

Evolution of Flowering Time in Native North American Grapevines

by

Ke Jiang

A dissertation submitted to the Graduate Faculty of
Auburn University
In partial fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 14, 2010

Keywords: grapevines, flowering time, herbarium specimen,
flowering time genes, molecular evolution, spliceosome intron

Copyright 2010 by Ke Jiang

Approved by

Leslie Goertzen, Chair, Assistant Professor of Biological Sciences
Kenneth Halanych, Associate Professor of Biological Sciences
Scott Santos, Assistant Professor of Biological Sciences
Narendra Singh, Professor of Biological Sciences

Abstract

Native North American grapevines represent a significant portion of the biodiversity within the genus *Vitis* (Vitaceae). Previous studies of taxonomy and natural history suggested patterns of biogeography, phenology and reproductive isolations among wild species. Here specimen informatics and molecular approaches are used to test hypotheses on wild grapevine phylogeny, ecological niches, phenology and their genetic basis. *RNA Polymerase II Subunit 2 (RPB2)* genes were characterized to study angiosperm evolution and genus-wide phylogeny. Biogeography, ecology and phenology were studied using techniques from the emerging field of specimen informatics. Geographic ranges were estimated to investigate geographic modes of speciation. Ecological niche models built with geographic and environmental variables revealed ecological niche partitioning. Phenological information summarized from specimens suggested significant phenological separation among certain species. Grapevine genes controlling flowering time were characterized based on genomic analyses and molecular experimental approaches. Alternative transcription of grapevine *FCA* was discovered, indicating a conserved *FCA* auto-regulation mechanism between grapevine and *Arabidopsis*. Variation in microsatellite region discovered immediately upstream of *TERMINAL FLOWER 1 (TFL1)* was revealed with potential to affect species-specific *TFL1* expression. Multi-locus phylogenetic analyses revealed congruence and conflicts among

phylogenetic signals from flowering time genes and other nuclear markers for grapevine systematics. Molecular evolutionary analyses suggested lack of natural selection on flowering time genes within *Vitis* as a result of prevalent purifying selection. Molecular evolution of flowering time genes in Angiosperms revealed variations in evolutionary rates among different genes. *CONSTANS* (*CO*) showed an elevated nucleotide substitution rate while *FLOWERING LOCUS C* (*FLC*) showed an accelerated amino acid substitution rate compared with other genes. All genes showed spatial variations of selective constraints within coding regions. Most genes showed shifts of selective constraints accompanied with diversification of eudicots, suggesting possible connections between fast flowering gene evolution and the radiation of eudicots. Genomic analyses and annotation of *FLC* revealed an intron of unusually large size. Genome-wide survey of intron size in domesticated grapevine, *Vitis vinifera*, identified extensive intron size expansion compared with *Arabidopsis* and *Populus*. Large introns reduced gene expression levels but are evenly distributed among genes of different functions. Over 80% of the expanded intron space contain repetitive elements with an enrichment of recently inserted LTR-retrotransposons, suggesting an association between intron expansion and grapevine domestication and vegetative propagation.

Acknowledgments

I am grateful to Dr. Leslie Goertzen, my major advisor, and all members of the Goertzen laboratory especially Kataren Johnson and Jennifer Trusty. I am also grateful to Curtis Hansen for his help on herbarium specimens and Dr. Scott Santos for his help with computational analyses. I also want to thank all of my committee members, Dr. Kenneth Halanych, Dr. Scott Santos and Dr. Narendra Singh. Finally, I am grateful to my wife, Yuzhou Wu, and my parents for their unstinting support.

Table of Contents

Abstract.....	ii
Acknowledgments	iv
List of Tables	x
List of Figures.....	xi
Chapter 1. Review of North American wild grapevine natural history and molecular aspects of flowering time control.....	1
Abstract.....	1
Natural history of North American wild grapevines	3
Molecular biology of flowering time	11
Chapter 2. Characterization of grapevine <i>RPB2</i> and phylogenetic analysis	23
Abstract.....	23
Introduction	25
Materials and methods	27
Molecular cloning and sequence analysis	27
Phylogenetic analysis	29
Results.....	31
Characterization of <i>RPB2</i> orthologous and paralogous genes	31
<i>RPB2</i> phylogeny and gene duplication in angiosperms	34
Combined phylogenetic analyses of <i>RPB2</i> and other nuclear marker	35

Discussion	37
<i>RPB2</i> duplication and angiosperm phylogeny	37
Multi-locus phylogenetic analyses based on concatenation and coalescence	40
Chapter 3. Ecological niche partitioning and divergence of flowering phenology of North American native grapevines (<i>Vitis</i> spp) revealed by herbarium specimens	43
Abstract	43
Introduction	45
Materials and methods	48
Data collection	48
Georeferencing process	49
Range estimations	49
Identification of flowering specimens	49
Climate data and ecological niche analysis	50
Flowering time analysis	51
Results	52
Range estimations	52
Ecological niche partitioning	55
Flowering time difference	56
Discussion	61
Range estimation	61
Environmental variables and ecological niches	62
Flowering time differences	64

Implications for research on plant phenology and evolution	66
Chapter 4. Characterization and analyses of grapevine genes controlling flowering time.....	68
Abstract	68
Introduction	70
Materials and methods	73
Molecular cloning and sequence analysis	73
Molecular evolutionary analyses	75
Results	75
Characterization of flowering time genes (<i>CO</i> , <i>FCA</i> , <i>FT</i> , <i>FRI</i> and <i>TFL1A</i>)	75
Alternative transcription of <i>FCA</i> in grapevine	78
Analysis of natural selection on flowering time genes	80
Discussion	82
<i>FT/TFL1A</i> family in grapevines	82
<i>FCA</i> alternative transcription	83
Selective constraints on flowering time genes	84
Chapter 5. Variations in evolutionary rates of Angiosperm flowering time genes	85
Abstract	85
Introduction	87
Materials and methods	90
Taxon sampling and data collection	90
Orthologous gene identification	91
Analyses of evolution rates	92

Molecular clock analysis	92
Genetic distances	92
Evolution rates	93
Results	94
Taxon samplings	94
Ortholog identifications	96
<i>API</i> and <i>FLC</i>	96
<i>CONSTANS</i> family	97
<i>FT</i> and <i>TFL1A</i> family	97
Evolutionary rates	98
Molecular clock	98
Variations in evolution rates	99
Selection analysis	103
Discussion	106
Rate variations relative to positions in pathways	106
Site and lineage specific changes of selective constraints	109
Chapter 6. Genome-wide intron size expansion in domesticated grapevine.....	112
Abstract.....	112
Introduction	114
Materials and methods	116
EST data collection, processing and assembly	116
Identification of large genes and introns	117
Analysis of large intron contents and selected individual introns	118

