF
AT1G58602.1	66.85	184	5028	5209	315	134	4.70E-15	short alignment
AT1G48690.1	88.16	76	1	76	77003	77077	6.60E-12	short alignment
AT1G48690.1	88.16	76	1	76	240978	241052	6.60E-12	short alignment
AT1G43665.1	73.33	105	427	531	222932	223036	4.10E-11	short alignment
AT1G58602.2	60.97	948	5098	6031	228923	227983	1.90E-66	TE
AT1G58602.2	63.08	409	3152	3559	1484	1091	2.30E-33	completely intron, no ORF
AT1G58602.2	66.85	184	5047	5228	315	134	4.70E-15	short alignment
AT1G65346.1	100	858	1	858	124363	125220	0	mitochondrial sequence transferred to nucleus
AT2G18320.1	73.73	118	101	215	222918	223035	3.40E-12	short alignment
AT2G34520.1	82.09	296	247	542	58341	58635	7.90E-74	short alignment
AT2G46505.1	76.73	159	377	535	219198	219356	2.20E-25	short alignment
AT2G01810.1	67.8	177	875	1051	151789	151613	3.40E-19	short alignment
AT2G36940.1	85.71	126	17	142	145829	145707	2.80E-26	short alignment
AT2G01010.1	65.4	237	1113	1344	362407	362173	1.50E-22	short alignment

gene_name	% identity	alignment_length	gene_start	gene_end	mt_start	mt_end	E-value	filtering results
AT2G01010.1	60.49	243	1413	1655	361759	361524	4.70E-15	short alignment
AT2G06645.1	66.86	175	10	183	310327	310153	9.60E-16	short alignment
AT2G24755.1	67.37	285	1857	2140	310826	310547	3.60E-33	short alignment
AT2G24755.2	66.6	494	1857	2349	310826	310342	7.20E-65	TE
AT2G24755.3	66.6	494	1857	2349	310826	310342	7.20E-65	TE
AT2G18323.1	99.06	213	1	213	2826	3038	4.00E-80	short alignment
AT2G18323.1	99.06	212	1	212	119701	119490	9.80E-80	short alignment
AT3G29800.1	97.7	174	1669	1842	119988	119815	1.90E-59	short alignment
AT3G58390.1	95.65	69	157	225	273173	273105	9.60E-16	short alignment
AT3G47020.1	92.54	67	500	566	324988	324922	4.40E-13	short alignment
AT3G10280.1	100	123	601	723	314587	314465	6.00E-41	short alignment
AT3G07610.1	78.08	219	3979	4191	105084	105296	3.90E-35	short alignment
AT3G27530.1	84.54	97	1336	1431	23707	23612	3.00E-15	short alignment
AT3G59360.1	87.64	89	1415	1503	260646	260560	4.90E-16	short alignment
AT3G59360.1	84.27	89	1415	1503	241998	241912	8.90E-14	short alignment
AT3G22234.1	80.45	133	442	567	78123	77991	1.20E-21	short alignment
AT3G22238.1	80.45	133	442	567	78123	77991	1.20E-21	short alignment
AT3G54350.1	97.67	86	1207	1292	210013	209928	3.20E-23	short alignment
AT3G29636.1	89.36	94	757	848	157128	157221	9.10E-21	short alignment
AT3G41768.1	65.82	237	1113	1344	362407	362173	2.50E-23	short alignment
AT3G41768.1	60.49	243	1413	1655	361759	361524	4.70E-15	short alignment
AT3G23780.1	97.14	70	2008	2077	274256	274187	2.00E-16	short alignment
AT3G20950.1	76.47	119	180	294	223035	222918	1.70E-12	short alignment
AT3G55930.1	75.22	113	398	507	223035	222924	5.10E-11	short alignment
AT3G54350.2	97.67	86	1207	1292	210013	209928	3.20E-23	short alignment
AT3G59360.2	87.64	89	1421	1509	260646	260560	4.90E-16	short alignment
AT3G59360.2	84.27	89	1421	1509	241998	241912	8.90E-14	short alignment
AT3G54350.3	97.67	86	1219	1304	210013	209928	3.20E-23	short alignment
AT3G11945.1	84.09	88	1864	1950	323579	323664	2.70E-12	short alignment
AT3G60961.1	65.1	341	309	648	285419	285758	2.20E-38	completely intron, no ORF
AT3G60961.1	60.39	361	6536	6891	228520	228878	2.00E-23	TE
AT3G60961.1	64.97	177	6718	6894	312180	312005	1.40E-12	short alignment
AT3G11945.2	84.09	88	1864	1950	323579	323664	2.70E-12	short alignment

gene_name	% identity	alignment_length	gene_start	gene_end	mt_start	mt_end	E-value	filtering results
AT3G31005.1	64.5	169	7592	7760	312004	312172	1.70E-12	short alignment
AT3G60164.1	55.14	350	1615	1962	228531	228880	1.70E-12	low identity
AT3G07610.3	78.08	219	4058	4270	105084	105296	3.90E-35	short alignment
AT3G23780.2	97.14	70	2027	2096	274256	274187	2.00E-16	short alignment
AT3G25820.2	100	65	3076	3140	58987	58923	1.60E-16	short alignment
AT3G25820.2	98.41	63	3015	3077	153481	153419	2.40E-15	short alignment
AT3G25820.2	100	55	2959	3013	165839	165785	4.20E-12	short alignment
AT4G09860.1	67.27	330	1	272	136005	136327	4.20E-37	mitochondrial sequence transferred to nucleus
AT4G06611.1	65.87	167	11	172	222207	222041	2.80E-13	short alignment
AT4G23160.1	62.71	649	952	1599	69446	68814	3.40E-63	TE
AT4G23160.1	60.94	361	292	644	228968	228614	5.60E-20	TE
AT4G23160.1	56.98	344	1398	1733	285754	285418	7.10E-14	low identity
AT4G30080.1	63.29	286	1570	1853	251727	251445	2.70E-25	short alignment
AT4G04840.1	95.38	65	1004	1068	172017	171953	3.60E-14	short alignment
AT5G22260.1	62.43	189	1136	1324	151815	151627	1.70E-12	short alignment
AT5G36180.1	79.17	96	1182	1276	222941	223035	2.10E-11	short alignment
AT5G35615.1	57.14	371	149	515	91452	91082	1.00E-17	low identity
AT5G34852.1	88.12	160	91	244	146188	146029	7.60E-41	short alignment
AT5G62165.1	79.57	93	1856	1947	222944	223035	8.00E-11	short alignment
AT5G37960.1	94.74	95	38	132	337930	338024	1.10E-24	short alignment
AT5G62165.2	79.57	93	1811	1902	222944	223035	8.00E-11	short alignment
AT5G06043.1	92	75	1	75	142495	142569	3.20E-17	short alignment
AT5G06043.1	96.83	63	214	276	305520	305458	2.80E-13	short alignment
AT5G62165.3	79.57	93	1796	1887	222944	223035	8.00E-11	short alignment
AT5G18404.1	98.28	58	202	258	164218	164161	6.40E-11	short alignment
AT5G24065.1	89.88	563	13	563	105583	105021	6.80E-165	mitochondrial sequence transferred to nucleus
AT5G32690.1	67.12	1673	10978	12628	221925	220270	4.40E-235	mitochondrial sequence transferred to nucleus
AT5G32690.1	59.14	793	1321	2097	228191	228981	1.40E-45	low identity
AT5G32690.1	63.55	310	12626	12935	219919	219619	3.00E-28	mitochondrial sequence transferred to nucleus
AT5G62165.4	79.57	93	1809	1900	222944	223035	8.00E-11	short alignment
AT5G62165.5	79.57	93	1809	1900	222944	223035	8.00E-11	short alignment

Fig. 3.1. Diagram of the *Arabidopsis thaliana* mitochondrial genome, with the region around *orf164* shown in detail. Arrows indicate the direction of transcription of the genes. Triangles followed by numbers indicate the sizes of the intergenic regions upstream and downstream of *orf164*.

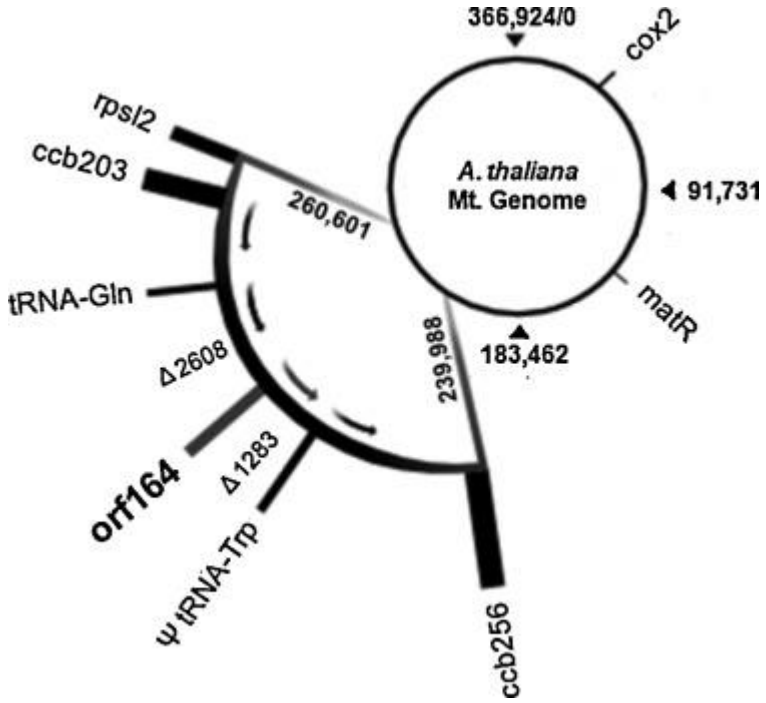


Fig. 3.2. Structures of *orf164* and *ARF17*. Arrows indicate the transcription start sites. Boxes indicate exons and bars indicate introns.

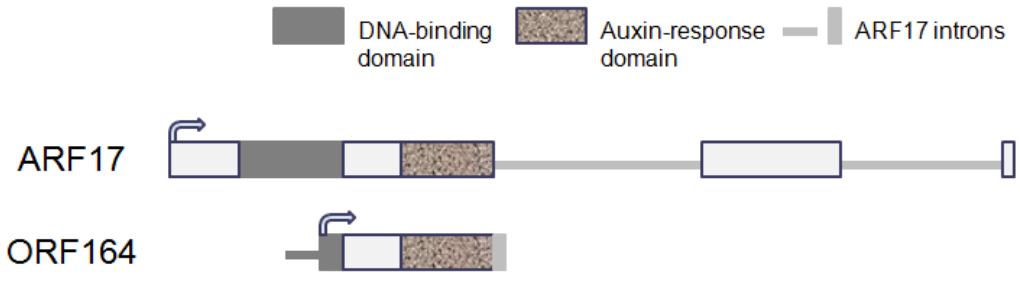


Fig. 3.3. Alignment of *ARF17* and *orf164* in the region of *orf164* corresponding to *ARF17*. Functional domains and the intron sequence are indicated by lines below the corresponding alignment region.

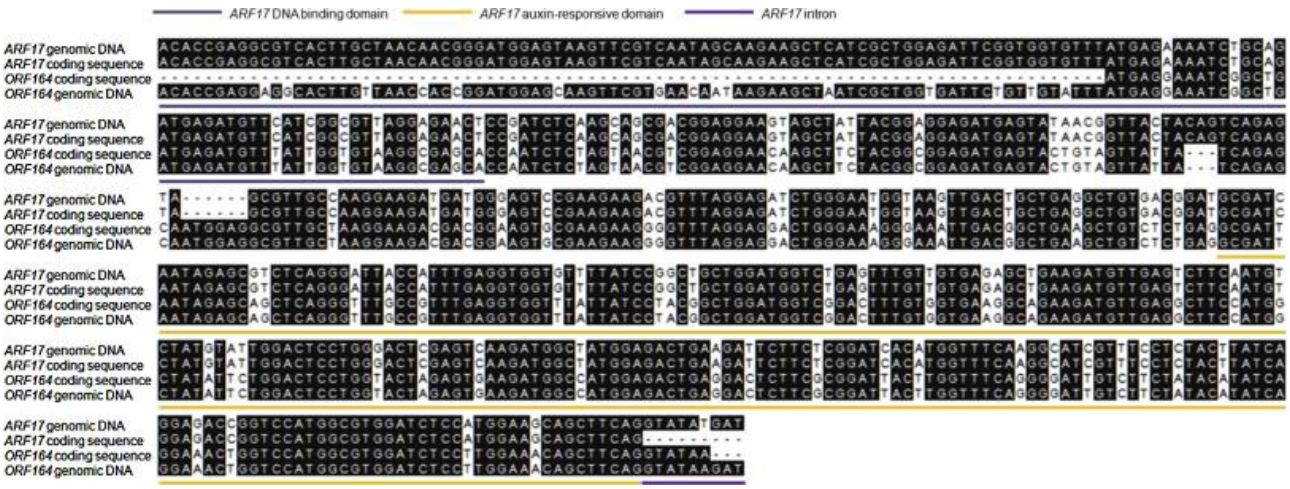


Fig. 3.4. Sequence evolution analysis of *orf164* and *ARF17* sequences. Numbers above the branches indicate dN/dS values and the scale bar indicates substitutions per codon. Taxa abbreviations include: Ath – *Arabidopsis thaliana*, Tha – *Tarenaya hassleriana*, Cpa – *Carica papaya*, Csi – *Citrus sinensis*, Egr – *Eucalyptus grandis*, Ptr – *Populus trichocarpa* and Ppe – *Prunus persica*.

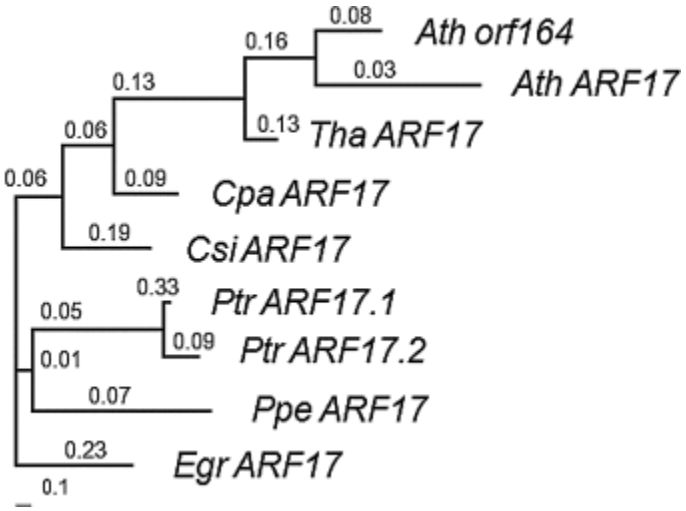


Fig. 3.5. Expression of *orf164*. (A) RT-PCR products of *orf164* in multiple organ types of *Arabidopsis thaliana*. Plus signs indicate the presence of reverse transcriptase and minus signs indicate absence of reverse transcriptase. (B) RT-PCR products of *orf164* in Brassicaceae species.

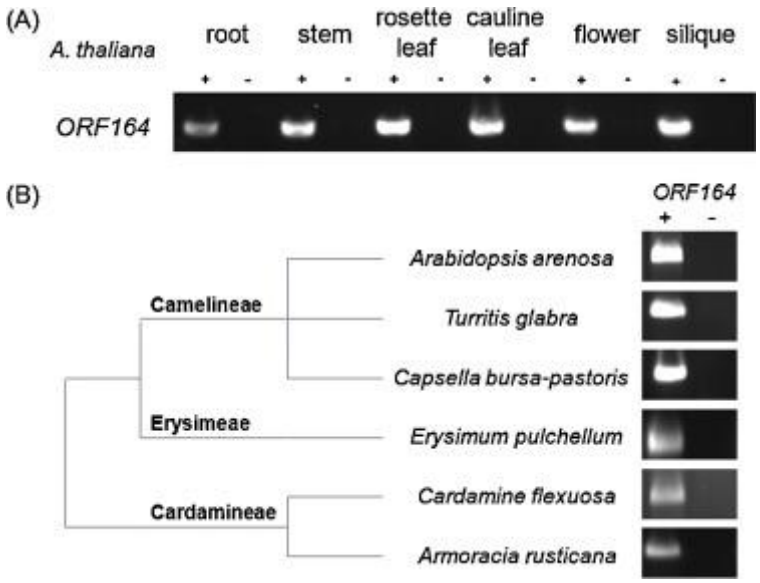
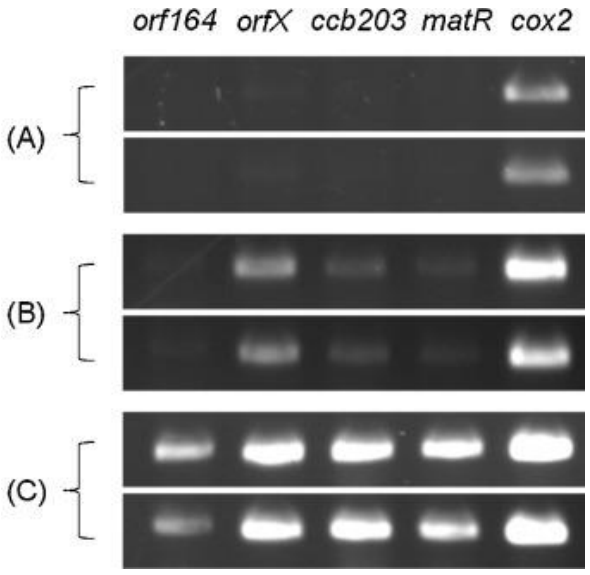


Fig. 3.6. Expression of *orf164* in comparison to four other mitochondrial genes in *A. thaliana*. PCR products amplified with the same amount of cDNA were generated using 20 cycles (A), 25 cycles (B), and 30 cycles (C). Two lanes in each section represent two replicates. Each column represents a mitochondrial gene labeled at the top.



4 Concerted Divergence after Gene Duplication in Polycomb Repressive Complexes

4.1 Introduction

Duplicated genes are continuously formed during evolution by various types of gene duplication events in eukaryotes, and they can have effects on morphological and physiological evolution (for review, see Van de Peer et al., 2009; Soltis and Soltis, 2016). Gene duplication can happen at small scales, such as tandem duplication, segmental duplication, and duplicative retroposition. The largest scale of gene duplication is whole-genome duplication (WGD), which gives rise to thousands of duplicated gene pairs. The genetic model plant Arabidopsis (*Arabidopsis thaliana*) has experienced five rounds of WGD events in the evolutionary history of seed plants (Jiao et al., 2011; Li et al., 2015). The most recent polyploidy event, the α WGD, is specific to the Brassicaceae family, which took place after the divergence of the closest sister family, Cleomaceae (Schranz and Mitchell-Olds, 2006). There are about 2,500 pairs of duplicated genes retained from this WGD in the *Arabidopsis* genome (Blanc et al., 2003; Bowers et al., 2003).

Fates of duplicated genes vary during evolutionary history. One duplicate may eventually be lost or become a pseudogene; thus, the once duplicated pair returns to a single copy status. Several mechanisms drive the retention of both copies. Duplicated pairs could preserve similar functions to maintain dosage balance (Birchler et al., 2005; Coate et al., 2016). Duplicated pairs

also can diverge through subfunctionalization or neofunctionalization, where two duplicated genes divide the ancestral function or gain a novel function, respectively (Force et al., 1999; Moore and Purugganan, 2005). These types of divergence also could be inferred from expression patterns. For example, two duplicates that together make up the preduplicate expression profile is referred to as regulatory subfunctionalization, and regulatory neofunctionalization indicates that one or both copies gain a new expression pattern (Duarte et al., 2006; Liu et al., 2011). Sometimes, these processes are difficult to distinguish, and there can be a combination of different mechanisms, such as subneofunctionalization (He and Zhang, 2005).

There are many protein complexes whose members are encoded by different gene families. If multiple components in a complex are duplicated simultaneously, such as in a WGD, the doubled components could redundantly cross-interact or go on to experience subsequent divergence (Capra et al., 2012; Aakre et al., 2015). Thus, a type of coevolution between the interacting gene products is hypothetically possible, but this has not been described in the plant kingdom. Extending the concept of concerted divergence, which is discussed in the context of coexpression patterns of duplicated genes in the same metabolic or regulatory pathways (Blanc and Wolfe, 2004), we here propose the evolutionary scenario that simultaneous duplication of two genes whose products function together in a complex, followed by parallel evolution and the divergence of each derived gene, can lead to functional divergence of the complexes.

In this study, we focus on genes in POLYCOMB REPRESSIVE COMPLEX2 (PRC2) in Brassicaceae species as a potential example to demonstrate the proposed scenario. Those complexes are histone modifiers and regulate gene expression primarily by the trimethylation of Lys-27 on histone H3 (H3K27me3) associated with target genes, which leads to transcriptional repression (Hennig and Derkacheva, 2009; Mozgova et al., 2015). One type of PRC2, the VERNALIZATION (VRN) complex, regulates vegetative tissue differentiation and, more importantly, the vernalization process to control flowering time in Arabidopsis (Chen et al., 2009; Hennig and Derkacheva, 2009; Mozgova et al., 2015). This complex also represses autonomous seed coat development (Roszak and Köhler, 2011), and it is present across rosids. The VRN complex consists of four subunits: REDUCED VERNALIZATION RESPONSE2 (VRN2), SET DOMAIN-CONTAINING PROTEIN10 (SWINGER [SWN]), and two WD-40 repeat proteins that act as the scaffold of the complex assemblies, FERTILIZATION-INDEPENDENT ENDOSPERM (FIE) and MULTICOPY SUPRESSOR OF IRA1 (MSI1). In Brassicaceae, the α WGD gave rise to a duplication of VRN2 to create its paralog FERTILIZATION INDEPENDENT SEED2 (FIS2) and a duplication of SWN to create its paralog SET DOMAIN-CONTAINING PROTEIN5 (MEDEA [MEA]; Fig. 4.1; Spillane et al., 2007; Luo et al., 2009). Substituting for their paralogous proteins, FIS2 and MEA, together with FIE and MSI1 (the α WGD paralogs of these two genes were lost), make up a new Brassicaceae-specific PRC2, referred to as the FIS complex (Fig. 4.1). The FIS complex functions in gametophyte and seed development, preventing female gamete proliferation before fertilization and facilitating endosperm cellularization after fertilization (Hennig and Derkacheva, 2009). A typical *fis* phenotype, caused by nonfunctional mutation in *FIS2*, *MEA* (also known as *FIS1*), or *FIE* (also known as *FIS3*), shows fertilization-independent

embryogenesis, and other types of mutants have abnormal seed development, even abolished seeds (Hennig and Derkacheva, 2009).

The observed divergence in the functions of the two kinds of PRC2 complexes leads to the hypothesis that *FIS2* and *MEA* have undergone divergence in a concerted way to give rise to the FIS complex. This study aimed to evaluate this hypothesis by examining expression patterns, DNA and histone methylation, and rates of sequence evolution in both genes compared with their paralogs. We found evidence for the parallel divergence of *FIS2* and *MEA* from their paralogs in multiple ways, which has accompanied functional divergence of the two complexes. This study supports a model of concerted divergence of simultaneously duplicated genes whose products function in a complex. To our knowledge, this is a previously unreported fate of duplicated genes.

4.2 Materials and Methods

4.2.1 Comparing Expression Specificity and Detecting Coexpression Using Microarray Data Analyses

Two sets of ATH1 microarray data from Arabidopsis (*Arabidopsis thaliana*) were obtained: the ADA (Arabidopsis developmental atlas) from the TAIR Web site (http://www.Arabidopsis.org/), which included 63 different organ types and developmental stages (Schmid et al., 2005), and the ASA (Arabidopsis seed atlas) from the Goldberg Lab *Arabidopsis thaliana* Gene Chip

Database (http://seedgenenetwork.net/arabidopsis), which included 42 different tissue types from seed developmental stages (Le et al., 2010; Belmonte et al., 2013). The data were GC-RMA normalized using the gcrma package in R. We used the expression specificity (τ) defined by Yang and Gaut (2011) to describe the expression patterns of *FIS2*, *VRN2*, *MEA*, and *SWN*:

$$\tau = \sum_{j=1}^{n} \left(\frac{[1 - S(i,j)/S(i,max)]}{n-1} \right)$$

where n is the total number of samples (63 or 42) and S(i,max) is the highest log2-transformed expression value for gene i across the n organ types. High values of expression specificity indicate genes with expression limited to few organ or tissue types or developmental stages, while low values of expression specificity indicate broad expression of genes with similar expression levels in most of the organ or tissue types and developmental stages. To test if there is any significant difference of expression specificity between any two of the four genes, we applied 1,000 Monte Carlo randomization tests to each two-gene comparison. For the Monte Carlo randomization test, we computed the following statistic: DIF = |τGENE1 − τGENE2|, where DIF indicates the absolute difference of expression specificity between two genes. Then, we compared the observed value (DIFobs) against the null distribution of simulated DIF value (DIFsim) from 1,000 randomized data. If the null hypothesis is rejected, the expression specificity of any two compared genes is significantly different. The cutoff of the significant P value was set to 0.05.

In addition to the comparison of expression specificity among gene pairs, we applied the Pearson correlation analysis to determine if the expression profile between any two genes showed any evidence of coexpression (i.e. correlated expression across different organ types or

tissue types). Coexpression is determined when the Pearson correlation coefficient (r) is significantly positive, and vice versa.

4.2.2 Inferring the Ancestral Expression States Using RT-PCR

Total RNA samples of Arabidopsis, *Tarenaya hassleriana* (formerly known as *Cleome spinosa*), *Carica papaya*, and *Vitis vinifera* were extracted from liquid N2-frozen tissue of five organ types: root, stem, leaf (rosette leaves in Arabidopsis), flower, and seed (whole siliques in Arabidopsis and *T. hassleriana*). A modified CTAB method was used for RNA extraction (Zhou et al., 2011). The quality of each RNA sample was checked on 2% agarose gels by electrophoresis, and the amount of each RNA sample was determined by a Nanodrop spectrophotometer. After DNaseI (Invitrogen) treatment to remove residual DNA, M-MLV reverse transcriptase (Invitrogen) was applied to the RNA samples to generate cDNA, according to the manufacturer's instructions. PCR was performed with cDNA templates to detect the organ-specific expression of Arabidopsis *FIS2*/*VRN2* and *MEA*/*SWN* paralogous pairs as well as orthologous genes in outgroup species for inference of the ancestral, preduplication expression states. Gene-specific primers were designed to amplify 250 to 1,000 bp of the cDNA of targeted genes (Table 4.1). For PCR, the cycling program was as follows: preheating at 94°C for 3 min; 30 to 35 cycles of denaturing at 94°C for 30 s, annealing at 53°C to 56°C for 30 s, and elongation at 72°C for 30 s or 1 min; and a final elongation at 72°C for 7 min. PCR products were checked on 1% agarose gels and sequenced to confirm identity.

4.2.3 Identifying Epigenetic Marks Associated with the Studied Genes

We investigated the epigenetic modifications around the genomic regions of Arabidopsis *FIS2*, *VRN2*, *MEA*, and *SWN*. We also used *EMF2* and *CLF*, which are members of the *FIS2*/*VRN2* and *MEA*/*SWN* families, respectively, to help assess the ancestral state. For DNA methylation, we obtained data from Schmitz et al. (2013), Stroud et al. (2013), and Zemach et al. (2013) from CoGe (https://genomevolution.org/CoGe/), visualized by JBrowse in Araport (https://www.araport.org/). Analyzed data included assayed genomic DNAs from leaves from Schmitz et al. (2013) and Stroud et al. (2013) and assayed genomic DNAs from seedlings and roots from Zemach et al. (2013), which were all vegetative organs. Cytosine methylation at CpG sites was analyzed along the genomic region of a target gene. For histone methylation, we extracted tiling array data from seedlings from Roudier et al. (2011) and chromatin immunoprecipitation-on-chip data from wild-type and fie mutant seedlings from Bouyer et al. (2011). Four histone marks were analyzed: H3K27me3, H3K4me3, H3K4me2, and H3K36me3. The epigenetic features in Arabidopsis seedlings were compared among the paralogous genes in a family and between the two interacting gene families.

4.2.4 Detecting Accelerated Sequence Evolution and Positive Selection by Ka/Ks Analyses

To analyze the selection acting on the gene pairs *FIS2*/*VRN2* and *MEA*/*SWN*, several rate analyses were performed using Codeml in the PAML package (Yang, 2007). We obtained the sequences of the four genes from Arabidopsis as well as some other Brassicaceae species,

including *Arabidopsis lyrata*, *Arabidopsis halleri*, *Capsella rubella*, *Brassica rapa*, *Brassica oleracea*, *Eutrema salsugineum* (formerly known as *Thellungiella halophila*), and *Schrenkiella parvula* (formerly known as *Thellungiella parvula*). We also identified orthologous sequences, by reciprocal best BLAST hits, from species outside of the Brassicaceae, including *T. hassleriana* (formerly known as *C. spinosa*), *C. papaya*, *Gossypium raimondii*, *Theobroma cacao*, *Citrus sinensis*, *Populus trichocarpa*, *Ricinus communis*, *Manihot esculenta*, and *V. vinifera*, from PLAZA version 3.0 Dicots (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/; Proost et al., 2015), Phytozome version 10 (http://phytozome.jgi.doe.gov/pz/portal.html; Goodstein et al., 2012), the BRAD database (http://brassicadb.org/brad/; Cheng et al., 2011), and NCBI's GenBank. Gene orthology was later confirmed by comparing the topology of the gene phylogeny with the species tree. Alignments of amino acid sequences were generated using MUSCLE under default parameters (Edgar, 2004) and then reverse translated into codon alignments using the customized Perl script. We generated the alignments for the full length of the two gene families as well as some documented functional domains, including the VEF and C2H2 domains in the FIS2 and VRN2 genes and the C5, SET, SANT, and CXC domains in the *MEA* and *SWN* genes. Phylogenies of the two gene families were analyzed by RAxML version 7.0.3 with GTR as the substitution matrix (Stamatakis, 2006). Maximum likelihood trees of the two gene families were generated based on codon alignments.

We first used a phylogeny-based free-ratio test to estimate branch-wise Ka/Ks ratios along the phylogenetic tree branches. For the full-length *FIS2*/*VRN2* genes, we implemented four different models to test if the Ka/Ks ratios of the Brassicaceae *FIS2* clade and the Brassicaceae

VRN2 clade display an asymmetric pattern and how conserved they are compared with the orthologous genes. The first model (model I, one-ratio model) assumes that all the genes have the same Ka/Ks ratio, bearing the hypothesis that all genes are under the same level of selection. The second model (model II, two-ratio model-1) assumes that the Brassicaceae *VRN2* clade and the orthologous genes have the same Ka/Ks ratio but the Brassicaceae *FIS2* clade can have a different one, suggesting that the Brassicaceae *VRN2* clade reflects the ancestral selection but *FIS2* evolved in a different manner. The third model (model III, two-ratio model-2) assumes that the duplicated *FIS2* and *VRN2* clades in Brassicaceae have the same Ka/Ks ratio while the orthologs can have a different ratio, which is a hypothesis that the two Brassicaceae copies evolved at the same rate. The fourth model (model IV, three-ratio model) assumes that the two Brassicaceae branches have different Ka/Ks ratios and, thus, the two genes evolved at different rates, with the third Ka/Ks ratio for the orthologous branches. A set of likelihood ratio tests was applied, where twice the difference of likelihood values was calculated and compared against a $\chi 2$ distribution with the degrees of freedom set at 1: comparison between model II and model IV can tell if the selection on the Brassicaceae *VRN2* is significantly different from the orthologous genes; and comparisons between model I and model II, as well as between model III and model IV, are used to see if the selection on the Brassicaceae *FIS2* is different from the Brassicaceae *VRN2* and/or the orthologous genes. When model II fits better than model I and model IV fits better than model III with statistical support, the evolutionary rate of the duplicated pair in Brassicaceae is considered to evolve asymmetrically. The same analyses were performed on the functional domains of the *FIS2*/*VRN2* genes and the full-length *MEA*/*SWN* genes and their functional domains (Table 4.2). We also applied a branch-site model

to detect positively selected sites along *FIS2* as well as *MEA*. Test 2 of model A with the Bayes Empirical Bayes analysis was applied to identify amino acid sites with a high posterior probability of positive selection (Zhang et al., 2005).

4.3 Results

4.3.1 *FIS2* and *MEA* Have Specific and Similar Expression Patterns in Reproductive Organs

FIS2 and *MEA* formed by the α WGD that is specific to the Brassicaceae family after the divergence of the Brassicaceae lineage from the Caricaceae lineage. After gene duplication, duplicated genes may experience expression divergence. We analyzed microarray data in Arabidopsis to compare the expression profiles of paralogous interacting gene pairs *FIS2*/*VRN2* and *MEA*/*SWN*. We obtained two sets of ATH1 microarray data and analyzed them separately: 63 different organ types and developmental stages (Schmid et al., 2005), referred to as the ADA (Arabidopsis developmental atlas) data set hereafter, and 42 different tissue types during seed developmental stages (Le et al., 2010; Belmonte et al., 2013), referred to as the ASA (Arabidopsis seed atlas). We first calculated the expression specificity (τ) of the four genes defined by Yang and Gaut (2011). *VRN2* and *SWN* have expression specificity values of 0.19 and 0.17, respectively, indicating that both genes have relatively broad expression in nearly all organ types included in the ADA data set (Fig. 4.2). In contrast, *FIS2* has an expression specificity value of 0.7 and that of *MEA* is 0.63, indicating an organ-specific expression pattern. We observed that the expression of *FIS2* and *MEA* is restricted to flowers and siliques, and the

absence of vegetative expression explains the high expression specificity. Yang and Gaut (2011) analyzed the ADA data set and found that the recent whole-genome duplicates have a median τ close to 0.2. Thus, what we observed for *FIS2* and *MEA* is quite high, and what we observed for *VRN2* and *SWN* is about average.

We also analyzed τ in the ASA data set (Fig. 4.2A). Similarly, the τ of *FIS2* is 0.48 and that of *MEA* is 0.56, while *VRN2* has τ of 0.21 and *SWN* has τ of 0.22. *FIS2* and *MEA* turn out to show more tissue-specific expression in seed tissues. We broke down the ASA data and observed that *FIS2* and *MEA* tend to be expressed in the triploid endosperm rather than in the diploid embryo or maternally derived seed coat. We did a 1,000-replicate permutation test and gained statistical support that the expression specificity differences in the *FIS2*-*MEA* and *VRN2*-*SWN* comparisons are not significant (Fig. 4.3), indicative of the concerted divergence in their expression profiles. In contrast, the tissue specificity expression profiles are significantly different in the two duplicated pairs, *VRN2*-*FIS2* and *SWN*-*MEA*, indicative of their regulatory divergence.

Not only did we analyze the expression index for those genes individually, we also performed a correlation test to examine the association of the expression profiles of the four genes, as their products function in a complex (Fig. 4.2B). We found that the expression patterns of *FIS2* and *MEA* are positively correlated in both the ADA and ASA data sets, while broadly expressed VRN2 and SWN are coexpressed. However, the expression of both *FIS2* and *MEA* is negatively correlated to the expression of *VRN2* and *SWN*. The negative coefficients are around −0.5 (Fig. 4.2B), which is below 1% of the total α whole-genome pairs analyzed by Blanc and Wolfe

(2004). Overall, the *FIS2-MEA* expression patterns indicate parallel divergence from the *VRN2-SWN* expression patterns in a concerted manner.