Identification and molecular evolutionary analyses of MADS-box	
genes	119
Results	119
EST collection, processing and assembly	119
Identification and properties of large genes and introns	120
Contents of large introns and TEs	126
Intron size expansion in MADS-box genes	131
Discussion	136
Transcriptome and intron identification	136
Properties of large gene	137
TEs within introns and gene expression	139
Intron size expansion, TEs, grapevine evolution and domestication	140
Summary	145
References	148
Appendices	161

List of Tables

Table 1 Primers used to characterize <i>Vitis RPB2</i> , <i>FRI</i> and <i>TFL1A</i>	29
Table 2 Ka/Ks of <i>RPB2</i> -Ds in Vitaceae	34
Table 3 Contrasts of environmental variables among groups.....	55
Table 4 Flowering time descriptive statistics.....	58
Table 5 Pairwise comparisons of flowering time and climate within Southeastern group.....	61
Table 6 Primers for characterization of flowering time genes in <i>Vitis</i> spp	74
Table 7 Tests of molecular clock	99
Table 8 LRT results of eudicot specific dN/dS ratio.....	106
Table 9 Characterization of selected introns in wild grapevines and <i>V. vinifera</i> cultivars.....	130

List of Figures

Fig. 1 Preliminary phylogenetic hypotheses about genus level relationship of native North American grapevines (<i>Vitis</i> spp) based on nuclear genes (<i>RPB2</i> , <i>ADH1</i> , <i>TFL1</i> , <i>FRI</i>).....	6
Fig. 2 Fig. 2 Pathways and key genes controlling flowering time in <i>Arabidopsis</i> ; circles indicate genes; arrows indicate activation, blunted lines indicate repression, <i>API</i> , <i>APETALA1</i> ; <i>CO</i> , <i>CONSTANS</i> ; <i>FLC</i> , <i>FLOWERING LOCUS C</i> ; <i>FT</i> , <i>FLOWERING LOCUS T</i> ; <i>LFY</i> , <i>LEAFY</i> ; <i>SOC1</i> , <i>SUPPRESSOR OF OVEREXPRESSION OF CO 1</i> ; <i>TFL1</i> , <i>TERMINAL FLOWER 1</i> (adapted from Alonso-Blanco <i>et al.</i> 2009)	15
Fig. 3 Correlation between frequencies of <i>FRI</i> null alleles and flowering time in natural <i>Arabidopsis</i> populations (adapted from Le Corre 2005).....	18
Fig. 4 Negative feedback loop of <i>FCA</i> self-regulation in <i>Arabidopsis</i> ; Boxes represent exons, lines represent introns, together they represent models of transcripts; protein domains are represented by shaded boxes, together they present protein models (adapted from Macknight <i>et al.</i> 2002)	20
Fig. 5 Alignment of flowering plant <i>RPB2</i> gene coding region (exon 20 to exon 25, numbers in the first line indicate positions in entire coding region); identical sites are shaded; ‘-‘ represent gaps; frameshift mutations in Vitaceae <i>RPB2</i> -D homologues are indicated by arrows	33
Fig. 6 D (red) and I (blue) lineages of <i>RPB2</i> in angiosperm, bootstrap (100 replicates) values are shown on the nodes.....	35

Fig. 7 Comparison of phylogenetic tree topologies within genus <i>Vitis</i> inferred from: a. concatenated multi-locus phylogenetic analyses, numbers on the nodes indicate bootstrap supports; and b. coalescence-based multi-locus phylogenetic analyses, numbers on the nodes indicate Bayesian posterior probabilities.....	37
Fig. 8 Estimated ranges of <i>Vitis</i> spp, species are grouped according to geography: a. Southeastern group, including <i>Vitis aestivalis</i> , <i>Vitis labrusca</i> , <i>Vitis monticola</i> , <i>Vitis mustangensis</i> and <i>Vitis shuttleworthii</i> ; b. Western group, including <i>Vitis arizonica</i> , <i>Vitis girdiana</i> and <i>Vitis californica</i> ; c. <i>Vitis palmata</i> and <i>Vitis vulpina</i> ; d. Southern species <i>Vitis rotundifolia</i> ; e. Northern species <i>Vitis riparia</i> and f. other species, including <i>Vitis acerifolia</i> , <i>Vitis cinerea</i> and <i>Vitis rupestris</i>	54
Fig. 9 Results of PCA (Plot of the first and second principle components, PC1 and PC2, explained over 95% variance in all groups) of bioclimatic variables within a. Southeastern group; b. Western group; c. Northern and Southern group and d. other species, different species are indicated by points of different shapes (legends shown in the up-right corner)	56
Fig. 10 Box plots of flowering time (Julian day) distribution for <i>Vitis</i> spp, species grouped by geography, vertical lines indicate mean flowering time, boxes indicate 75% quartiles, horizontal lines show lower and upper limits.....	59
Fig. 11 Comparison of ecological niches (frequency plots of LD1) and flowering time (frequency plots in Julian day) in geographic groups, a. Southeastern; b. Western; c. Northern/Southern. LD1, the first linear discriminator, a linear combination of principle components (also environmental variables) specifying characteristics of ecological niches	60

Fig. 12 Gene models of characterized flowering time genes. Boxes represent exons and lines represent introns.....	76
Fig. 13 Microsatellite regions (boxed TC repeats) with variable length in front of <i>TFL1A</i> gene start codon in <i>Vitis</i> spp	77
Fig. 14 Diagram of annotation of <i>Vitis vinifera FCA</i> gene. Dashed bar: genomic coordinates (kbp); dark bar: genomic sequences; shade boxes: exons; lines: introns; filled boxes: primers. ESTs and cDNAs annotated with GenBank accession numbers and putative transcript form (β , γ or δ)	79
Fig. 15 Site-specific dN/dS ratio estimates based on codon evolution models, a. <i>FRI</i> exon 1; b. <i>TFL1A</i> complete coding regions, blue and red lines (connected data points) indicate estimated median and mean values of dN/dS ratio for individual amino acid site.....	81
Fig. 16 Simplified diagram of major pathways (vernalization and photoperiod) controlling flowering time; arrows indicate activation, crosses indicate inhibition; <i>API</i> , <i>APETALA1</i> ; <i>CO</i> , <i>COSTANS</i> ; <i>FLC</i> , <i>FLOWERING LOCUS C</i> ; <i>FT</i> , <i>FLOWERING LOCUS T</i>	88
Fig. 17 Distribution of genetic distances estimated from the plastid gene <i>rbcL</i>	95
Fig. 18 Gene trees of a. <i>API</i> (<i>APETALA1</i>) and <i>FLC</i> (<i>FLOWERING LOCUS C</i>); b. <i>FT</i> (<i>FLOWERING LOCUS T</i>) and <i>TFL1A</i> (<i>TERMINAL FLOWER 1</i>) reconstructed by identified homologues, monophyletic groups of orthologous genes are indicated by brackets with gene names (only taxa within the monophyletic groups are included in the molecular evolutionary analyses).....	98
Fig. 19 Distribution of dS and dN across the genes investigated, a. dS rates; b. dN rates.....	101

Fig. 20 dS and dN plots against genetic distances (<i>rbcL</i>), showing locus-specific regression lines between evolution rate and genetic distance, a. dS; b. dN	102
Fig. 21 Site-specific dN/dS ratio (w) in a. <i>API</i> (<i>APETALA1</i>); b. <i>FLC</i> (<i>FLOWERING LOCUS C</i>) and c. <i>CO</i> (<i>CONSTANS</i>), x-axis shows the amino acid sites with functional protein domains indicated by shaded boxes	104
Fig. 22 Expression level represented by EST number in clusters. Bars represent number of clusters with certain number of ESTs: a. 996 large genes identified in <i>Vitis vinifera</i> ; b. 996 genes randomly selected from all assembled clusters	121
Fig. 23 a. Size distribution of large introns identified in <i>Vitis vinifera</i> (>3000 bp). Bars represent number of introns of certain size (bp); b. Distribution of large introns within genes. Bars represent proportions of intron positions within genes.....	123
Fig. 24 Intron and genome size comparisons among <i>Arabidopsis</i> , <i>Populus</i> and <i>Vitis</i> : a. Intron size variation at gene-by-gene base; b. Overall intron size (kbp) and genome size (mbp) comparison	125
Fig. 25 Annotation of <i>Vitis vinifera FLC</i> . Filled bars: exons; lines: introns; open bars: unknown sequences (Ns); hatched bars: LTR retrotransposons; cross-hatched bars: LINES; square-hatched bars: DNA transposons; numbers indicate the orders of TEs annotated within introns (DNA transposon number 1 inserted into LTR retrotransposon number 7 so LTR retrotransposon number 7 is divided into two parts).....	126
Fig. 26 Distribution of LTR retrotransposon age represented by similarity between LTR regions (the higher similarity, the younger the elements) in a. large introns; b. <i>Vitis vinifera</i> chromosome 1	128

Chapter 1

Review of North American wild grapevine natural history and molecular aspects of flowering time control

Abstract

Grapevines (*Vitis*, family Vitaceae) represent an early branching core eudicot group with unique morphological characters such as leaf-opposed tendril. North America harbors a significant portion of wild grapevine diversity with over 15 native species belonging to several series according to current taxonomy. However, evolutionary history has not been well incorporated into grapevine taxonomic treatments. Characterization and phylogenetic analyses of nuclear molecular markers are combined with information from morphology-based taxonomic studies to investigate native North America grapevine evolutionary history and build a phylogenetic framework for future evolutionary studies. With preliminary genus-level phylogeny, ecological and phenological information, hypotheses that native grapevines are reproductively separated by ecological and phenological factors are tested, as the first part of a comprehensive study of grapevine speciation. This study is an example of integrating specimen informatics, Geographic Information Systems (GIS) and molecular phylogenetics to address natural history questions. In addition, I attempt to identify the genetic basis underlying phenological difference in native grapevines, as mechanisms separating species were rarely identified in plants. Recent progress on genetic processes underlying speciation suggested genetic variations associated with population divergence, adaptation and speciation. With

expanding plant genomic information, genome heterogeneity can be assessed to search for signature of divergence and selection related to speciation process. For native grapevines, the major isolation factor, phenological difference, may be influenced by divergence of genes controlling flowering time. In model system *Arabidopsis*, regulatory networks and key genes controlling flowering time were well characterized, including four pathways and several signal integrators. In native grapevines, a candidate gene approach is adopted to characterize genes controlling flowering time. Comparative and association studies are conducted to identify the genetic basis of phenological divergence. Essentially, potential natural variations in grapevine orthologous flowering time genes are identified and analyzed in relation to flowering time differences.

Natural history of North American wild grapevines

Grapevine (genus *Vitis* L., family Vitaceae) is a widespread group of vines common in temperate regions of the Northern Hemisphere. The family Vitaceae consists mainly of woody vines and is characterized by leaf-opposed tendrils. Within the family, the genus *Vitis* bears a unique floral structure, the calyptra, which falls off as a unit at anthesis. The calyptra is formed by secondarily fused petals at the apex as a consequence of development of an abscission layer encircling the corolla base (Comeaux 1984). More than 50 *Vitis* species are recognized worldwide with significant diversity in Asia and North America. The common cultivated species, *Vitis vinifera*, was domesticated around 7000 years ago (Zohary and Spiegel-Roy 1975). Since then, these grapevines are the source of wine and fruit products, contributing in total \$162 billion to the US economy each year (NAAW 2007).

North America harbors a significant portion of *Vitis* biodiversity, including 15~20 wild species genetically distinct from Old World *Vitis* species. However, because of the great morphological variability within the genus, taxonomists made constant and substantial efforts to produce taxonomic treatments for *Vitis* in the past century. T.V. Munson proposed the eight series, or groups of species, as a starting point for modern *Vitis* taxonomy in his 1909 book 'Foundations of American Grape Culture' (Munson 1909). These series are followed by most subsequent treatments or descriptions of *Vitis*. A recent *Vitis* taxonomic system by B. L. Comeaux followed the same series but proposed an improved genus description and key to identification of wild grapevines in East North

America (including *Vitis rotundifolia*, *Vitis labrusca*, *Vitis aestivalis*, *Vitis vulpina*, *Vitis cinerea*, Comeaux 1984). Lately, M. O. Moore focused specifically on the *Vitis* species in the Southeast United States. He eliminated errors in *Vitis* specimen typifications and clarified many synonyms with extensive herbaria data analysis and fieldwork (Moore 1991). Summarizing all previous taxonomic studies, a simplified taxonomic system was adopted in this study, which includes 15 species without considering sub-specific taxonomic levels: *Vitis monticola*, *V. aestivalis*, *V. labrusca*, *Vitis mustangensis*, *Vitis shuttleworthii*, *Vitis californica*, *Vitis arizonica*, *Vitis girdiana*, *V. rotundifolia*, *Vitis riparia*, *Vitis acerifolia*, *V. cinerea*, *V. vulpina*, *Vitis palmata*, *Vitis rupestris*.