4.3.2 *FIS2* and *MEA* Acquired New Expression Patterns

As the microarray data from the ADA indicated, *FIS2* and *MEA* both have expression patterns that are restricted to reproductive organs, such as flowers and siliques, but not vegetative organs, including roots, stems, and leaves. We confirmed this result with reverse transcription (RT)-PCR (Fig. 4.4). In contrast, their paralogs, *VRN2* and *SWN*, have a broad expression pattern in both vegetative and reproductive organs, and they are expressed ubiquitously in all examined organ types in our RT-PCR results (Fig. 4.4). To infer the ancestral expression patterns of the two gene pairs, we assayed the expression patterns of orthologs in *Tarenaya hassleriana* (formerly known as *Cleome spinosa*), *Carica papaya*, and *Vitis vinifera*. Among those species with sequenced genomes, *T. hassleriana* belongs to Cleomaceae, the most closely related sister group to Brassicaceae. Although *T. hassleriana* had its own genome triplication after the divergence between Cleomaceae and Brassicaceae (Cheng et al., 2013), only a single copy each of the orthologous *VRN2* and *SWN* has been retained. *C. papaya* also is in the order Brassicales. *V. vinifera* was chosen because its lineage has not experienced any WGD events since the γ WGD during early eudicot evolution, which applies to *C. papaya* as well; thus, genes are frequently single copy in these taxa. These single copy orthologs can facilitate the inference of ancestral expression patterns. We confirmed that these sequences are true orthologs of *FIS2*/*VRN2* and *MEA*/*SWN* by phylogenetic analysis of the gene families.

For both the *FIS2*/*VRN2* and *MEA*/*SWN* pairs, their orthologs in *T. hassleriana*, *C. papaya*, and *V. vinifera* are widely expressed in all examined organ types, which is the same as *VRN2* and *SWN* in Arabidopsis (Fig. 4.4). The absence of expression in vegetative organs is observed only in *FIS2* and *MEA*. Collectively, we inferred that the preduplicated expression state is likely to be a broad expression pattern, which is reflected by *VRN2* and *SWN*. The Brassicaceae *FIS2* and *MEA* both lost expression in vegetative organs to become expressed specifically in reproductive organs.

4.3.3 *FIS2* and *MEA* Acquired Novel Epigenetic Modifications

The epigenetic features of cytosine methylation and histone methylation often are associated with the expression or silencing of genes. To examine the patterns of cytosine and histone methylation in organ types where the expression of *FIS2* and *MEA* was lost, we investigated the epigenetic variation among these genes in vegetative tissues, including leaves, roots, and seedlings, of Arabidopsis (for details, see "Materials and Methods"). For DNA methylation, we found that cytosine methylation at CpG sites is enriched in the promoter region (defined as 1,500 bp upstream of the transcription start site) of the *FIS2* genomic sequence but not in the gene body (Fig. 4.5). The opposite is found for *VRN2*, with the promoter region unmarked but the gene body highly methylated (Fig. 4.5). The same divergence of DNA methylation was found for *MEA* and *SWN* (Fig. 4.5). Cytosine methylation is enriched in the promoter region of *MEA* but only in the gene body of *SWN*. The DNA methylation patterns in *EMBRYONIC FLOWER 2* (*EMF2*) and *CURLY LEAF* (*CLF*), the more distant paralogs of *VRN2* and *SWN*, respectively, are

also gene body enrichment, the same as *VRN2* and *SWN*, suggesting that the pattern of DNA methylation for *FIS2* and *MEA* changed after duplication. As promoter cytosine methylation is associated with transcriptional repression and gene body methylation is indicative of expression activation (Suzuki and Bird, 2008), this finding is consistent with the expression data. We did not examine methylation patterns in whole endosperm, because in the ASA data set, *FIS2* and *MEA* showed variable expression patterns in different parts of the endosperm and different developmental stages.

We also examined histone methylation in the region of these genes in the seedlings of Arabidopsis based on the data generated by Roudier et al. (2011). Similar to DNA methylation, we found that *VRN2*, *SWN*, *EMF2*, and *CLF* have the same types of histone methylation, which are different from *FIS2* and *MEA* (Table 4.3). We noticed that *FIS2* and *MEA* lost trimethylation of Lys-4 on histone H3 (H3K4me3), which is shared by all the other genes. Instead, they gained a novel mark of H3K27me3. H3K4me3 is an activating mark, while its antagonistic mark, H3K27me3, is repressive. This could help explain the expression of *VRN2*, *SWN*, *EMF2*, and *CLF* in the vegetative tissue but the lack of expression of *FIS2* and *MEA*. It is also notable that, in the fie mutant, where the PRC2 function was supposed to be abolished, *FIS2* and *MEA* lost their H3K27me3 but, instead, *VRN2*, *SWN*, *EMF2*, and *CLF* were marked by H3K27me3 (Bouyer et al., 2011). As H3K27me3 is regulated by PRC2 complexes, this finding suggests the self- and cross-regulation among these genes. With both DNA and histone modification comparative analyses, we observed the convergent evolution of epigenetic features in *FIS2* and *MEA*, divergent from their preduplicated and postduplicated paralogs.

4.3.4 Gene Structural Changes in *FIS2* and *MEA*

FIS2 formed from *VRN2* by duplication, and *MEA* duplicated from *SWN*, during the α WGD. *FIS2* in Arabidopsis lost three exons, called the E15-E17 region (corresponding to the 15th to 17th exons in Arabidopsis *EMF2*, not named after *VRN2*), compared with *VRN2* (Fig. 4.6A; Chen et al., 2009). FIS2 has a large Ser-rich domain that is not shared with any other VEF genes in any species, indicating gain of the domain in Brassicaceae (Fig. 4.6A; Chen et al., 2009). Our sequence analysis showed that the Ser-rich domain is highly variable among FIS2 sequences from different Brassicaceae species (Fig. 4.6A). The lost E15-E17 domain and the gained Ser-rich domain are both neighboring the VEF domain that interacts with the C5 domain in MEA.

MEA is about 150 amino acids shorter than SWN, and the deleted region is just downstream of the C5 domain that interacts with the VEF domain in FIS2, due to a large shrinkage in a single exon (the ninth in Arabidopsis MEA and SWN) where Brassicaceae SWN and orthologous SWN-like sequences are not conserved (Fig. 4.6B). How the structural changes affect the physical interaction of FIS2 and MEA remains to be tested. In addition to the rearrangement of functional domains, those shared domains show different levels of amino acid sequence divergence. In contrast, VRN2/EMF2-like sequences and SWN/CLF-like sequences show relative conservation across all flowering plants in amino acid sequences and functional domains (Chen et al., 2009; Qian et al., 2014).

4.3.5 *FIS2* and *MEA* Show Accelerated Amino Acid Substitution Rates and Evidence for Positive Selection

Duplicated genes diverge not only in expression pattern but also in their sequences. We first analyzed by Ka/Ks analysis (the ratio of the number of nonsynonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site) the full-length coding regions of *FIS2*, *VRN2*, *MEA*, and *SWN* genes (Fig. 4.7; Fig. 4.8). The Brassicaceae *FIS2* clade had a much higher average Ka/Ks than *VRN2* lineages, 3.5-fold greater than the paralogous Brassicaceae *VRN2* clade and 10-fold greater than the orthologous preduplicate *VRN2* sequences. Similarly, the Brassicaceae *MEA* clade had a high average Ka/Ks comparable to the *FIS2* clade, which is 3.5-fold greater than the paralogous Brassicaceae *SWN* clade and 4.5-fold greater than the orthologous preduplicated *SWN* sequences. We implemented different models assuming similar versus different Ka/Ks ratios in these clades, described in "Materials and Methods," and the likelihood ratio tests indicated that the divergence in sequence rate is significant (Table 4.2). These analyses indicate that, while the paralogous Brassicaceae *VRN2* and *SWN* lineages are under stronger purifying selection along with the orthologous genes in outgroup species, *FIS2* and *MEA* in the Brassicaceae have experienced relaxation of purifying selection. Asymmetric Ka/Ks ratios are seen in a minority of duplicated gene pairs in Arabidopsis; for example, Gossmann and Schmid (2011) estimated that 7% of the duplicated pairs they analyzed have asymmetric Ka/Ks ratios.

Additionally, among the branch-wise Ka/Ks of specific *FIS2* and *MEA* sequences, we detected possible positive selection, indicated by Ka/Ks greater than 1, acting on the sequences from certain lineages (Fig. 4.8). In order to distinguish certain amino acid sites evolving under positive selection from relaxed purifying selection, we also applied a branch-site model, which suggested that both branches leading to Arabidopsis *FIS2* ($P < 0.0001$) and *MEA* ($P = 0.007$) have positively selected amino acid sites across different functional domains (Fig. 4.9).

Thus, we further studied the sequence evolution of characterized functional domains of *FIS2*/*VRN2* and *MEA*/*SWN* genes, including the VEF and C2H2 domains in the FIS2 and VRN2 genes and the C5, SET, SANT, and CXC domains in the *MEA* and *SWN* genes (Fig. 4.7; Fig. 4.8). We observed that the trend of acceleration in sequence evolution of *FIS2* and *MEA*, and the evolutionary constraint resulting in the conservation of *VRN2* and *SWN*, were reflected by all the functional domains we analyzed individually. The VEF domain in *FIS2*/*VRN2* genes and the C5 domain in *MEA*/*SWN* genes interact physically with each other; thus, the comparison between the two sets of Ka/Ks ratios best describes the coevolution between *FIS2* and *MEA* at the coding sequence level from a protein-protein interaction perspective (Fig. 4.7). Consistent with the full-length gene analyses, the VEF domain in the FIS2 lineages and the C5 domain in the MEA lineages both have accelerated amino acid substitution rates, with evidence (Ka/Ks > 1) suggesting positive selection on a few branches (Fig. 4.8; Table 4.2). Similar results were found in the DNA binding-related domains, C2H2 in FIS2/VRN2 and CXC and SANT in *MEA*/*SWN* genes (Fig. 4.8), indicating that the PRC2 complexes with FIS2 and MEA may have affinity to specific DNA regions, regulating a novel network of gene expression. The SET domain plays the

role of methyltransferase in the PRC2 complex and is usually highly conserved across eukaryotes (Baumbusch et al., 2001). This is reflected by the low Ka/Ks ratios detected in the SWN SET domains (Fig. 4.8). Instead, the SET domain in the Brassicaceae MEA shows evidence for positive selection (Fig. 4.9). The rapid amino acid substitution rates in the PRC2 functional domains together likely relate to the functional divergence of the PRC2 complexes containing FIS2 and MEA.

4.3.6 VEL2 and VEL1, Which Interact with PRC2 Complexes, Show Corresponding Divergence Patterns to FIS2/VRN2 and MEA/SWN

A family of five PHD finger proteins is necessary for the core PRC2 complex to maintain the repressed status of chromatin (Kim and Sung, 2010, 2013). Among them, *VERNALIZATION5*/*VIN3-LIKE1* (*VEL1*) and *VEL2* are a pair of α whole genome duplicates. VEL2 is a maternally expressed imprinted gene (Wolff et al., 2011). We analyzed their expression profiles in the ADA and ASA microarray data sets and detected that *VEL1* shows a coexpression pattern with VRN2 and SWN that is similar to the broadly expressed *VEL* homologs, whereas *VEL2* has a similar expression pattern to *FIS2* and *MEA* due to the loss of vegetative expression (Fig. 4.10A; Qiu et al., 2014). *VEL2* has a higher specificity than its paralog *VEL1* (Fig. 4.10B). Thus, the observed concerted divergence in expression pattern in the FIS complex is not limited to the core complex but also includes other associated proteins.

For cytosine methylation in the vegetative tissue, *VEL1* is marked through the coding exons but not the promoter region, whereas *VEL2* has cytosine methylation enriched in the upstream promoter region and the first two introns located between the 5' untranslated region exons (Schmitz et al., 2013; Stroud et al., 2013; Zemach et al., 2013). For histone methylation in the vegetative tissue, *VEL1* is marked by activating marks, including H3K4me3, trimethylation of Lys-36 on histone H3 (H3K36me3), and dimethylation of Lys-4 on histone H3 (H3K4me2; Roudier et al., 2011). *VEL2* has lost the H3K4me3 and H3K36me3 but gained the repressive mark H3K27me3. Those epigenetic features not only correspond to the vegetative expression level but also are consistent with the divergence of the core PRC2 components FIS2 and MEA (Table 4.3).

We further analyzed the sequence evolution of the *VEL* genes. The *VEL2* sequences have an elevated average Ka/Ks ratio compared with the Brassicaceae *VEL1* and orthologous *VEL* genes (Fig. 4.10B). While *VEL1* and orthologous sequences have a low Ka/Ks ratio close to 0, indicating strong purifying selection, a 3-fold change in *VEL2* sequences suggests the relaxation of purifying selection. This coincides with the accelerated amino acid substitution rates of *FIS2* and *MEA*.

4.4 Discussion

4.4.1 Concerted Divergence of FIS2 and MEA in the FIS-PRC2 Complex

Upon gene duplication, hypothetically, two duplicates are identical in function as well as expression pattern if the cis-elements also are entirely duplicated. Considering that many proteins function through interactions with other proteins, in a regulatory or metabolic pathway, through protein-protein interaction, or form an integral complex, either the duplicates are redundant or both duplicates could integrate into either complex and affect the function of the complex if they have divergence. A shift in expression pattern would be one way to avoid potentially disadvantageous cross talk between interacting members (Aakre et al., 2015). Blanc and Wolfe (2004) described a process of concerted divergence of gene expression in Arabidopsis, in which pairs of duplicates, whose protein products interact, diverge in a parallel manner in expression pattern. However, as *FIS2* and *VRN2* were not identified as α whole-genome duplicates by the genome-wide study (Blanc et al., 2003), their concerted divergence in expression pattern with *MEA* and *SWN* was not included.

Here, we show that *FIS2* and *MEA* diverged in expression pattern in a concerted manner, modified from coexpressed *VRN2* and *SWN*, whose expression pattern resembles the ancestral status. In addition, we show that cytosine methylation and histone methylation patterns in *FIS2* and *MEA* also diverged in a concerted manner. It is possible that the methylation change contributed to the changes in expression patterns, although mutations in regulatory elements

also may have played a role in the expression pattern changes. *FIS2* and *MEA* are marked by H3K27me3 in the vegetative tissue, suggesting they both became the targets of a vegetative PRC2 complex after formation by gene duplication (Bouyer et al., 2011). In addition to the vegetative epigenetic divergence, *FIS2* and *MEA* are well known as imprinted genes during seed development, both of which are maternally expressed genes (Berger and Chaudhury, 2009). Based on the genome-wide data sets from Hsieh et al. (2011) and Gehring et al. (2011), we determined that *VRN2* and *SWN* are not imprinted, while the more distant relatives in their gene families, *EMF2* and *CLF*, also lack evidence for imprinting. Thus, we infer that *FIS2* and *MEA* became imprinted genes after their divergence from *VRN2* and *SWN*. This concerted change in the regulation of both genes ensures the dosage balance between the interacting proteins. The concerted divergence of *FIS2* and *MEA* from their paralogs also is reflected by the elevated Ka/Ks ratios in the coding sequences at comparable levels, suggesting that similar relaxed purifying selection is acting on the two genes. Altogether, these changes indicate that *FIS2* and *MEA* have been diverging in concert in multiple ways, which likely contributed to the divergence in functions between the FIS-PRC2 complex and the VRN-PRC2 complex.

4.4.2 Functional Divergence in the FIS-PRC2 Complex

VRN2, SWN/CLF, FIE, and MSI1 form the VRN complex, which regulates vernalization to control flowering time in Arabidopsis (Fig. 4.1; Hennig and Derkacheva, 2009). The complex also represses autonomous seed coat development (Roszak and Köhler, 2011). The FIS complex contains FIS2, MEA, FIE, and MSI1. The FIS complex is important in gametophyte and seed

development and has two major functions. A prefertilization role for the FIS complex is that it prevents proliferation of the central cell of the female gametophyte until after fertilization, so that seed development does not start until after fertilization (Hennig and Derkacheva, 2009). The FIS complex also acts postfertilization. It is needed for regulating endosperm cellularization during seed development (Hehenberger et al., 2012). *FIS2* mutants show a phenotype of abnormal female gametophyte development into embryos and are defective in controlling central cell proliferation in the female gametophyte, suggesting that FIS2 is not redundant with VRN2 in the prefertilization function (Roszak and Köhler, 2011). Thus, the FIS complex function in the female gametophyte is specific to the FIS complex and not the VRN complex. MEA also was shown to not be redundant with SWN (Roszak and Köhler, 2011). Unlike all the key components in the FIS complex, a *SWN* mutant failed to lead to autonomous seed development in the absence of fertilization, nor to seed abortion with embryo and endosperm overgrowth (Luo et al., 1999); thus, it is possible that MEA is functionally specialized for the prefertilization function of the FIS complex and cannot be complemented by SWN. As for the postfertilization function, SWN was shown to be not essential in seed development (Spillane et al., 2007). Thus, it was proposed that *MEA* underwent neofunctionalization to gain a postfertilization role in regulating seed development after its duplication from *SWN* (Spillane et al., 2007).