As all *Vitis* taxonomic studies are based on morphological characters, there is instability and difference among all of the systems due to high morphological variability of *Vitis*. Therefore, although all previous taxonomic studies tried to incorporate or use systematics and evolutionary histories in their classifications, none of the phylogenetic hypotheses, *i. e.* the hypotheses concerning genealogical relationships among species, have been subjected to vigorous testing. With the advancement of molecular techniques in systematics, molecular markers become ideal tools to test the specific phylogenetic hypotheses within the genus. Plastid genes and two nuclear genes (*ADHI*, *GDHI*) have been characterized as molecular markers to improve the understanding of phylogenetic relationship among native North American grapevines. Use of nuclear markers provided opportunities to identify and investigate natural hybrids and clarify ambiguous species identifications as parentage of alleles from heterozygous loci can be compared and traced. Multiple individuals were usually sampled for each species in an attempt to not

only test species-level phylogenetic hypothesis but also preliminarily delineate species boundaries. Preliminary analyses of molecular data identified several well-supported monophyletic groups: the Muscadine grapes, *V. rotundifolia*; the Southeastern group, consisting of *V. aestivalis*, *V. labrusca*, *V. mustangensis* and *V. shuttleworthii*; the Western group with *V. californica*, *V. arizonica* and *V. girdiana*. A monophyletic group (also suggested by Comeaux (1984)) including *V. vulpina* and *V. palmata* was also weakly supported in the multi-locus phylogenetic analyses. *V. monticola* was traditionally placed in the Western group but the molecular evidence was ambiguous with almost equivalent supports to group this species with either Southeastern species or Western species. The phylogenetic relationship of the remaining species (including *V. riparia*, *V. acerifolia*, *V. cinerea*) to other monophyletic groups are still not well-resolved. Among these species and groups, *V. rotundifolia* (the Muscadine grape) has a different chromosome number than other species. As a result, all *Vitis* species other than *V. rotundifolia* form a monophyletic group (*Euvitis*) based on karyotype. Within *Euvitis*, the Southeastern group occupies a basal position, suggesting an early divergence of this group in native North American grapevines (Fig. 1).

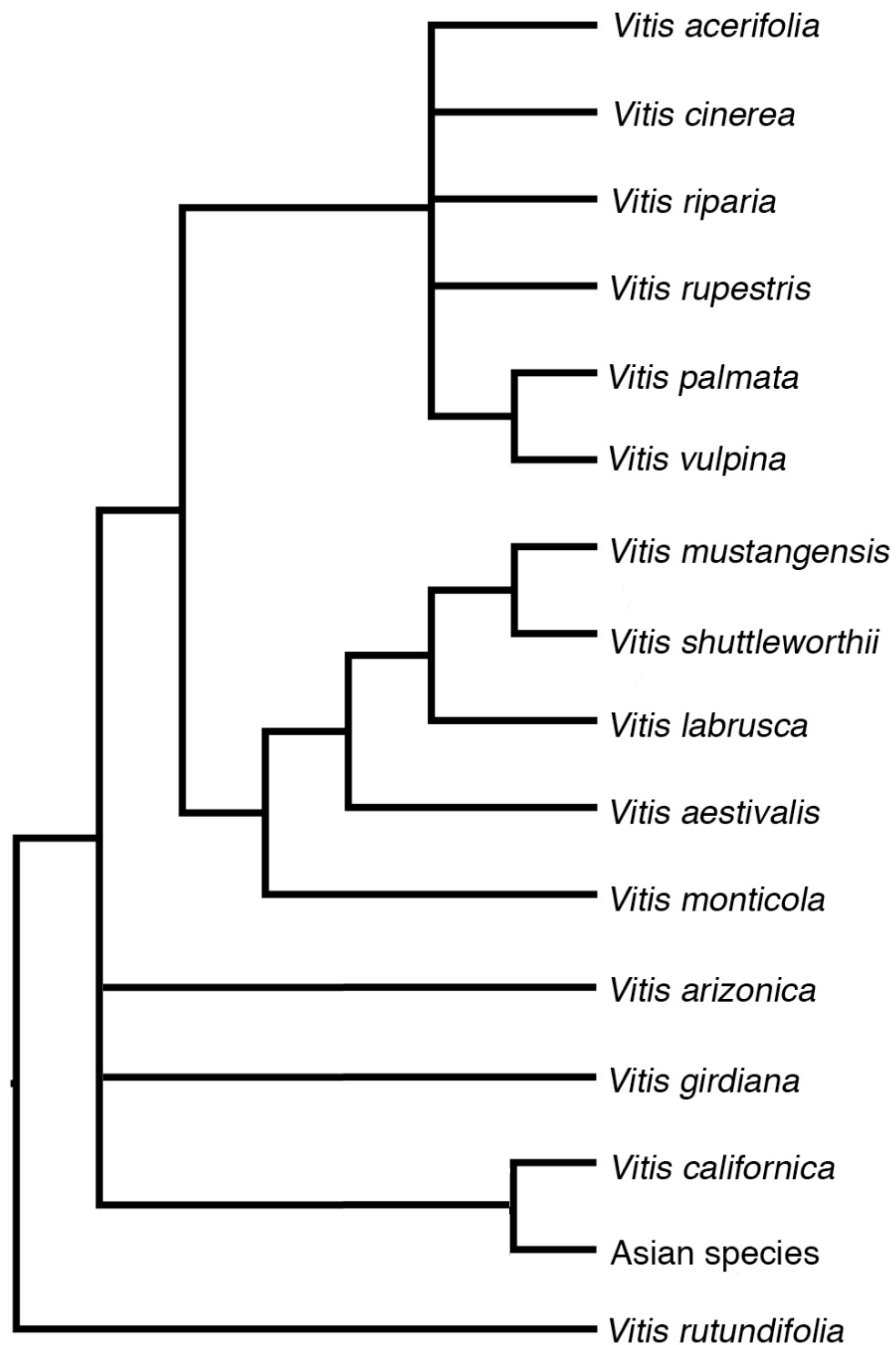


Fig. 1 Preliminary phylogenetic hypotheses about genus level relationship of native North American grapevines (*Vitis* spp.) based on nuclear genes (*RPB2*, *ADH1*, *TFL1*, *FRI*)

In addition to the taxonomic treatment, from his extensive analyses of grapevine natural history, Comeaux studied pollination biology, dispersal, phenology (phenology is the study of seasonal lifestyle events in plants and animals and their interactions with climate and environments), habitats, as well as hybridization compatibilities of native grapevines with a focus on the species present in East North America. He concluded that natural hybrids of North American wild grapevines were very rare due to selection at the seedling stage as a result of strong local adaptations. Because of the presence of various insect pollinators and genetic compatibility of native species, the lack of natural hybrids indicates that phenological and ecological differences function collectively to form the major component of reproductive isolations among native species, instead of pollinator specificity or intrinsic barriers. In the summary section of Comeaux's work about East North American native grapevines, he proposed that the pattern of reproductive isolation among these species generally agreed with Grant's (1981) theory about isolation barriers in perennial plants in the Northern Hemisphere; *i. e.*, species are largely compatible genetically but mainly separated by phenological, geographic and ecological barriers. Essentially, based on his preliminary observations on eastern species, Comeaux proposed two important hypotheses, which need to be further tested in more native North American grapevines: 1) sympatric species belong to different series, *i. e.* are distantly related, and closely related species (members of the same series) do not have overlapping distributions; 2) interfertile, sympatric species are reproductively separated by phenological and ecological barriers (Grant 1981).

One important hypothesis proposed by B. L. Comeaux (1984) is the isolation of

sympatric species by ecological and phenological barriers. One specific case is the late flowering phenotype which was thought to separate *V. palmata* from its sister species *V. vulpina* (Comeaux 1984). Other than this, no other ecological and phenological barriers have been described in previous studies. This is probably due to the large amount of time and effort needed in field surveys for ecological and phenological records. Instead of traditional fieldwork, recent biodiversity research involved Geographic Information System (GIS) technology (Hijmans and Spooner 2001). GIS consists of three components: map, data and analysis tools. In biology-related studies, data include ranges of organisms as well as parameters characterizing the habitats. The analysis tools, incorporated in many GIS applications, made it possible to process and analyze large scale data, *i. e.* to study the relationship between organisms and the environments in which they are living. The ability to process these data is necessary in biodiversity research, one example of which estimated complete ranges of all known wild potato species and identified geographic origin of potatoes by evaluating species richness (Hajmans *et al.* 2001). Furthermore, the highly integrated information and powerful analytic tools are ideal for ecological niche modeling, the field of predicting organisms' geographic ranges from occurrence records and environmental data layers. Ecological niche modeling was used in conservation, macro-ecology and evolution (Knowles *et al.* 2007).

Because phenological phenomena are sensitive and strongly associated with temporal and spatial climate variations, phenological records were used as a proxy to investigate historical climate changes (Fitter and Fitter 2002). Phenological records can also be

utilized to study evolution of plants and animals as they reflect organisms' phenotypical changes in response to environments and adaptations to specific habitats. Phenological information can be obtained from direct observation in the field, but can also be extracted from other sources. Plant specimen collection has been a standard scientific practice for over three hundreds years, and an unparalleled source of information on plant distribution and diversity. Unfortunately, before the widespread use of database management and the rise of the Internet, valuable information was scattered around the world in thousands of herbaria with very limited sharing and accessibility. Only recently have herbaria started to database information associated with their collections and make them publicly available. This transition has allowed many opportunities to re-examine and analyze the collections at an unprecedented scale and synthetic level. For example, Bradley *et al.* (1999) examined climate change as reflected by phenological changes in a small flora in southern Wisconsin, United States. With advanced GIS technology, Leimbeck *et al.* (2004) analyzed herbarium specimens with landscape and climate variables to investigate endemism of Araceae in Ecuador. Plant specimens can provide information such as geographic localities, altitude and climate in addition to their taxonomic values. With this additional information, herbarium collections are ideal for applying GIS to ecological and phenological problems *e.g.* Lavoie and Lachance (2006); Miller-Rushing *et al.* (2006); Gallaher and Leishman (2009). Specifically, the status of the specimens provides phenological records in the form of different reproductive stages. Therefore, digitization of herbaria data provide many potential high quality data sources for ecological and phenological research.

To test Comeaux's hypothesis about phenological and ecological isolation barriers among wild grapevines, specimen data with both phenological and geographic information are necessary. With phenological information, one can test the uniformity or differences of phenological phenomenon. Flowering time, a key reproductive trait, is at center stage of phenological studies of flowering plants. Flowering time differences are common prezygotic reproductive isolation mechanisms in flowering plants (Antonovics *et al.*, 2006, Silvertown *et al.*, 2006). Because grapevine flowering periods are assumed to be relatively short (about two weeks), a slight time shift in flowering time can effectively prevent gene flow, hybridization and maintain species boundaries. With detailed analyses of specimen annotation, the phenological records (flowering periods) can be extracted from plant specimen collections in herbarium databases, and used to investigate contemporary flowering time variations. The availability of large sample size has the potential to provide relatively accurate estimates of temporal distribution and reveal significant differences among taxa. For example, analysis of herbaria data of a Canadian Asteraceae species discovered delaying flowering time from 19th to 20th century, which matched the trend of climate change (Miller-Rushing *et al.*, 2006). As an important horticultural crop, grapevine and its wild relatives received consistent attention in North America. Tens of thousands of grapevine specimens are stored in herbaria throughout North America, and many of them are being digitized in recent biodiversity informatics projects. This provides us with a rare opportunity to integrate all the information collected by previous botanists. These data will help us further test hypotheses on grapevine ecology and phenology at an unprecedented scale.

Molecular biology of flowering time

Although previous studies found that flowering time divergence is among the earliest-appearing isolation mechanisms when populations begin to diverge (Antonovics *et al.*, 2006, Silvertown *et al.*, 2006), the molecular evolution accompanying flowering time divergence, is completely unknown in these plants. Broadly, genetic changes accompanying the speciation process, *i.e.*, the establishment of reproductive isolation, is the least studied area in plant speciation. Few genes directly affecting reproductive isolation, serving as speciation genes, have been recovered from plant systems (Bomblies *et al.* 2007, Presgraves 2009). All speciation genes identified so far were characterized using fine-scale Quantitative Traits Locus (QTL) mapping with the aid of genomic sequences, which represent the phenotype-genotype forward genetic approach. With the accumulation of functional studies of specific genes and advancement of molecular evolutionary analysis approaches, a candidate gene approach becomes more attractive as a reverse genetic way to connect genotype and phenotype (Templeton 1994).

The candidate gene approach to study speciation, especially the molecular basis of flowering time divergence, requires sound understanding of the genetic aspects of speciation (theories and hypotheses) and highly sophisticated analytic methods (sequence data and analytic tools). Before any studies of speciation, species concepts should be clarified, which is not an easy task in plants. If the biological species concept is to be adopted, the extent of reproductive isolation among groups needs to be evaluated and used to define species. In animals, the degree of isolation can be assessed because

intrinsic barriers can be identified from viability of hybrids. Loci responsible for Dobzhansky-Muller incompatibilities can be identified via introgression lines and QTL mappings (Presgraves 2009). However, in plants at least in the temperate Northern Hemisphere, as proposed by Grant (1981), 'species' are often interfertile without significant postzygotic isolations such as hybrid inviability or breakdown. In contrast, reproductive isolations among 'species' are usually dominated by external barriers such as geography, phenology and partitioning of ecological habitats (Grant 1981). As a result, in terms of the genic view of plant speciation, the 'speciation genes' are not limited to those causing Dobzhansky-Muller incompatibilities with intrinsic genetic effects, but also genes contributing to adaptations to local environments. Specifically, plant genomes are heterogeneous in divergence associated with speciation process. There are genes with selective advantages as species identity genes that spread quickly to the entire species range and hold the species together. These genes will have elevated divergence between species compared with other genomic regions (Morjan and Rieseberg 2004). At the same time, there are genes contributing to or responsible for local adaptations for specific populations, which will show divergence at the population level. The genome heterogeneity can be assessed by conducting genome scans of divergence within and between species or compare divergence across different genomic regions (Via 2009). Therefore, genes contributing to local adaptations and species identities can be revealed by population genetic and molecular evolutionary analyses of genomic regions without *a priori* knowledge about their functions.