Taking the two parts of the FIS complex functions together, it appears that the novel PRC2 made up by FIS2 and MEA created a Brassicaceae-specific complex for preventing seed development prior to fertilization and facilitating seed development after fertilization in Brassicaceae. This functional divergence complements the concerted divergence of FIS2 and

MEA in other ways that we show in this study. The FIS complex also plays an important role in establishing the imprinted expression of many genes in the endosperm, especially paternally expressed imprinted genes, as the differentially methylated paternal or maternal allele can affect the targeting by this complex (Wolff et al., 2011; Köhler et al., 2012). The concerted divergence of *FIS2* and *MEA* in expression patterns, methylation patterns, and accelerated sequence evolution may have contributed to functional diversification or, potentially, neofunctionalization of the FIS-PRC2 complex. An alternative to neofunctionalization of the FIS-PRC2 complex is subfunctionalization after the formation of *FIS2* and *MEA* from their paralogs. Without knowledge of the ancestral function of the PRC2 complex in plants closely related to the Brassicaceae, discussed below, we cannot say for sure if there has been neofunctionalization or subfunctionalization. We show in this study that there has been regulatory neofunctionalization of *FIS2* and *MEA*, which leads us to favor the possibility of neofunctionalization of the complex. Nonetheless, under a scenario of subfunctionalization, *FIS2* and *MEA* still show concerted divergence in their expression patterns, cytosine and histone methylation, and accelerated sequence evolution. In order to distinguish the two possible hypotheses, more research on VRN complexes in rosid species will provide valuable information to indicate the function of the ancestral rosid PRC2 complex.

How are the FIS complex functions performed in other angiosperms outside of Brassicaceae? Some clues come from studies of FIE, which is a member of the FIS complex, in *Hieracium pilosselloides* (Asteraceae). The central cell proliferation phenotype of Arabidopsis *fie* mutants is not seen in sexual *H. pilosselloides* FIE RNA interference lines; thus, a PRC2 complex does not

regulate central cell proliferation in the female gametophyte of *H. piloselloides*, in contrast to Arabidopsis (Rodrigues et al., 2008). This might indicate that parts of the prefertilization function of FIS-PRC2 in Brassicaceae is an evolutionary innovation; at the same time, it is possible that the unknown mechanism repressing central cell proliferation is specific to the *H. piloselloides* lineage. *FIE* down-regulation in *H. piloselloides* leads to seed abortion (Rodrigues et al., 2008); thus, FIE is important for seed development, presumably as part of a PRC2 complex. Asterids do not contain FIS2, VRN2, or MEA. Thus, if there is a PRC2 complex regulating seed development in asterids, it probably contains the product of lineage-specific polycomb proteins and a mechanism independently evolved from Brassicaceae. In maize (*Zea mays*) and rice (*Oryza sativa*), there has been duplication of *FIE* (Luo et al., 2009; Li et al., 2014). Thus, the grasses may have PRC2 complexes that are divergent from the ancestral state. The requirement of H3K27me3 in rice and maize endosperm for the establishment of imprinting suggests the functional conservation or convergence of a PRC2 complex in Brassicaceae and Poaceae (Makarevitch et al., 2013; Zhang et al., 2014).

4.4.3 Evolution of Protein Complexes after the Duplication of Components

We propose a model of simultaneous gene duplication and concerted divergence of one copy of each duplicated pair (Fig. 4.11). Following formation by duplication, two genes whose products function together in a complex diverge in similar ways, and the complex diverges in function. This divergence pattern is not limited to neofunctionalization/subfunctionalization but includes some other modifications of these scenarios, such as escape from adaptive conflict. To

our knowledge, the PRC2 complexes in Brassicaceae we examined in this study provide the first example of this type of divergence of duplicated genes. We contrast this scenario with single-gene duplication and divergence, where one component in the complex undergoes gene duplication and then the paralog diverges, driving the two complexes with either paralog to diverge in function as a result. Intuitively, many described functionally divergent paralogs may contribute to this type of divergence of their protein complexes. One example is the centromere-defining histone variant CENH3 in the histone core octamers that show duplication specific to the genus *Mimulus* and sequence divergence, whereas other components in the histone core octamers do not show duplications specific to *Mimulus* (Finseth et al., 2015). Another case is the telomere-associated proteins POT1a and POT1b in the telomerase RNP complexes in Brassicaceae, where POT1a experienced positive selection that enhanced its affinity with interacting proteins (Beilstein et al., 2015). A variation on this model is when there is a subsequent gene duplication at a later time of another gene whose product functions in the complex, followed by divergence. An example is the plant-specific RNA polymerase IV and V, where rounds of independent lineage-specific duplications and subsequent divergence of varying kinds of subunits have increased RNA polymerase complexity and specificity among different plant groups (Wang and Ma, 2015).

4.4.4 Concerted Divergence of the Functionally Associated VELs and Some PRC2 Targets

The *VEL* genes, *VEL1* and *VEL2*, which are required to maintain and facilitate polycomb transcriptional repression, interact with the PRC2 complex but are not part of the complex

itself. Our expression, methylation, and sequence analysis results indicate that *VEL2* has similar patterns to *FIS2* and *MEA*, whereas *VEL1* has similar patterns to *VRN2* and *SWN*. Thus, *VEL2* appears to be diverging in concert with *FIS2* and *MEA*. *VEL2* also is a maternally expressed gene and regulated by the FIS complex in the endosperm (Wolff et al., 2011), and VEL2 works together with the FIS core complex to impose maternal regulation in seed development similar to FIS2 and MEA.

Several PRC2 targets duplicated through the α WGD show similar patterns of divergence as well. *PKR2* and *JMJ15* are FIS-PRC2-regulated imprinted genes (Hsieh et al., 2011; Wolff et al., 2011), whereas their paralogs, *PKL* and *JMJ18*, show broad expression, are not imprinted, and are associated with a vegetative PRC2 complex (Aichinger et al., 2011; Yang et al., 2012; Zhang et al., 2012). Out of 46 imprinted genes regulated by FIS2 (Wolff et al., 2011), we identified 41 Brassicaceae-specific duplicated genes. Some of those genes have roles in seed development, such as *PHERES1* (Köhler et al., 2003; Villar et al., 2009) and *ADMETOS* (Kradolfer et al., 2013). Thus, there are new Brassicaceae-specific genes involved in seed development that are regulated by the FIS-PRC2 complex. The functional innovation of the FIS complex appears to have rewired, to some extent, the regulatory pathway of seed development specific to Brassicaceae.

Simultaneous gene duplication events, such as polyploidy, give rise to pairs of duplicated genes that can then codiverge (Shan et al., 2009). Many of the genes that are PRC2 targets, included in the previous paragraph, were derived by the α WGD. *FIS2*, *MEA*, and *VEL2* also were derived

from that WGD. Thus, this study illustrates the potential of concerted divergence after simultaneous gene duplication to affect the functions as well as the regulation of other genes.

Table 4.1. Gene-specific primers used in this study.

Species	Genes	Primers	Sequences
Arabidopsis thaliana	*FIS2*	F	ACCACGACTCAACTAGCAATAG
		R	CTACTAACACGACGCACCTTAG
	VRN2	F	CTAGGCAACCCATCGTTTCT
		R	CATCGACTTCATCCTCGCTATC
	MEA	F	GAGAAGTATGAGCCCGAGTCTA
		R	CAGGCGGATAACGAGCATATT
	SWN	F	GGTAGTGGAAACGGAGCAATAA
		R	GATGACCAGCAGACTTTGTAGAG
Tarenaya hassleriana	*VRN2*	F	CAGAACGATCTGAGGCTAGAAG
		R	CGTGTTCTTGGCTGAAGTTATC
	SWN	F	AGAGCTCGGAAGCTGAAATAC
		R	ATTGGCCTTCTCCTCGTTTAG
Carica papaya	*VRN2*	F	GCCAATGGCGTTGGAGCAAGTAAT
		R	AGCATCGAGAAGCCCATGATTCCA
	SWN	F	TAAGGCAACAGCAGAGGATTC
		R	TCTCCTGCCACAAGCATTAC
Vitis vinefera	*VRN2*	F	AAGGCCTGTTGCAGAAAGCTATGC
		R	AATGCCTCACAAGCCCAAGGAATG
	SWN	F	CCCACCGAGAAGCAGATAAG
		R	TCTTTGCTCGACCTTGAGATAC

Table 4.2. Ka/Ks ratios under different branch models for full-length *FIS2*/*VRN2* and *MEA*/*SWN* genes and functional domains. Models are described in Methods.

Full-length *VEF* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *FIS2*	Brassicaceae *VRN2*	Orthologous *VRN2*	
Model I	one-ratio	0.21			-3444.244921
Model II	two-ratio-1	0.72	0.11		-3394.380830
Model III	two-ratio-2	0.41		0.08	-3398.101173
Model IV	three-ratio	0.71	0.20	0.07	-3384.019232

likelihood ratio test		
2ΔL		p-value
L2-L1	99.728182	0.00E+00
L4-L3	28.163882	1.12E-07

Full-length *SET* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *MEA*	Brassicaceae *SWN*	Orthologous *SWN*	
Model I	one-ratio	0.26			-14232.671512
Model II	two-ratio-1	0.64	0.15		-14105.165432
Model III	two-ratio-2	0.45		0.14	-14143.077944
Model IV	three-ratio	0.64	0.19	0.14	-14103.092763

likelihood ratio test		
2ΔL		p-value
L2-L1	255.012160	0.00E+00
L4-L3	79.970362	0.00E+00

VEF domain in *VEF* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *FIS2*	Brassicaceae *VRN2*	Orthologous *VRN2*	
Model I	one-ratio	0.25			-3656.719905
Model II	two-ratio-1	0.81	0.14		-3611.299659
Model III	two-ratio-2	0.5		0.07	-3596.160590
Model IV	three-ratio	0.82	0.31	0.07	-3586.137471

likelihood ratio test		
2ΔL		p-value
L2-L1	90.840492	0.00E+00
L4-L3	20.046238	4.44E-05

C2H2 domain in *VEF* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *FIS2*	Brassicaceae *VRN2*	Orthologous *VRN2*	
Model I	one-ratio	0.33			-961.637951
Model II	two-ratio-1	1.91	0.16		-942.313988
Model III	two-ratio-2	0.88		0.10	-943.325275
Model IV	three-ratio	1.90	0.35	0.10	-938.379267

likelihood ratio test		
2ΔL		p-value
L2-L1	38.647926	5.08E-10
L4-L3	9.892061	1.66E-03

C5 domain in *SET* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *MEA*	Brassicaceae *SWN*	Orthologous *SWN*	
Model I	one-ratio	0.22			-1062.365498
Model II	two-ratio-1	0.64	0.10		-1047.329548
Model III	two-ratio-2	0.53		0.09	-1048.169020
Model IV	three-ratio	0.64	0.21	0.09	-1046.619073

likelihood ratio test		
2ΔL		p-value
L2-L1	30.071900	4.16E-08
L4-L3	3.099894	7.83E-02

SET domain in *SET* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *MEA*	Brassicaceae *SWN*	Orthologous *SWN*	
Model I	one-ratio	0.11			-2827.475761
Model II	two-ratio-1	0.49	0.03		-2738.980584
Model III	two-ratio-2	0.28		0.03	-2773.219591
Model IV	three-ratio	0.49	0.02	0.03	-2738.754361

likelihood ratio test		
2ΔL		p-value
L2-L1	176.990354	0.00E+00
L4-L3	68.930460	1.11E-16

CXC domain in *SET* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *MEA*	Brassicaceae *SWN*	Orthologous *SWN*	
Model I	one-ratio	0.13			-1533.796582
Model II	two-ratio-1	0.50	0.04		-1496.736733
Model III	two-ratio-2	0.29		0.04	-1510.919259
Model IV	three-ratio	0.50	0.04	0.04	-1496.731740

likelihood ratio test		
2ΔL		p-value
L2-L1	74.119698	0.00E+00
L4-L3	28.375038	1.00E-07

SANT domain in *SET* genes					
		average Ka/Ks ratio			likelihood
		Brassicaceae *MEA*	Brassicaceae *SWN*	Orthologous *SWN*	
Model I	one-ratio	0.23			-1174.141534
Model II	two-ratio-1	0.67	0.12		-1157.704776
Model III	two-ratio-2	0.53		0.11	-1159.831847
Model IV	three-ratio	0.67	0.18	0.11	-1157.322338

likelihood ratio test		
2ΔL		p-value
L2-L1	32.873516	9.84E-09
L4-L3	5.019018	2.51E-02

Table 4.3. Histone methylation of studied genes. x's indicate presence of a particular type of histone methylation.

		H3K27me3	H3K4me3	H3K4me2	H3K36me3
VEF genes	***FIS2***	x			
	VRN2		x	x	x
	EMF2		x	x	x
SET genes	***MEA***	x		x	
	SWN		x	x	x
	CLF		x	x	x
VEL genes	***VEL2***	x		x	
	VEL1		x	x	x

Fig. 4.1. Two PRC2 complexes in Brassicaceae, the VRN-complex and the Brassicaceae-specific FIS-complex, arose by the alpha whole genome duplication where VRN2 duplicated to form FIS2, and SWN duplicated to form MEA.

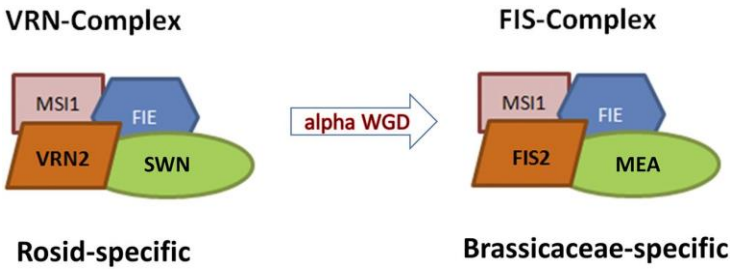


Fig. 4.2. Microarray analyses. **A.** Organ/tissue-specific expression indices based on two sets of microarray data. A large value indicates expression is restricted to fewer organ or tissue types while a low value indicates broad expression. **B.** Correlation of expression profile of each gene pair. Left: ADA set (63 organ types and developmental stages); right: ASA set (42 seed tissue types and developmental stages). Black arrows indicate a positive correlation and grey arrows indicate a negative correlation. The thickness of arrows indicates the level of the correlation coefficient. The correlation coefficient and p-value of expression profile of each gene pair are labeled along the arrows. Bold values indicate positive correlation.

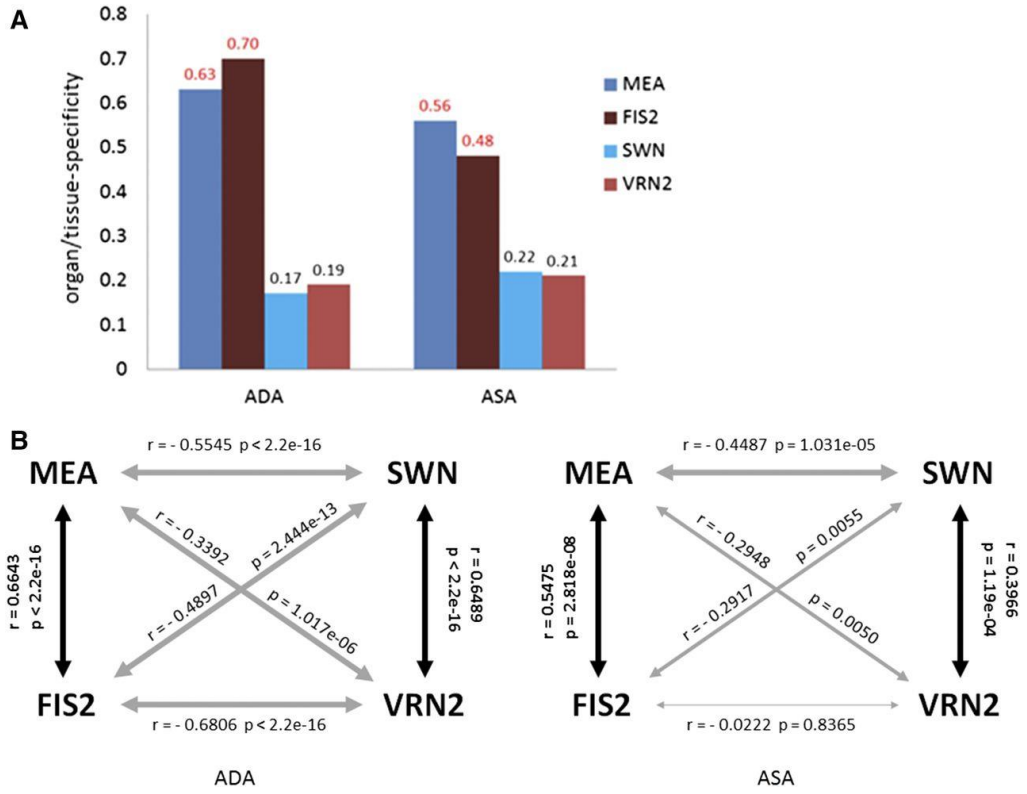


Fig. 4.3. Permutation test for microarray data to detect the difference in expression profile for all sets of comparisons of gene pairs in the ADA (A) and ASA (B) datasets. Dashed lines indicate the observed DIF.