Previously, such studies were limited by the lack of plant genomic information. Now

the advancement of DNA sequencing technology and availability of plant whole genome sequences will contribute to the progress in this area. With large scale sequence data at the population and species level, complicated population genetic and molecular evolutionary analyses can be conducted to reveal temporal and spatial genome heterogeneity in relation to plant speciation. For example, with molecular markers, relatively accurate phylogenies can be constructed to facilitate character evolution related to speciation. One of the best examples of such studies is illustrated by the stepwise evolution and co-evolution of pollinators and nectar spurs in the genus *Aquilegia* that led to a rapid species radiation (Hodges and Arnold 1994, Whittall and Hodges 2007). The combination of phylogenetic and biogeographic information made it possible to assess the evolutionary histories in relation to geography, as the rise of the field called phylogeography (Avice 2004). In addition, the use of molecular data greatly enhanced the resolution of low level phylogenies, which resulted in sophisticated analyses to address questions at species boundaries (Nested Clade Analysis, Templeton 1998). With these tools, the evolutionary histories of candidate genes can be investigated in combination with phenotypic, geographic and ecological information, which has the potential to identify the molecular basis of speciation. Especially, with *a priori* information on genes affecting plant flowering time, molecular evolutionary analyses of these genes may reveal molecular basis of flowering time differences.

The other key component of the candidate gene approach is the functional study of candidate genes. Flowering time, as one of the most important components in plant reproductive success, has been under intensive investigation in the plant model system

Arabidopsis, and extending to non-model systems (Glover 2007). With the fully sequenced genome of model system *Arabidopsis*, key genes in molecular pathways underlying flowering time control were identified by integrating genetic and functional information, thus, the pathways promoting flowering are largely deciphered (Simpson and Dean 2002, Putterill *et al.* 2004). Flowering in *Arabidopsis* is controlled by four pathways (autonomous, Gibberellins (GA), vernalization and light-dependent), which complement each other to ensure the transition to reproductive stage. The signals from these pathways are integrated into expression control of several genes. This process, sometimes called the integration pathway, includes flowering-time integrators *FLOWERING LOCUS T (FT)*, *SUPPRESSION OF OVEREXPRESSION OF CONSTANS 1 (SOC1)* and *LEAFY (LFY)*. For every pathway, key genes are recovered by screening flowering time mutants (Koornneef *et al.* 1991), and subsequent functional studies and gene isolation (Koornneef *et al.* 1998, Fig. 2).

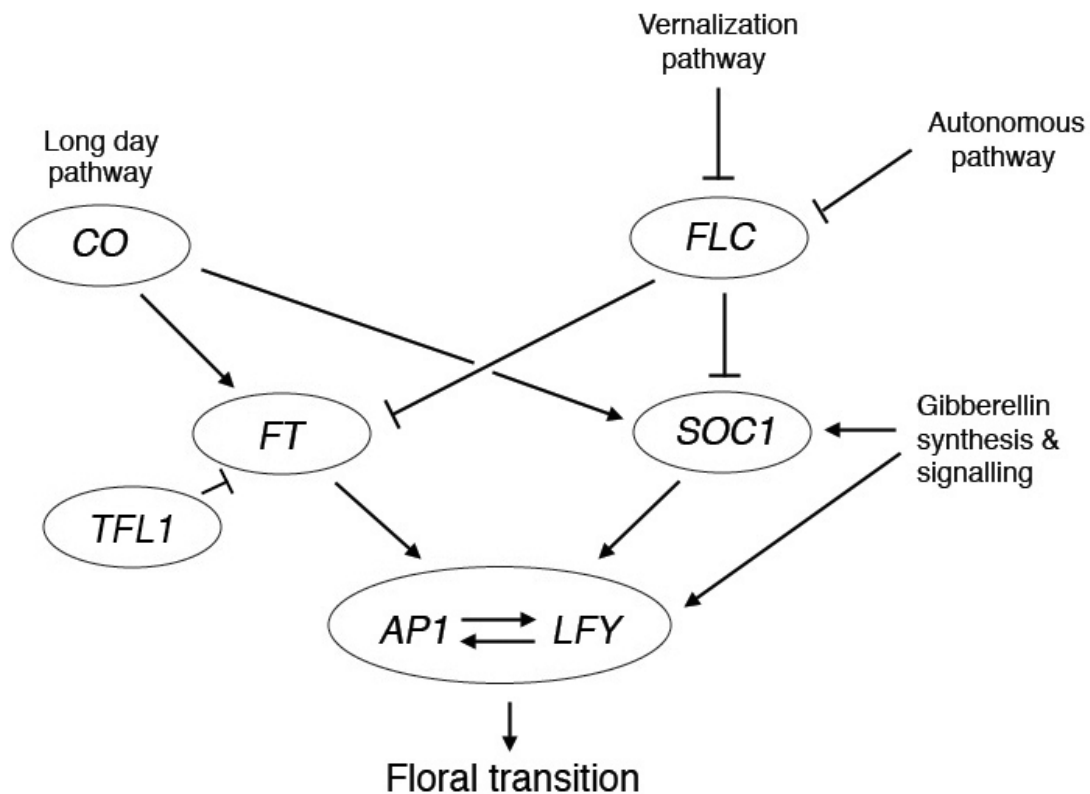


Fig. 2 Pathways and key genes controlling flowering time in *Arabidopsis*; circles indicate genes; arrows indicate activation, blunted lines indicate repression, *AP1*, *APETALA1*; *CO*, *CONSTANS*; *FLC*, *FLOWERING LOCUS C*; *FT*, *FLOWERING LOCUS T*; *LFY*, *LEAFY*; *SOC1*, *SUPPRESSOR OF OVEREXPRESSION OF CO 1*; *TFL1*, *TERMINAL FLOWER 1* (adapted from Alonso-Blanco *et al.* 2009)

In the center of flowering time pathways is the *FT* gene, which integrates signals from other pathways and interacts with meristem identity genes (*APETALA1*(*AP1*), *LEAFY*, *etc.*) to initiate floral meristem development. Interestingly, florigen, the mysterious flowering signal generated in leaves and transported into floral meristems, was recently found to be the protein encoded by *FT*, and transported through phloem (Jaeger and

Wigge 2007). Once FT protein arrives in the meristem, it triggers local *FT* expression via a positive feedback loop so that its expression can be maintained and stabilized locally (Huang et al. 2005). *FT* orthologs play the same role in flowering of a tree, *Populus*, and affect flowering time in both rice and wheat, suggesting conservation of *FT* function across plant taxa (Kojima et al. 2002 and Yan et al. 2006). Interestingly, *FT* belongs to a gene family containing a phosphatidylethanolamine-binding domain. Another member of the family, *TERMINAL FLOWER 1 (TFL1)*, one of the floral identity genes, is also a mobile signal controlling meristem identity throughout plant life cycle (Conti and Bradley 2007). *FT* and *TFL1* show antagonistic functions: *FT* usually acts as flowering promoter while *TFL1* acts as flowering repressor (Kobayashi et al. 1999). Indeed, Hanzawa et al. (2005) suggested that as few as one single amino acid change can reverse the functions of *FT* and *TFL1*.