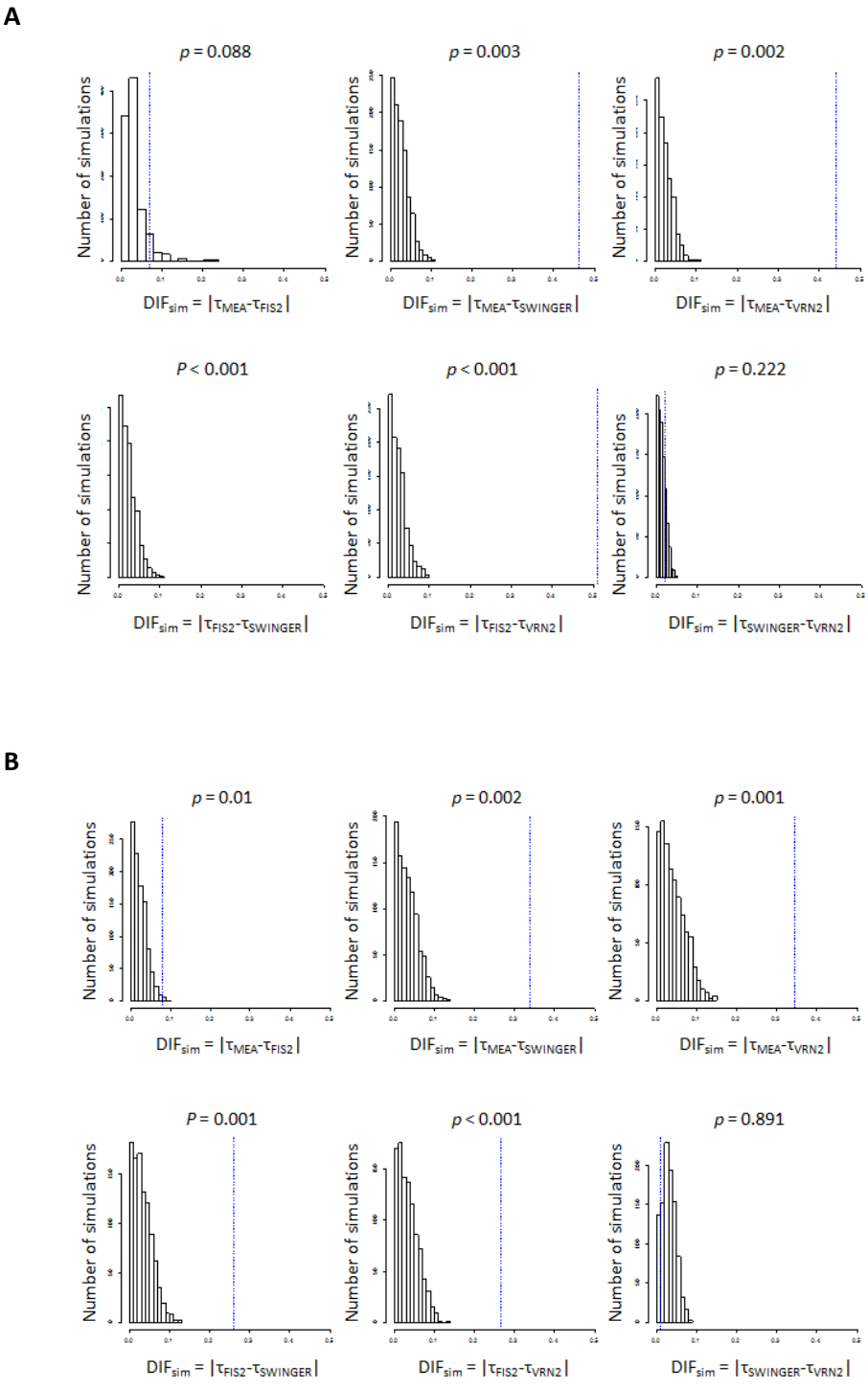


Fig. 4.4. RT-PCR assays indicate that *FIS2* and *MEA* have lost the ancestral vegetative expression pattern after duplication. Plus signs indicate reactions with reverse transcriptase and minus signs indicate controls with no reverse transcriptase. Species abbreviations include: At - *Arabidopsis thaliana*, Th - *Tarenaya hassleriana*, Cp - *Carica papaya*, and Vv - *Vitis vinifera.*

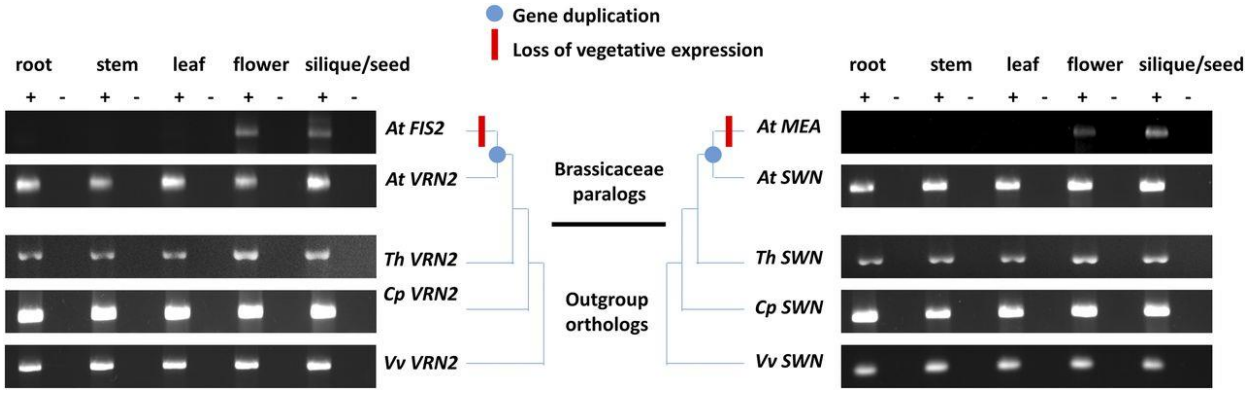


Fig. 4.5. DNA methylation at the genomic region of the VEF-domain genes and SET-domain genes. *CLF* and *EMF2* are ancient paralogs of *SWN* and *VRN2*, respectively. For each gene, four rows represent four replicates, and the dashed line separates 1500 bp upstream of the translation start codon. Vertical bars in each row represent the level of methylation.

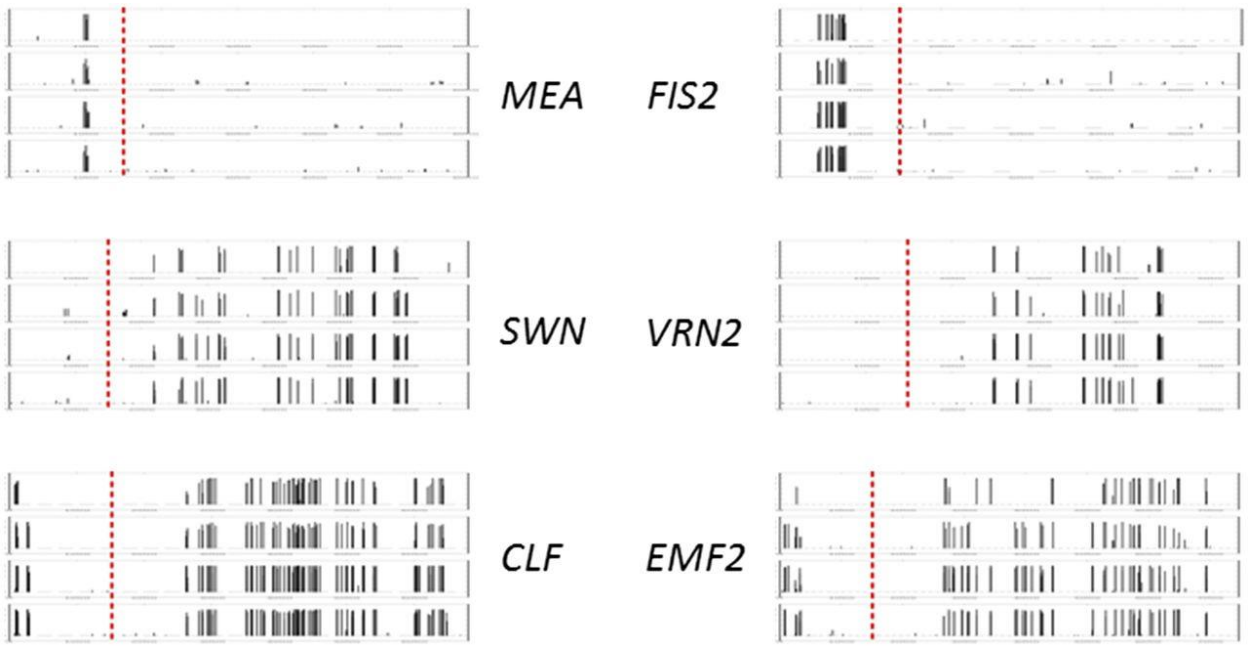
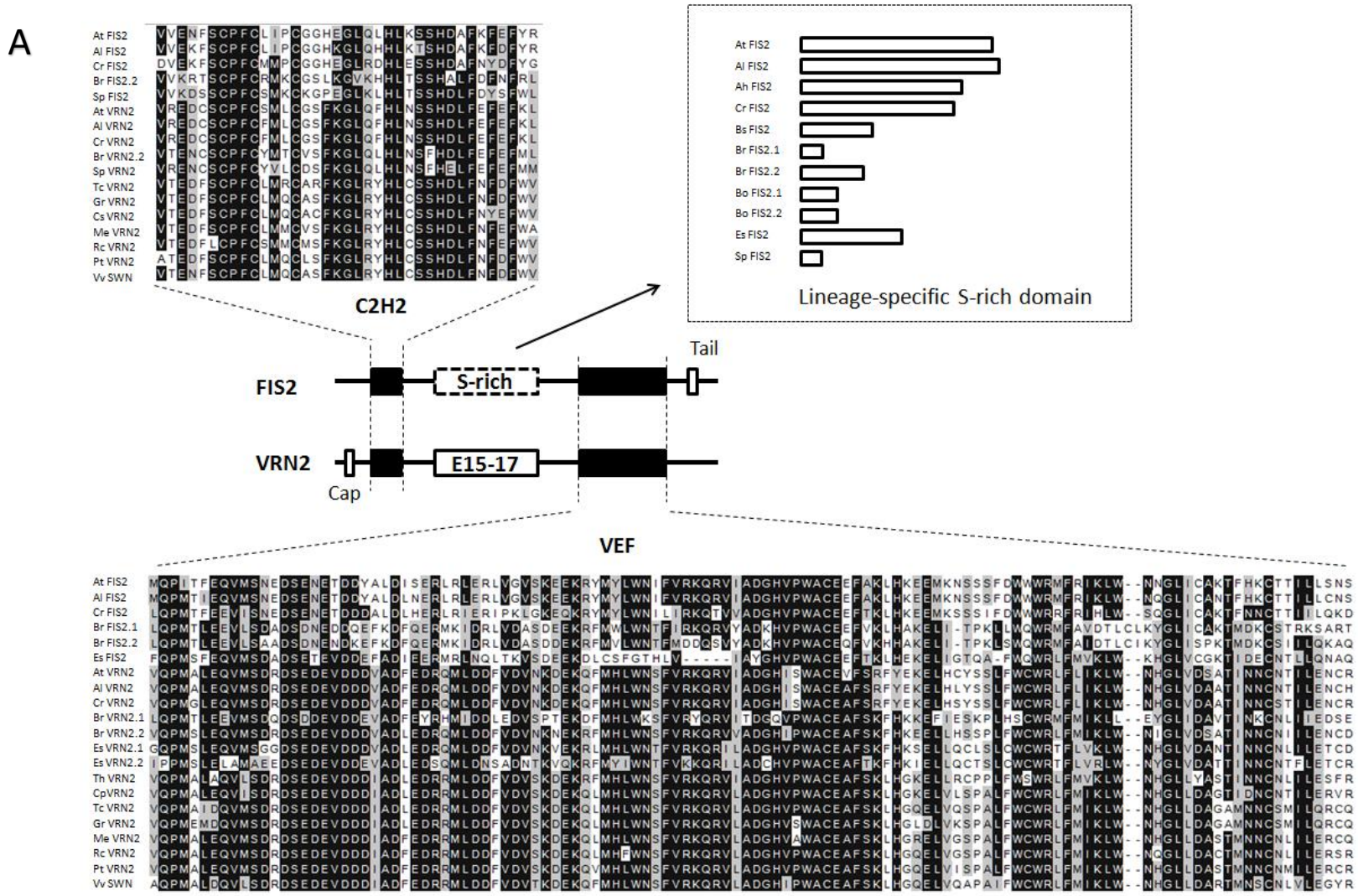


Fig. 4.6. Structures of FIS2 and VRN2 **(A)**, along with MEA and SWN **(B)** in Brassicaceae and other eurosids. Only coding regions are shown, and boxes indicate functional domains, not exons. Alignments of amino acid sequences are shown for domains analyzed in sequence rate evolution. The bars for the S-rich domain of FIS2 sequences are scaled by the length. Species abbreviations include: At - *Arabidopsis thaliana*, Al - *Arabidopsis lyrata*, Cr - *Capsella rubella*, Sp - *Schrenkiella parvula*, Es - *Eutrema salsugineum*, Br - *Brassica rapa*, Bo - *Brassica oleracea*, Th - *Tarenaya hassleriana*, Cp - *Carica papaya*, Gr – *Gossypium raimondii*, Tc - *Theobroma cacao*, Pt - *Populus trichocarpa*, Rc - *Ricinus communis*, Me - *Manihot esculenta* and Vv - *Vitis vinifera*.



B

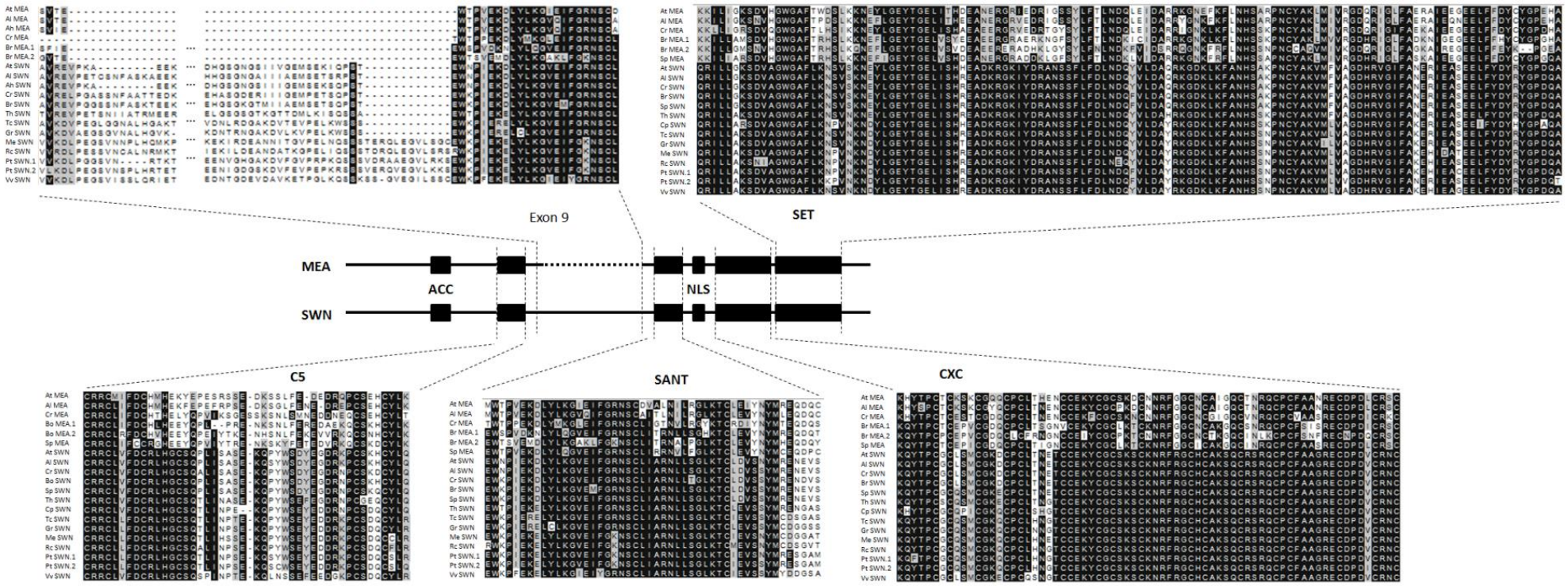


Fig. 4.7. Ka/Ks values of the interacting domains: VEF domain in FIS2/VRN2 and C5 domain in MEA/SWN. Estimated average Ka/Ks ratio of each clade is shown between the two trees. The black dots indicate the alpha WGD at the base of the Brassicaceae. The scale bars indicate 0.1 substitution per codon. Species abbreviations include: At - *Arabidopsis thaliana*, Al - *Arabidopsis lyrata*, Cr - *Capsella rubella*, Br - *Brassica rapa*, Bo - *Brassica oleracea*, Es - *Eutrema salsugineum*, Sp - *Schrenkiella parvula*, Th - *Tarenaya hassleriana,* Cp - *Carica papaya*, Gr - *Gossypium raimondii*, Th - *Theobroma cacao*, Pt - *Populus trichocarpa*, Rc - *Ricinus communis*, Me - *Manihot esculenta* and Vv - *Vitis vinifera.*

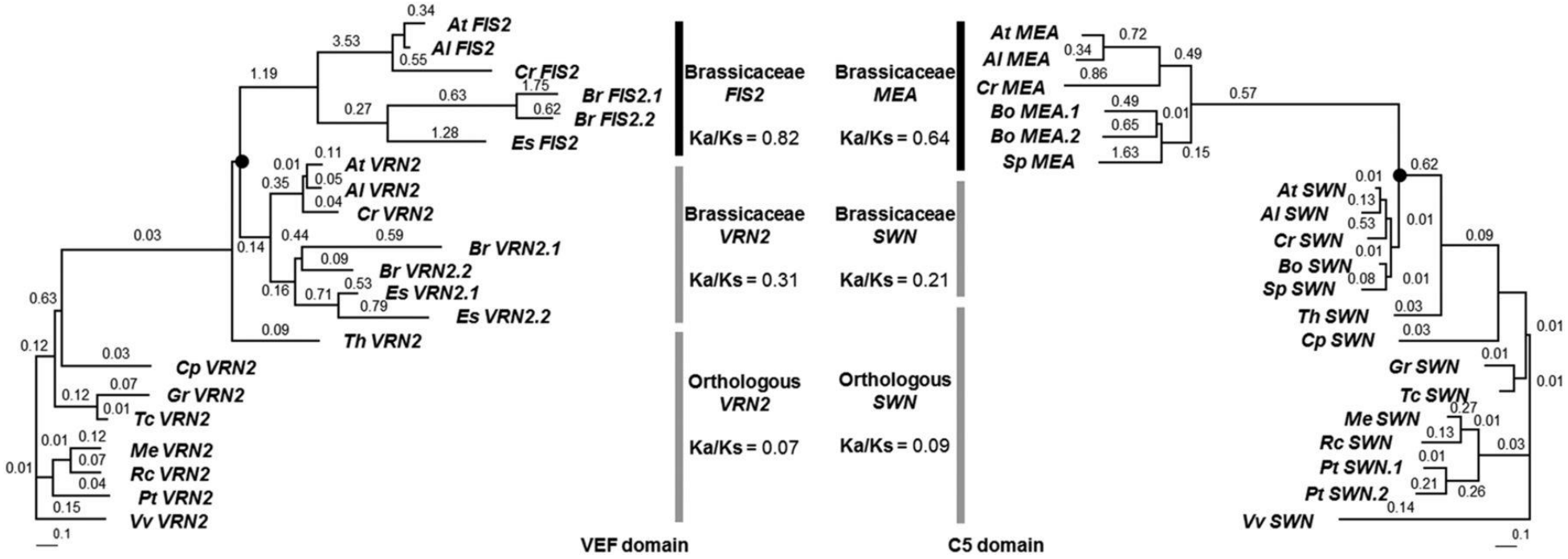


Fig. 4.8. Ka/Ks ratios of full-length FIS2/VRN2 and MEA/SWN genes and functional domains. Estimated average Ka/Ks ratio of each clade is shown between the two trees. The values above branches are Ka/Ks ratios. The black dots indicate the alpha WGD at the base of the Brassicaceae. The scale bars indicate 0.1 substitution per codon. Species abbreviations include: At - *Arabidopsis thaliana*, Al - *Arabidopsis lyrata*, Ah - *Arabidopsis halleri*, Cr - *Capsella rubella*, Sp - *Schrenkiella parvula*, Es - *Eutrema salsugineum*, Br - *Brassica rapa*, Bo - *Brassica oleracea*, Th - *Tarenaya hassleriana*, Cp - *Carica papaya*, Gr – *Gossypium raimondii*, Tc - *Theobroma cacao*, Cs - *Citrus sinensis*, Pt - *Populus trichocarpa*, Rc - *Ricinus communis*, Me - *Manihot esculenta* and Vv - *Vitis vinifera*.

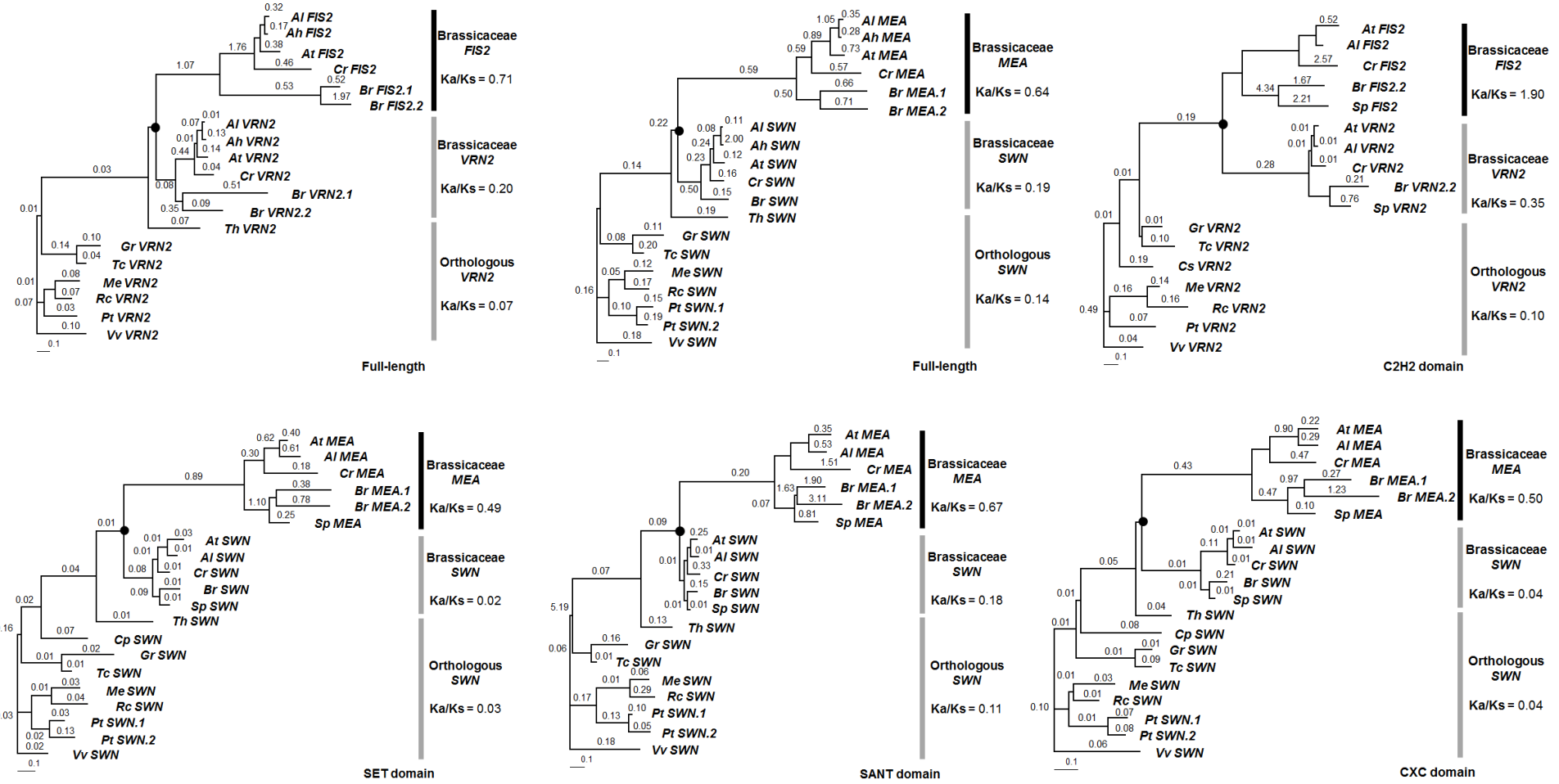
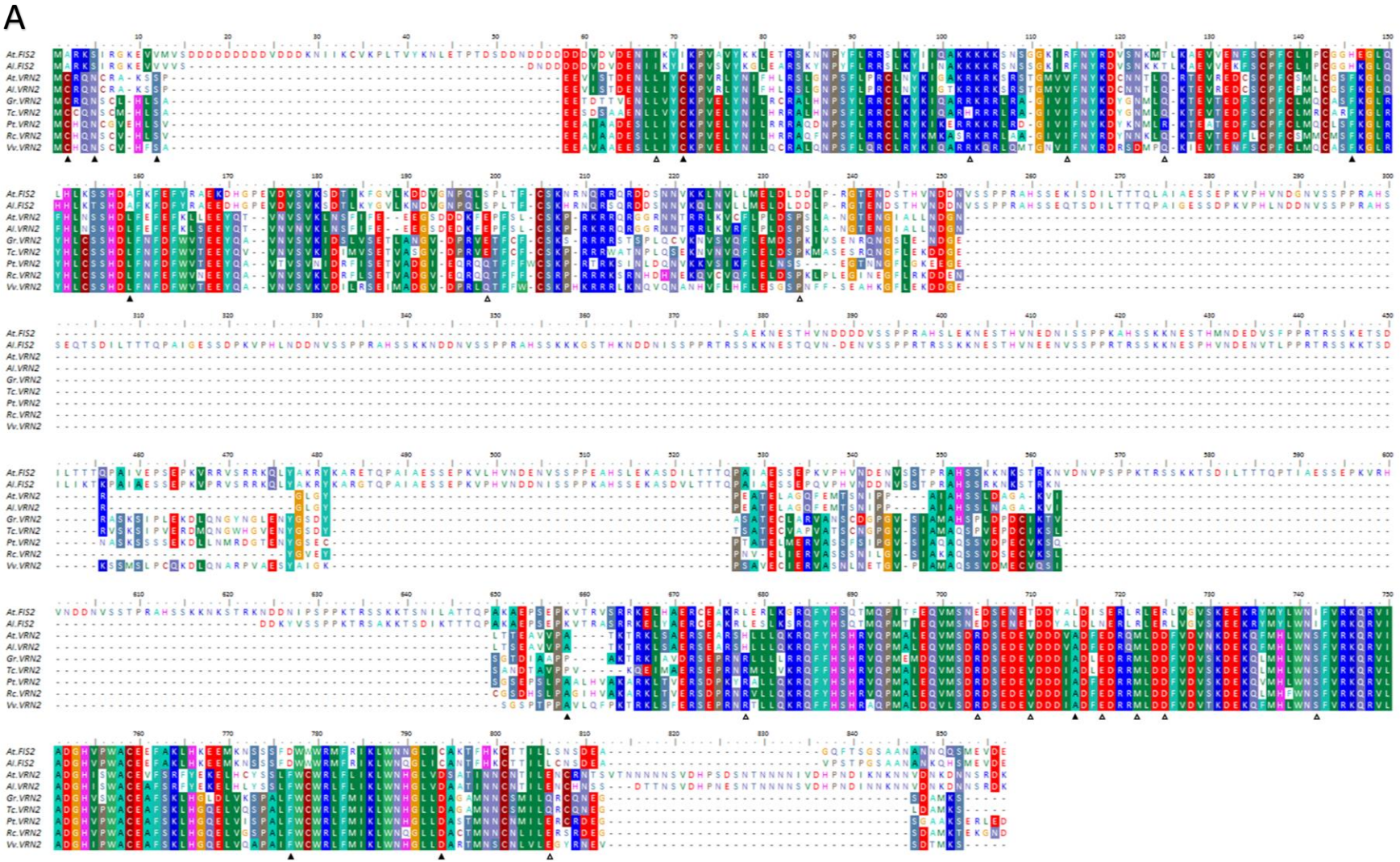


Fig. 4.9. Positive selection on specific sites of MEA and FIS2 genes. Sequence alignments of the protein sequences of FIS2, VRN2 and orthologs **(A)**; MEA, SWN and orthologs **(B)**. Amino acid residues positively selected in the lineage leading to the MEA or FIS2 are pointed by open triangles (with a posterior probability of positive selection larger than 0.95), and solid triangles (with a posterior probability larger than 0.99). Species abbreviations include: At - *Arabidopsis thaliana*, Al - *Arabidopsis lyrata*, Gr – *Gossypium raimondii*, Tc - *Theobroma cacao*, Pt - *Populus trichocarpa*, Rc - *Ricinus communis* and Vv - *Vitis vinifera*.



B

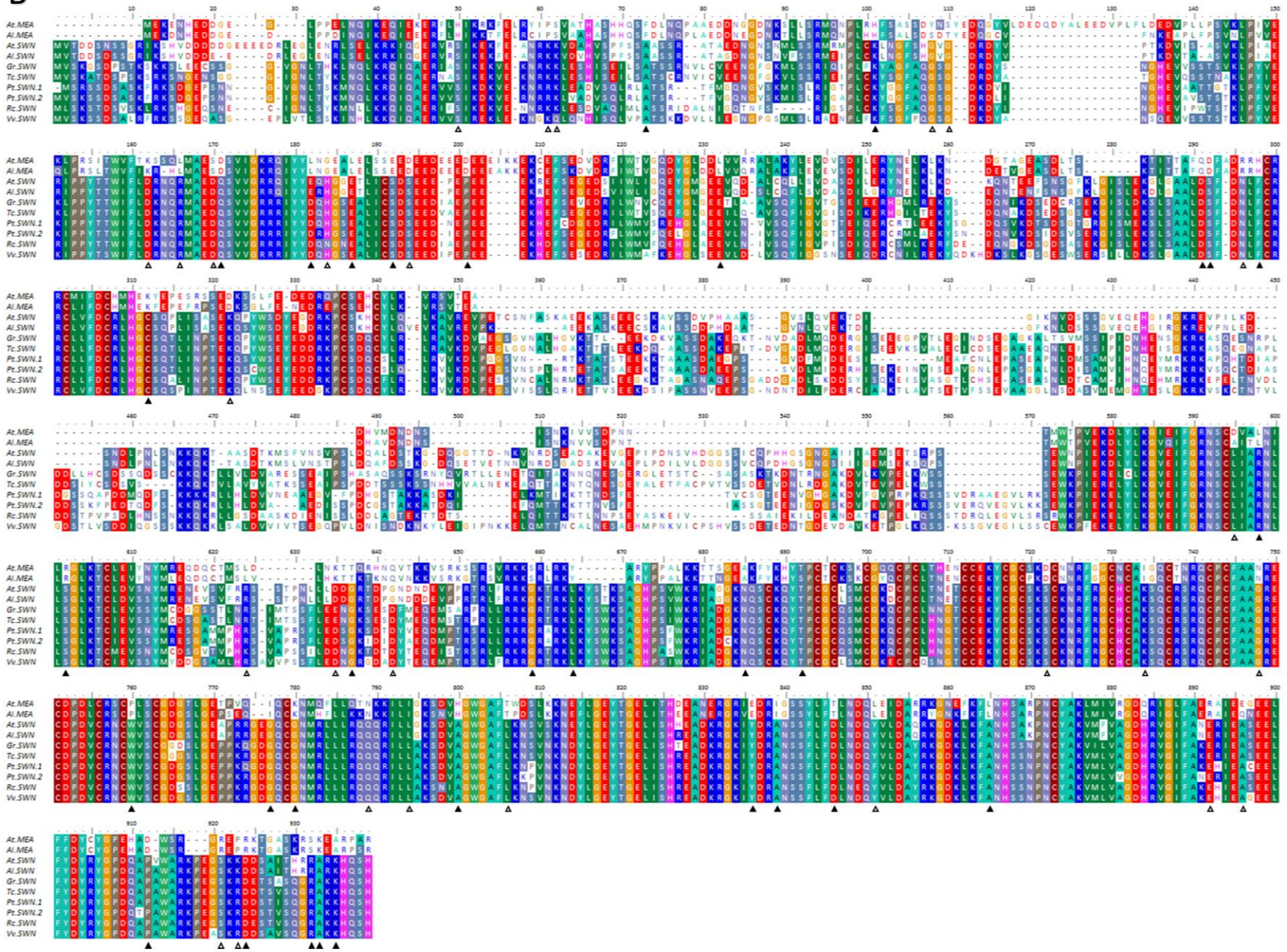

At.MEA
Al.MEA
At.SWN
Al.SWN
Gr.SWN
Tc.SWN
Pt.SWN.1
Pt.SWN.2
Rc.SWN
Vv.SWN

Fig. 4.10. *VEL2* and *VEL1* expression and sequence evolution. **A.** Organ/tissue specificity of *VEL* genes. **B.** Correlation of expression profile between *VEL* genes and PRC2 core components. Left: ADA set (63 organ types and developmental stages); right: ASA set (42 seed tissue types and developmental stages). Black arrows indicate positive correlation, and grey arrows indicate negative correlation. The thickness of arrows indicates the level of the correlation coefficient. The correlation coefficient and p-value of expression profile of each gene pair are labeled along the arrows. Bold values indicate positive correlations. **C.** DNA methylation at the genomic region of *VEL* genes (as in Fig.4.5). **D.** Ka/Ks values of the *VEL* genes. Average Ka/Ks ratio of each clade is shown. The black dot at the node indicates gene duplication events.

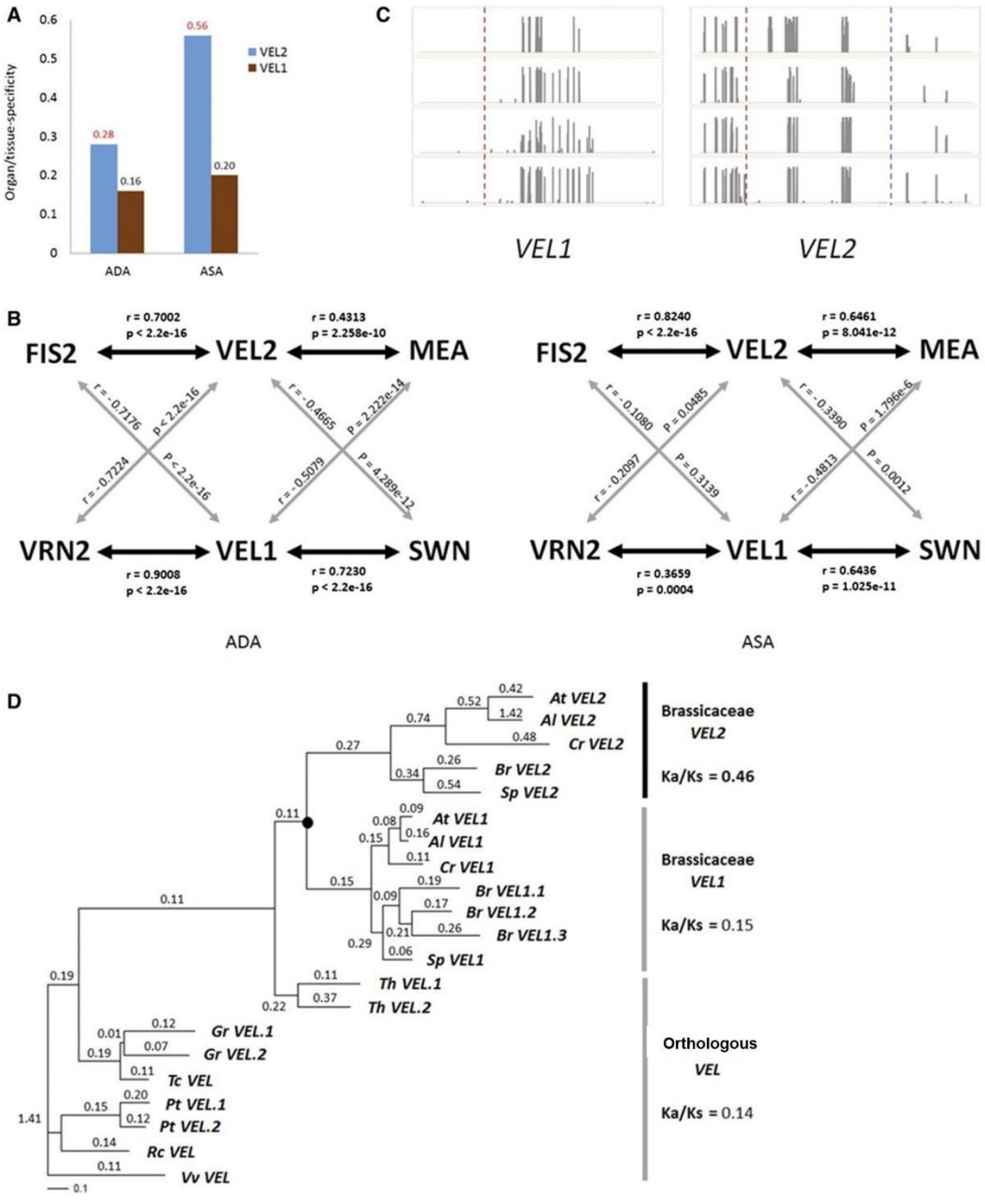
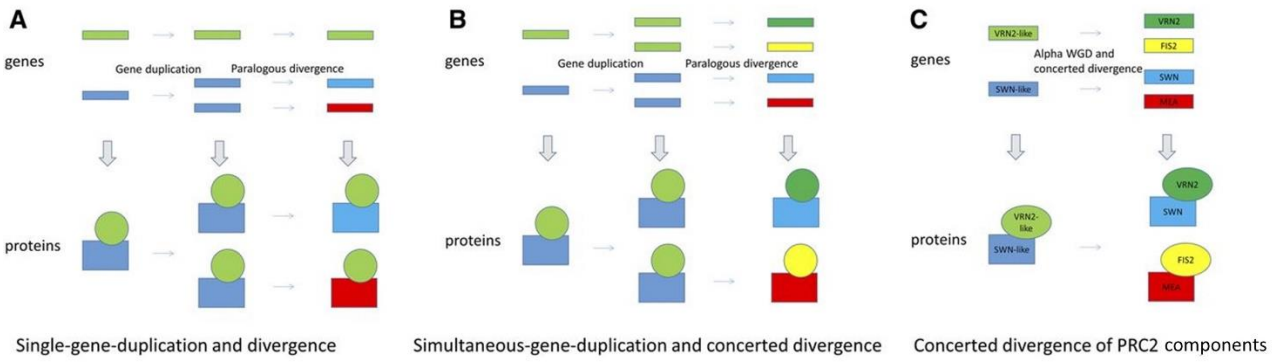