Unlike *FT*, which does not respond to the signal from GA pathway, another flowering-time integrator *SOC1* integrate signal input from every flowering-time pathway as it directly responds to Gibberellins (Borner et al. 2000). It has been shown that the *Arabidopsis* *SOC1* promoter responds to both *FLOWERING LOCUS C (FLC)* and *CONSTANS (CO)* signals, which repress and promote flowering, respectively, suggesting that *SOC1* can handle conflicts of signal as an integrator (Hepworth et al. 2002). The third signal integrator, *LFY*, is also a floral meristem identity gene. *LFY* acts downstream of *SOC1* and is directly influenced by Gibberellins, so it is also considered an integrator (Jack 2004).

The signal integrator of the photoperiod pathway, *CO* encodes a protein with two zinc finger domains, suggesting transcription factor activities. Mutations in this gene were shown to directly affect flowering time in *Arabidopsis* (Putterill *et al.* 1995). Furthermore, *CO* was shown to directly interact with *FT* and *SOC1*, the integrators, and promote flowering process (Kardailsky *et al.* 1999, Hepworth *et al.* 2002). Some plants have to experience a period of low temperature to make the transition from vegetative growth to flowering, a process called vernalization. A series of genes respond to the environmental cues such as low temperature and control the signal for the vernalization process. *FLC* is the signal integrator of not only vernalization but also autonomous response genes (Putterill *et al.* 2004). It has been shown that the FLC protein binds to the first intron of *FT* and inhibits its transcription (Helliwell *et al.* 2006). Interestingly, *FLC* expression and repression was affected by upstream genes via various types of chromatin modifications. The 'active' modification mechanisms for promoting *FLC* expression includes acetylation of core histone tails, histone H3 lysine-4 (H3K4) methylation, H2B mono-ubiquitination, H3 lysine-36 (H3K36) di- and tri-methylation and deposition of the histone variant H2A.Z, while the 'repressive' activations include histone deacetylation, H3K4 demethylation, histone H3 lysine-9 (H3K9) and H3 lysine-27 (H3K27) methylation, and histone arginine methylation (He 2009).

Some genes directly interact with *FLC* thus indirectly affect flowering time in *Arabidopsis*. For example, in natural *Arabidopsis* populations, Le Corrie (2005) revealed that frequency of null alleles disrupted by indels in exon 1 of *FRIGIDA* (*FRI*) gene was related to flowering time. In a population, the higher frequency of null alleles, the later

the average flowering time of that population. Several geographically isolated natural populations already diverged in terms of null allele frequencies and as a consequence, they showed significantly different flowering times (Le Corrie 2005, Fig. 3).

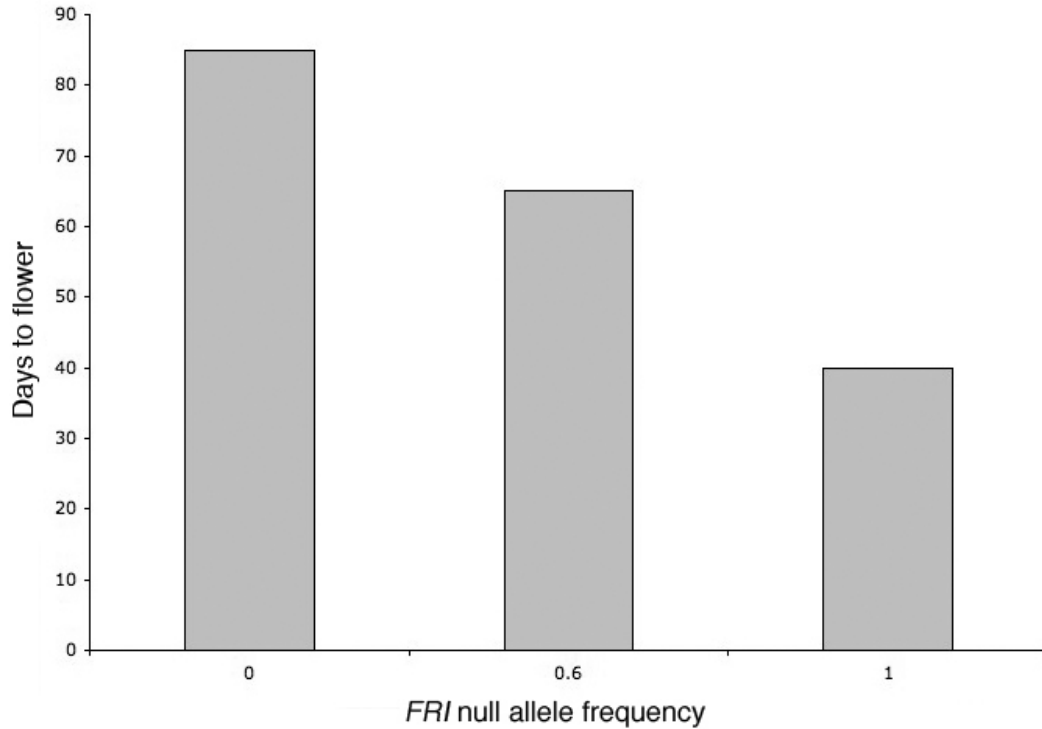


Fig. 3 Correlation between frequencies of *FRI* null alleles and flowering time in natural *Arabidopsis* populations (adapted from Le Corrie 2005)

Another *FLC* upstream regulator, *FCA*, works in a self-regulatory way in *Arabidopsis* (Macknight *et al.* 1997). In *Arabidopsis*, *FCA* exhibits alternative transcription in an autoregulated fashion to control the amount of fully functional protein products, and thus flowering time (Quesada *et al.* 2003). As many as four transcript forms have been identified although 90% of all transcripts were of two kinds, transcript α and γ (MacKnight *et al.* 1997). The functional FCA protein encoded by transcript γ contains two RNA recognition motifs (RRM) and one WW domain. The FCA protein interacts

with FY, a polyadenylation factor, forming a protein complex with both RNA recognition and cleavage-polyadenylation abilities. The FCA-FY complex directs cleavage and polyadenylation within intron 3 of *FCA* pre-mature mRNA, producing transcript form β , which leads to a non-functional protein in flower development (MacKnight *et al.* 1997, Macknight *et al.* 2002, Fig. 4). As a result, FCA protein promotes flowering but limits its own amount by negative-feedback autoregulation of mature mRNAs. Multiple transcript forms also have been identified in cabbage, pea and rice, suggesting that *FCA* autoregulation may be conserved among these plants (Macknight *et al.* 2002, Lee *et al.* 2005).