Fig. 4.11. Schematic diagrams illustrating models of protein complex divergence. Colors indicate conservation vs. divergence (could be neofunctionalization, subfunctionalization, loss of partial function, and other types of divergence). **A.** Single-gene-duplication and divergence: a single gene (dark blue) in a complex is duplicated. After duplication there is subsequent divergence (light blue vs. red) of the ancestral gene (dark blue) to give rise to divergent protein complexes. **B.** Simultaneous-gene-duplication and concerted divergence: two (or more) genes (dark green + dark blue) were duplicated simultaneously. After duplication there is parallel divergence (light green + light blue vs. yellow + red) to give rise to divergent protein complexes. **C.** The PRC2 complexes in this study are an example of simultaneous-gene-duplication and concerted divergence.



Single-gene-duplication and divergence | Simultaneous-gene-duplication and concerted divergence | Concerted divergence of PRC2 components

5 Conclusion

My thesis studied fates of duplicated genes with a focus on new models and new variations on existing models. I characterized examples of each of the duplicate gene fates.

In Chapter 2, I reported a case study of a new model of paralog divergence, sub-localization of duplicates by partitioning of alternative splice forms from the original single copy gene between the two duplicates. I identified more than ten cases of duplicated plastid *APX* gene pairs that were sub-localized during angiosperm evolution. After gene duplication, paralogs may diverge by changes in exon-intron structures such as new intron formation, shrinkage or expansion of introns, gain and loss of splicing elements and sites, exon shuffling, and other mutations along exons and/or introns (Xu et al., 2012). Consequently, the splicing pattern could diverge substantially, and gain or loss of certain alternatively spliced variants or the relative proportion of different variants is subjected to divergence by post-transcriptional regulation (Zhang et al., 2010; Tack et al., 2014). After gene duplication, paralogs may also diverge by changes in subcellular localization (Byun-McKay and Geeta, 2007). This study is the first to document examples of sub-localization of paralogs by partitioning of alternative splice forms. This study presents a type of functional divergence that facilitates the retention of both duplicated genes, which provides insights into understanding the dynamics of gene loss and retention after duplication, especially to functionally important organellar associated genes. The cpAPX examples in my study use alternative splicing of a final exon in the gene. Other mechanisms, such as an alternative translation start site, or alternative splicing of the first exon, could also be

possible to cause inclusion and exclusion of a transit peptide in the final protein products thus leading to dual targeting (Yogev and Pines, 2011).

Although I documented over ten independent cases of sub-localization of paralogs, with more newly sequenced genomes becoming available in the future, there might be more cases of duplication and sub-localization of APX to be found. New genome data may also allow for more precise inference of the phylogenetic timing of the duplications and sub-localizations as well as the time lag between duplication and sub-localization. It also might be possible to estimate the frequency of sub-localization in comparison to gene loss. Future studies could also identify other genes that are candidates for sub-localization after duplication. Good candidate genes are those that have a single gene with alternative splicing that produces gene products that are localized to different subcellular compartments. With a broader survey of alternative splicing and dual subcellular localization in additional genes in a variety of taxa, more cases may be found and further information about the frequency of this evolutionary process could be obtained.

A broader definition of gene duplication is not limited to those found within a genome. There have been studies showing duplicated genes located in different subcellular compartments through intercellular gene transfer (e.g., Liu et al., 2009). Chapter 3 presents a new direction of intracellular gene transfer involving a duplicated nuclear gene. Previously only fragments of nuclear genes, pseudogenes, and transposable elements had been found in plant mitochondrial genomes, but none were found to be transcribed. I found the first case of a gene transferred

from the nucleus to the mitochondrial genome that is transcribed with an intact open reading frame. Expression likely was gained by use of the promoter sequences of a nearby tRNA gene. I did not determine if ORF164 is translated, but even if it is, it is unlikely to be functional in mitochondria. It is also possible that *ORF164* could function as a non-coding RNA. After my study was published in 2014, another case involving a gene transferred from the nuclear genome to the mitochondrial genome, which finally ended up in the plastid genome, was reported in carrot (Spooner et al., 2017). Coincidentally it is another AUXIN RESPONSIVE FACTOR gene. With more nuclear and mitochondrial genomes from the same species sequenced in the future, additional cases of gene transfer from the nucleus to the mitochondrion might be identified.

In Chapter 4, I proposed a model for protein complex divergence involving duplicated gene products. Previously, functional and regulatory divergence of duplicated genes have been extensively studied, but were described with respect to of a single gene pair or a gene family. However, it should not be ignored that many gene products interact with each other, and the potential co-evolution between the products of simultaneously duplicated genes has not been examined. Several initial attempts to study the connection between duplicated genes pairs involved expression analyses. The concept of concerted divergence was initially proposed based on co-expression data (Blanc and Wolfe, 2004). The model in my study provides a concrete case further extending the expression co-divergence into functional concerted evolution, with emphasis on the evolution of protein complexes with components coded for by different types of genes. There are many protein complexes within the cell with members derived from

different genes. Among them, PRC2 demonstrates this new model. The key patterns in this model are that there are multiple gene products functioning together, and two or more genes are duplicated at the same time such as through a WGD, and most importantly, the duplicates diverge with pairwise coordination. Within the PRC2 components, *FIS2* has been shown to have diverged in function with its paralog, *VRN2*, and *MEA* has also diverged in comparison to *SWN*. I provided multiple lines of evidence showing that the divergence between *FIS2* and *VRN2* is parallel to the divergence between *MEA* and *SWN*, which drives the functional divergence of the whole complexes. This provides the first example of the functional divergence of protein complexes whose components duplicated and diverged in parallel. I also used two models to summarize the potential evolutionary trajectories of protein complexes after their components are duplicated, distinguished by single-gene-duplication or simultaneous-gene duplication. The key difference is the timing of gene duplication which can be resolved by finer phylogenetic inference.

With better knowledge of non-coding RNAs that make up the RNP complexes, the proposed model for complex divergence may not limit to protein-protein interaction and could be generalized into other kinds of molecular interactions. It can be predicted that many other protein complexes will fit in one of the two models, especially those with many components, such as RNA polymerase, and there are expected rounds of gene duplication in ribosome associated or spliceosome associated protein or RNA genes. With better understanding of more genomes of diverse species, and more detailed elaboration of protein complexes, a future

direction could be to identify more cases of each type of protein complex duplication, and the relative abundance or frequency could be estimated.

My thesis aimed to provide new insights into the mechanisms of duplicated gene retention and divergence. There are other mechanisms of divergence of paralogs that could occur and should be further considered. Epigenetic factors, including DNA and histone modifications, which are parts of my analyses in the case reported in Chapter 4, have also been addressed in Berke et al. (2012) and Wang et al. (2014). One type of post-translational regulation, phosphorylation, was studied in Amoutzias et al. (2010). Shifts between homo- or hetero- dimerization after gene duplication have been shown in Bartlett et al. (2016). MicroRNA regulation is addressed in Wang and Adams (2015). Epigenetic regulation is not only pre-transcriptional, but it could also regulate alternative splicing post-transcriptionally. An alternatively spliced region could correspond to a microRNA binding site or a phosphorylation site. The mechanisms mentioned above could all potentially cause functional divergence of duplicated genes. Future studies of a variety of divergence mechanisms will provide additional perspectives on the fates of duplicated genes.

References

Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. 2015. Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell. 163 (3):594-606.

Abdelsamad A, Pecinka A. 2014. Pollen-specific activation of *Arabidopsis* retrogenes is associated with global transcriptional reprogramming. Plant Cell. 26(8):3299-313.

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(8):1268-75.

Adams KL, Palmer JD. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol. 29(3):380-95.

Aichinger E, Villar CB, Di Mambro R, Sabatini S, Köhler C. 2011. The CHD3 chromatin remodeler PICKLE and polycomb group proteins antagonistically regulate meristem activity in the *Arabidopsis* root. Plant Cell. (3):1047-60.

Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell. 23(7):2499-513.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol. 27 (6):1436-48.

Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, Van de Peer Y. 2010. Posttranslational regulation impacts the fate of duplicated genes. Proc Natl Acad Sci U S A. 107(7):2967-71.

Asada K. 1999. The water-water cycle in chloroplasts: scavenging of active oxygens and dissipation of excess photons. Annu Rev Plant Physiol Plant Mol Biol. 50:601-639.

Arsovski AA, Pradinuk J, Guo XQ, Wang S, Adams KL. 2015. Evolution of cis-regulatory elements and regulatory networks in duplicated genes of *Arabidopsis*. Plant Physiol. 169(4):2982-91.

Barker MS, Li Z, Kidder TI, Reardon CR, Lai Z, Oliveira LO, Scascitelli M, Rieseberg LH. 2016. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am J Bot. 103(7):1203-11.

Barbazuk WB, Fu Y, McGinnis KM. 2008. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res. 18(9):1381-92.

Bartlett M, Thompson B, Brabazon H, Del Gizzi R, Zhang T, Whipple C. 2016. Evolutionary dynamics of floral homeotic transcription factor protein-protein interactions. Mol Biol Evol. 33(6):1486-501.

Baumbusch LO, Thorstensen T, Krauss V, Fischer A, Naumann K, Assalkhou R, Schulz I, Reuter G, Aalen RB. 2001. The *Arabidopsis thaliana* genome contains at least 29 active genes encoding SET domain

proteins that can be assigned to four evolutionarily conserved classes. Nucleic Acids Res. 29 (21):4319-33.

Beilstein MA, Brinegar AE, Shippen DE. 2012. Evolution of the *Arabidopsis* telomerase RNA. Front Genet. 3:188.

Beilstein MA, Renfrew KB, Song X, Shakirov EV, Zanis MJ, Shippen DE. 2015. Evolution of the telomere-associated protein POT1a in *Arabidopsis thaliana* is characterized by positive selection to reinforce protein-protein interaction. Mol Biol Evol. 32(5):1329-41.

Bensasson D, Zhang D, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 16(6):314-321.

Berger F, Chaudhury A. 2009. Parental memories shape seeds. Trends Plant Sci.14(10):550-6.

Berke L, Sanchez-Perez GF, Snel B. 2012. Contribution of the epigenetic mark H3K27me3 to functional divergence after whole genome duplication in *Arabidopsis*. Genome Biol. 13(10):R94.

Bertioli DJ, Moretzsohn MC, Madsen LH, Sandal N, Leal-Bertioli SC, Guimarães PM, Hougaard BK, Fredslund J, Schauser L, Nielsen AM, Sato S, Tabata S, Cannon SB, Stougaard J. 2009. An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics. 10:45.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet. 21(4):219-26.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. 13(2):137-44.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell. 16(7):1679-91.

Bonen L, Calixte S. 2006. Comparative analysis of bacterial- origin genes for plant mitochondrial ribosomal proteins. Mol Biol Evol. 23(3):701-12.

Bouyer D, Roudier F, Heese M, Andersen ED, Gey D, Nowack MK, Goodrich J, Renou JP, Grini PE, Colot V, Schnittger A. 2011. Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. PLoS Genet. 7(3): e1002014.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 422(6930):433-8.

Buggs RJ, Zhang L, Miles N, Tate JA, Gao L, Wei W, Schnable PS, Barbazuk WB, Soltis PS, Soltis DE. 2011. Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol. 21(7):551-6.

Byun SA, Singh S. 2013. Protein subcellular relocalization increases the retention of eukaryotic duplicate genes. Genome Biol Evol. 5(12):2402-9.

Byun-McKay SA, Geeta R. 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. Trends Ecol Evol. 22(7):338-44.

Capra EJ, Perchuk BS, Skerker JM, Laub MT. 2012. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell. 150(1):222-32.

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

Ceci LR, Veronico P, Gallerani R. 1996. Identification and mapping of tRNA genes on the *Helianthus annuus* mitochondrial genome. DNA Seq. 6(3):159-66.

Cenci A, Combes MC, Lashermes P. 2010. Comparative sequence analyses indicate that *Coffea* (Asterids) and *Vitis* (Rosids) derive from the same paleo-hexaploid ancestral genome. Mol Genet Genomics. 283(5):493-501.

Chang S, Yang T, Du T, Huang Y, Chen J, Yan J, He J, Guan R. 2011. Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in *Brassica*. BMC Genomics. 12:497.

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

Chen LJ, Diao ZY, Specht C, Sung ZR. 2009. Molecular evolution of VEF-domain-containing PcG genes in plants. Mol Plant. 2(4):738-54.

Cheng F, Liu S, Wu J, Fang L, Sun S, Liu B, Li P, Hua W, Wang X. 2011. BRAD, the genetics and genomics database for *Brassica* plants. BMC Plant Biol. 11: 136.

Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, Hofberger J, de Bruijn S, Bhide AS, Kuelahoglu C, Bian C, Chen J, Fan G, Kaufmann K, Hall JC, Becker A, Bräutigam A, Weber AP, Shi C, Zheng Z, Li W, Lv M, Tao Y, Wang J, Zou H, Quan Z, Hibberd JM, Zhang G, Zhu XG, Xu X, Schranz ME. 2013. The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. Plant Cell. 25(8): 2813-30.

Clarkson JJ, Lim KY, Kovarik A, Chase MW, Knapp S, Leitch AR. 2005. Long-term genome diploidization in allopolyploid *Nicotiana* section Repandae (Solanaceae). New Phytol. 168 (1): 241-252.

Coate JE, Song MJ, Bombarely A, Doyle JJ. 2016. Expression-level support for gene dosage sensitivity in three *Glycine* subgenus *Glycine* polyploids and their diploid progenitors. New Phytol. 212(4):1083-1093.

Couvreur TL, Franzke A, Al-Shehbaz IA, Bakker FT, Koch MA, Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol Biol Evol. 27(1):55-71.

Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ, Weigel D, Mackenzie SA. 2011. Double-strand break repair processes drive evolution of the mitochondrial genome in *Arabidopsis*. BMC Biol. 9:64.

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature. 454(7205):762-5.

De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci U S A. 110(8):2898-903.

De Smet R, Sabaghian E, Li Z, Saeys Y, Van de Peer Y. 2017. Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*. Plant Cell. 29(11):2786-2800.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, Mckain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci AM, Weissenbach J, Ruiz M, Glaszmann JC, Quétier F, Yahiaoui N, Wincker P. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature. 488(7410):213-7.

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Mol Biol Evol. 23(2):469-78.

Duchêne AM, Maréchal-Drouard L. 2001. The chloroplast-derived trnW and trnM-e genes are not expressed in *Arabidopsis* mitochondria. Biochem Biophys Res Commun. 285(5):1213-6.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-7.

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.

Finseth FR, Dong Y, Saunders A, Fishman L. 2015. Duplication and adaptive evolution of a key centromeric protein in *Mimulus*, a genus with female meiotic drive. Mol Biol Evol. 32(10):2694-706.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol. 183(3):557-64.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151(4):1531-45.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 60:433-53.

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome Res. 18(12):1924-37.

Fujii S, Toda T, Kikuchi S, Suzuki R, Yokoyama K, Tsuchida H, Yano K, Toriyama K. 2011. Transcriptome map of plant mitochondria reveals islands of unexpected transcribed regions. BMC Genomics. 12:279.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. Mol Biol Evol. 24(10):2298-309.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol. 31(2):448-54.

Gehring M, Missirian V, Henikoff S. 2011. Genomic analysis of parent-of-origin allelic expression in *Arabidopsis thaliana* seeds. PLoS One. 6(8): e23687.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40 (Database issue): D1178-86.

Goremykin VV, Lockhart PJ, Viola R, Velasco R. 2012.The mitochondrial genome of *Malus domestica* and the import-driven hypothesis of mitochondrial genome expansion in seed plants. Plant J. 71(4):615-26.

Gossmann TI, Schmid KJ. 2011. Selection-driven divergence after gene duplication in *Arabidopsis thaliana*. J Mol Evol. 73(3-4):153-65.

Grimes BT, Sisay AK, Carroll HD, Cahoon AB. 2014. Deep sequencing of the tobacco mitochondrial transcriptome reveals expressed ORFs and numerous editing sites outside coding regions. BMC Genomics. 15:31.

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(2):3009-22.

Hagen G, Guilfoyle T. 2002. Auxin-responsive gene expression: genes, promoters and regulatory factors. Plant Mol Biol. 49(3-4):373-85.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 169(2):1157-64.

Hehenberger E, Kradolfer D, Köhler C. 2012. Endosperm cellularization defines an important developmental transition for embryo development. Development. 139(11):2031-2039.

Hennig L, Derkacheva M. 2009. Diversity of Polycomb group complexes in plants: same rules, different players? Trends Genet. 25(9):414-23.

Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ. 2006. Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. Curr Biol. 16(16):1652-9.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature. 449(7163):677-81.

Hsieh TF, Shin JY, Uzawa R, Silva P, Cohen S, Bauer MJ, Hashimoto M, Kirkbride RC, Harada JJ, Zilberman D, Fischer RL. 2011. Regulation of imprinted gene expression in *Arabidopsis* endosperm. Proc Natl Acad Sci U S A. 108(5):1755-62.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43(5):476-81.

Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138 (3):1723-1733.

Ishikawa T, Shigeoka S. 2008. Recent advances in ascorbate biosynthesis and the physiological significance of ascorbate peroxidase in photosynthesizing organisms. Biosci Biotechnol Biochem 72(5):1143–54.

Ishikawa T, Yoshimura K, Tamoi M, Takeda T, Shigeoka S. 1997. Alternative mRNA splicing of 3'-terminal exons generates ascorbate peroxidase isoenzymes in spinach (*Spinacia oleracea*) chloroplasts. Biochem. J. 328(Pt 3), 795–800.

Jakobsson M, Hagenblad J, Tavare S, Sall T, Hallden C, Lind-Hallden C, Nordborg M. 2006. A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. Mol Biol Evol. 23:1217-1231.

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJ, Ohyanagi H, Mineta K, Michell CT, Saber N, Kharbatia NM, Rupper RR, Sharp AR, Dally N, Boughton BA, Woo YH, Gao G, Schijlen EG, Guo X, Momin AA, Negrão S, Al-Babili S, Gehring C, Roessner U, Jung C, Murphy K, Arold ST, Gojobori T, Linden CG, van Loo EN, Jellen EN, Maughan PJ, Tester M. 2017. The genome of *Chenopodium quinoa*. Nature. 542(7641):307-312. 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. 473(7345):97-100.

Joyce PB, Gray MW. 1989. Chloroplast-like transfer RNA genes expressed in wheat mitochondria. Nucleic Acids Res. 17(14):5461-76.

Kangasjarvi, S., Lepisto, A., Hannikainen, K., Piippo, M., Luomala, E.M., Aro, E.M., Rintamaki E. 2008. Diverse roles for chloroplast stromal and thylakoidbound ascorbate peroxidases in plant stress responses. Biochem. J. 412 (2): 275–285.

Kim DH, Sung S. 2010. The Plant Homeo Domain finger protein, VIN3-LIKE 2, is necessary for photoperiod-mediated epigenetic regulation of the floral repressor, MAF5. Proc Natl Acad Sci U S A. 107(39):17029-34.

Kim DH, Sung S. 2013. Coordination of the vernalization response through a *VIN3* and *FLC* gene family regulatory network in *Arabidopsis*. Plant Cell. 25(2):454-469.

Kleine T, Maier UG, Leister D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol. 60:115-38.

Köhler C, Hennig L, Spillane C, Pien S, Gruissem W, Grossniklaus U. 2003. The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene *PHERES1*. Genes Dev. 17(12): 1540-53.

Köhler C, Wolff P, Spillane C. 2012. Epigenetic mechanisms underlying genomic imprinting in plants. Annu Rev Plant Biol. 63:331-52.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biol. 3(2):R8.

Kradolfer D, Wolff P, Jiang H, Siretskiy A, Köhler C. 2013. An imprinted gene underlies postzygotic reproductive isolation in *Arabidopsis thaliana*. Dev Cell. 26(5):525-35.

Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T. 2000. The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* L.) reveals a novel gene for tRNA(Cys)(GCA). Nucleic Acids Res. 28(13):2571-6.

Kubo T, Yanai Y, Kinoshita T, Mikami T. 1995. The chloroplast trnP-trnW-petG gene cluster in the mitochondrial genomes of *Beta vulgaris*, *B. trigyna* and *B. webbiana*: evolutionary aspects. Curr Genet. 27(3):285-9.

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB. 2010. Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A. 107(18):8063-70.

Leon P, Walbot V, Bedinger P. 1989. Molecular analysis of the linear 2.3 kb plasmid of maize mitochondria: apparent capture of tRNA genes. Nucleic Acids Res. 17(11):4089-99.

Li S, Zhou B, Peng X, Kuang Q, Huang X, Yao J, Du B, Sun MX. 2014. OsFIE2 plays an essential role in the regulation of rice vegetative and reproductive development. New Phytol. 201(1):66-79.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. Sci Adv. 1(10):e1501084.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell. 28(2):326-44.

Liang Z, Schnable JC. Functional divergence between subgenomes and gene pairs after whole genome duplications. Mol Plant. 2018 Mar 5;11(3):388-397.

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Adams MD, Carrera AJ, Creasy TH, Goodman HM, Somerville CR, Copenhaver GP, Preuss D, Nierman WC, White O, Eisen JA, Salzberg SL, Fraser CM, Venter JC. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature. 402(6763):761-8.

Liscum E, Reed JW. 2002. Genetics of Aux/IAA and ARF action in plant growth and development. Plant Mol Biol. 49(3-4):387-400.

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

Liu SL, Baute GJ, Adams KL. 2011. Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. Genome Biol Evol. 3:1419-36.

Liu SL, Pan AQ, Adams KL. 2014. Protein subcellular relocalization of duplicated genes in *Arabidopsis*. Genome Biol Evol. 6(9):2501-15.

Liu SL, Zhuang Y, Zhang P, Adams KL. 2009. Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. Mol Biol Evol. 26(4):875-91.

Lough AN, Roark LM, Kato A, Ream TS, Lamb JC, Birchler JA, Newton KJ. 2008. Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics. 178(1):47-55.

Luo M, Bilodeau P, Koltunow A, Dennis ES, Peacock WJ, Chaudhury AM. 1999. Genes controlling fertilization-independent seed development in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 96(1):296-301.

Luo M, Platten D, Chaudhury A, Peacock WJ, Dennis ES. 2009. Expression, imprinting, and evolution of rice homologs of the polycomb group genes. Mol Plant. 2(4):711-23.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290(5494):1151-5.

Makarevitch I, Eichten SR, Briskine R, Waters AJ, Danilevskaya ON, Meeley RB, Myers CL, Vaughn MW, Springer NM. 2013. Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3K27. Plant Cell. 25(3):780-793.

Mano S, Yamaguchi K, Hayashi M, Nishimura M. 1997. Stromal and thylakoid-bound ascorbate peroxidases are produced by alternative splicing in pumpkin. FEBS Lett. 413(1):21-6.

Marechal-Drouard L, Guillemaut P, Cosset A, Arbogast M, Weber F, Weil JH, Dietrich A. 1990. Transfer RNAs of potato (*Solanum tuberosum*) mitochondria have different genetic origins. Nucleic Acids Res. 11(13):3689-96.

Marienfeld J, Unseld M, and Brennicke A. 1999. The mitochondrial genome of *Arabidopsis* is composed of both native and immigrant information. Trends Plant Sci. 4(12):495-502.

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

Maruta T, Tanouchi A, Tamoi M, Yabuta Y, Yoshimura K, Ishikawa T, Shigeoka S. 2010. *Arabidopsis* chloroplastic ascorbate peroxidase isoenzymes play a dual role in photoprotection and gene regulation under photooxidative stress. Plant Cell Physiol. 51(2):190-200.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol. 8(2):122-8.

Mozgova I, Köhler C, Hennig L. 2015. Keeping the gate closed: functions of the polycomb repressive complex PRC2 in development. Plant J. 83(1):121-32.

Noe L, Kucherov G. 2005. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 33:W540-3.

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K. 2002. The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol Genet Genomics. 268(4):434-45.

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

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. Plant Physiol. 171(4):2294-316.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A. 101(26):9903-8.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MF, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KF, Peterson DG, Rokhsar DS, Wang X, Schmutz J. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature. 492(7429):423-7.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet. 3(11):827-37.

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K. 2015. PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. 43(Database issue): D974-81.

Qian Y, Xi Y, Cheng B, Zhu S, Kan X. 2014. Identification and characterization of the SET domain gene family in maize. Mol Biol Rep. 41(3):1341-54.

Qiu Y, Liu SL, Adams KL. 2014. Frequent changes in expression profile and accelerated sequence evolution of duplicated imprinted genes in *Arabidopsis*. Genome Biol Evol. 6(7):1830-42.

Reddy AS. 2007. Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol. 58:267-94.

Reddy AS, Marquez Y, Kalyna M, Barta A. 2013. Complexity of the alternative splicing landscape in plants. Plant Cell. 25(10):3657-83.

Rösti S, Denyer K. 2007. Two paralogous genes encoding small subunits of ADP-glucose pyrophosphorylase in maize, *Bt2* and *L2*, replace the single alternatively spliced gene found in other cereal species. J Mol Evol. 65(3):316-27.

Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta MV, Munzinger J, Barry K, Boore JL, Zhang Y, dePamphilis CW, Knox EB, Palmer JD. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science. 342(6165):1468-73.

Roark LM, Hui AY, Donnelly L, Birchler JA, Newton KJ. 2010. Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet Genome Res. 129 (1-3):17-23.

Rodrigues JC, Tucker MR, Johnson SD, Hrmova M, Koltunow AM. 2008. Sexual and apomictic seed formation in Hieracium requires the plant polycomb-group gene *FERTILIZATION INDEPENDENT ENDOSPERM*. Plant Cell. 20(9):2372-86.

Roszak P, Köhler C. 2011. Polycomb group proteins are required to couple seed coat initiation to fertilization. Proc Natl Acad Sci U S A. 108(51):20826-31.

Roudier F, Ahmed I, Bérard C, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L, Giraut L, Després B, Drevensek S, Barneche F, Dèrozier S, Brunaud V, Aubourg S, Schnittger A, Bowler C, Martin-Magniette ML, Robin S, Caboche M, Colot V. 2011. Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. EMBO J. 30 (10):1928-38.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. Nat Genet. 37(5):501-6.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR. 2013. Patterns of population epigenomic diversity. Nature. 495(7440):193-8.

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(7278):178-83.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 108(10):4069-74.

Schranz M, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Planc Cell. 18(5): 1152-65.

Shan H, Zahn L, Guindon S, Wall PK, Kong H, Ma H, DePamphilis CW, Leebens-Mack J. 2009. Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. Mol Biol Evol. 26(10): 2229-44.

She X, Cheng Z, Zöllner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. Nat Genet. 40(7):909-14.

Silva-Filho MC. 2003. One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. Curr Opin Plant Biol. 6(6):589-95.

Spooner DM, Ruess H, Iorizzo M, Senalik D, Simon P. 2017. Entire plastid phylogeny of the carrot genus (*Daucus*, Apiaceae): concordance with nuclear data and mitochondrial and nuclear DNA insertions to the plastid. Am J Bot. 104(2):296-312.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. Curr Opin Genet Dev. 35:119-25.

Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 30:159-65.

Spillane C, Schmid KJ, Laoueille-Duprat S, Pien S, Escobar-Restrepo J-M, Baroux C, Gagliardini V, Page DR, Wolfe KH, Grossniklaus U. 2007. Positive darwinian selection at the imprinted *MEDEA* locus in plants. Nature. 448(7151): 349–52.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22(21):2688-90.

Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. Cell. 152(1-2):352-64.

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J. 2001. Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci U S A. 98(9):5099-103.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 9(6):465-76.

Tanaka Y, Tsuda M, Yasumoto K, Yamagishi H, Terachi T. 2012. A complete mitochondrial genome sequence of Ogura-type male-sterile cytoplasm and its comparative analysis with that of normal cytoplasm in radish (*Raphanus sativus* L.). BMC Genomics. 13:352.

Tack DC, Pitchers WR, Adams KL. 2014. Transcriptome analysis indicates considerable divergence in alternative splicing between duplicated genes in *Arabidopsis thaliana*. Genetics. 198(4):1473-81.

Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A. 107(1):472-7.

Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. Plant Cell. 29(11):2766-2785.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet. 38:615-43.

The Angiosperm Phylogeny Group, M. W. Chase, M. J. M. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, D. J. Mabberley, A. N. Sennikov, P. S. Soltis, P. F. Stevens. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 181(1):1–20.

The tomato genome sequence provides insights into fleshy fruit evolution. 2012. Tomato Genome Consortium. Nature. 485(7400):635-41.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala

J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 313(5793):1596-604.

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K. 2012. Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol. 158(2):590-600.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725-32.

Veitia RA. 2010. A generalized model of gene dosage and dominant negative effects in macromolecular complexes. FASEB J. 24(4):994-1002.

Villar CB, Erilova A, Makarevich G, Trösch R, Köhler C. 2009. Control of *PHERES1* imprinting in Arabidopsis by direct tandem repeats. Mol Plant. 2: 654-60.

Wang J, Marowsky NC, Fan C. 2014. Divergence of gene body DNA methylation and evolution of plant duplicate genes. PLoS One. 9(10):e110357.

Wang J, Tao F, Marowsky NC, Fan C. 2016. Evolutionary fates and dynamic functionalization of young duplicate genes in *Arabidopsis* genomes. Plant Physiol. 172(1):427-40.

Wang S, Adams KL. 2015. Duplicate gene divergence by changes in microRNA binding sites in *Arabidopsis* and *Brassica*. Genome Biol Evol. 7(3):646-55.

Wang X, Guo H, Wang J, Lei T, Liu T, Wang Z, Li Y, Lee TH, Li J, Tang H, Jin D, Paterson AH. 2016. Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. New Phytol. 209(3):1252-63.

Wang Y, Ma H. 2015. Step-wise and lineage-specific diversification of plant RNA polymerase genes and origin of the largest plant-specific subunits. New Phytol. 207(4): 1198-212.

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, Lambert G, Galbraith DW, Grassa CJ, Geraldes A, Cronk QC, Cullis C, Dash PK, Kumar PA, Cloutier S, Sharpe AG, Wong GK, Wang J, Deyholos MK. 2012. The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J. 72(3):461-73.

Wolff P, Weinhofer I, Seguin J, Roszak P, Beisel C, Donoghue MT, Spillane C, Nordborg M, Rehmsmeier M, Köhler C. 2011. High-resolution analysis of parent-of-origin allelic expression in the *Arabidopsis* endosperm. PLoS Genet. 7(6):e1002126.

Xiao J, Sekhwal MK, Li P, Ragupathy R, Cloutier S, Wang X, You FM. 2016. Pseudogenes and their genome-wide prediction in plants. Int J Mol Sci. 17(12):E1991.

Xu G, Guo C, Shan H, Kong H. 2012. Divergence of duplicate genes in exon-intron structure. Proc Natl Acad Sci U S A. 109(4):1187-92.

Xu, L., Carrie, C., Law, S.R., Murcha, M.W. and Whelan, J. 2013. Acquisition, conservation, and loss of dual-targeted proteins in land plants. Plant Physiol. 161: 644–62.

Yang H, Han Z, Cao Y, Fan D, Li H, Mo H, Feng Y, Liu L, Wang Z, Yue Y, Cui S, Chen S, Chai J, Ma L. 2012. A companion cell-dominant and developmentally regulated H3K4 demethylase controls flowering time in *Arabidopsis* via the repression of *FLC* expression. PLoS Genet. 8(4):e1002664.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol Biol Evol. 28(8):2359-69.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586-91.

Yogev O, Pines O. 2011. Dual targeting of mitochondrial proteins: mechanism, regulation and function. Biochim Biophys Acta. 1808(3):1012-20.

Yoo MJ, Liu X, Pires JC, Soltis PS, Soltis DE. 2014. Nonadditive gene expression in polyploids. Annu Rev Genet. 48:485-517.

Yoshimura K, Yabuta Y, Ishikawa T, Shigeoka S. 2002. Identification of a cis element for tissue-specific alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. J Biol Chem. 277(43):40623-32.

Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. 2013. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell. 153(1):193-205.

Zhang H, Bishop B, Ringenberg W, Muir WM, Ogas J. 2012. The CHD3 remodeler PICKLE associates with genes enriched for trimethylation of histone H3 lysine 27. Plant Physiol. 159(1):418-32.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18 (6): 292-298.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22(12):2472-9.

Zhang M, Xie S, Dong X, Zhao X, Zeng B, Chen J, Li H, Yang W, Zhao H, Wang G, Chen Z, Sun S, Hauck A, Jin W, Lai J. 2014. Genome-wide high resolution parental-specific DNA and histone methylation maps uncover patterns of imprinting regulation in maize. Genome Res. 24(1):167-76.

Zhang PG, Huang SZ, Pin AL, Adams KL. 2010. Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of *Arabidopsis*. Mol Biol Evol. 27(7):1686-97.

Zhou R, Moshgabadi N, Adams KL. 2011. Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. Proc Natl Acad Sci U S A. 108(38):16122-7.