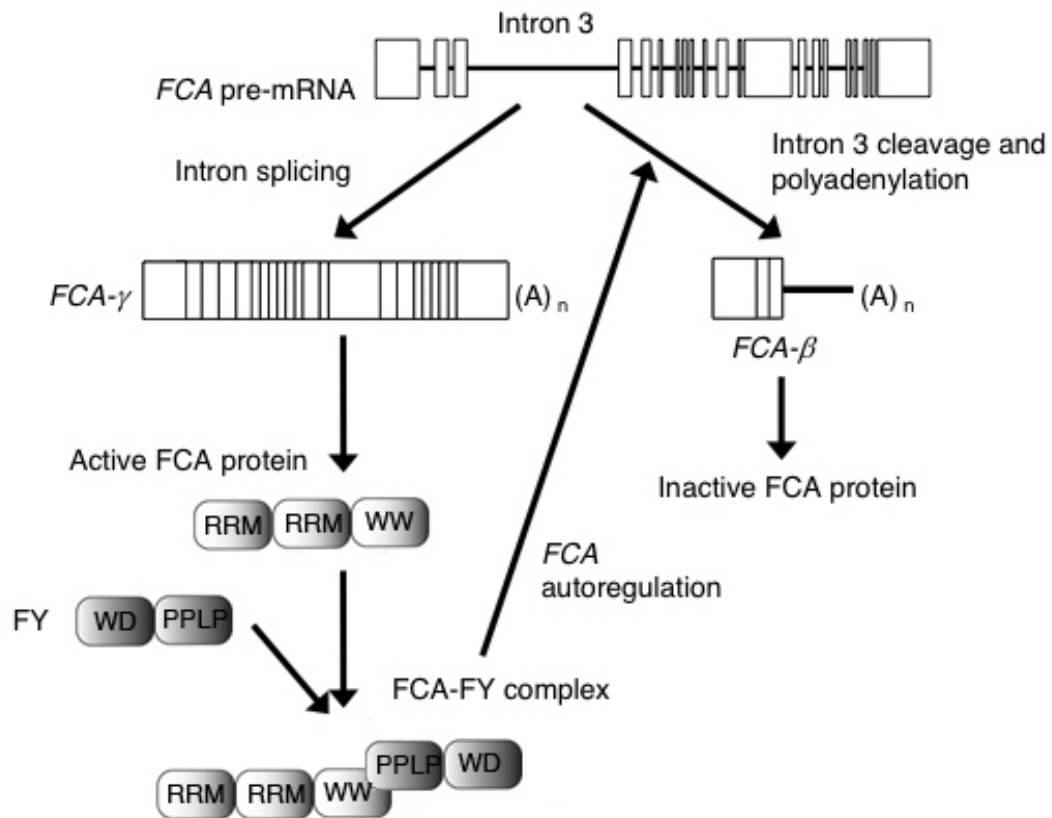


Fig. 4 Negative feedback loop of *FCA* self-regulation in *Arabidopsis*; Boxes represent exons, lines represent introns, together they represent models of transcripts; protein

domains are represented by shaded boxes, together they present protein models (adapted from Macknight *et al.* 2002)

Known natural genetic variations underlying plant development involve various types of mutations in key genes, including point mutations in coding regions (amino acid replacements), mutations in promoter regions and introns, transposable element (TE) insertions in promoter or intronic regions and attenuated or abolished gene expression due to unknown reasons. Over half of identified genetic basis of plant development variations are caused by point mutations that change amino acids in proteins (Koornneef *et al.* 1998). In addition, because of the prevalence of TEs in plant genomes (Bennetzen 2005), plant phenotypes are affected by TE insertions into functional important genes, and flowering time genes are no exceptions. TE insertion in *Arabidopsis FLC* intron 1 caused mis-expression of this gene and altered flowering time (Michaels *et al.* 2003). All of these data showed that the changes in flowering time could be traced down to molecular level. Along with the accumulation of plant genomic data (Tuskan *et al.* 2006, Jaillon *et al.* 2007), the search-for-molecular-basis process is shifting from time consuming phenotype-to-genotype forward genetic approaches to association studies between molecular features and phenotypes.

In summary, this study is a comprehensive analysis of flowering time differences and their possible molecular basis using a candidate gene approach and associative method. The study aims to test the following hypotheses: hypotheses of phylogenetic relationships among native North American grapevines proposed based on morphology-based studies;

Native North American grapevines are reproductively isolated by ecological and phenological factors, specifically, flowering times are different among closely related species, and serves as a major component of reproductive isolation; molecular evolution of genes controlling flowering is correlated to the evolution of flowering time divergence in North American *Vitis* spp. Molecular phylogeny of North American *Vitis* spp are inferred using nuclear markers and a full scale phylogenetic analysis. The second component involves large-scale analysis of herbarium specimens of native North American grapevines, in order to estimate geographic ranges, ecological niches and phenological divergence among these species, the results of which will help test Comeaux's hypothesis about phenological and ecological reproductive isolations among grapevines. The third section of the study characterizes flowering time genes in native North American grapevines and conduct molecular evolutionary analyses of these genes at different taxonomic levels. Combining results from all components, the connections between flowering phenology and evolutionary genetic changes underlying these phenological changes may be discovered in wild grapevines to improve our understanding of plant speciation process.

Chapter 2

Characterization of grapevine *RPB2* and phylogenetic analysis

Abstract

Nuclear markers are increasingly involved in plant phylogenetic analyses at all taxonomic levels. The *RPB2* gene, encoding the second largest subunit of RNA polymerase II, is becoming an important nuclear phylogenetic marker because of its conserved coding regions and variable intronic regions. Native North American grapevines represent a significant component of biodiversity in *Vitis*. Despite repeated and extensive taxonomic study, the natural history and phylogenetic relationships of this economically significant genus are still poorly understood. *RPB2* gene in North American grapevines are characterized and incorporated into phylogenetic analyses at both higher taxonomic levels and the genus level, to investigate the molecular evolution of *RPB2*, the position of the Vitaceae in angiosperm phylogeny, as well as phylogenetic hypotheses of species relationship within *Vitis*. In addition to the functional *RPB2* gene, pseudogene fragments of *RPB2* duplicates, belonging to a distinct angiosperm *RPB2* lineage, are discovered in Vitaceae plants *Cyphostemma*, *Parthenocissus* and *Vitis*. Higher level phylogenetic analyses place the duplication events close to the diversification of core eudicots. Within *Vitis*, combined multi-locus phylogenetic analyses based on both concatenated data and coalescence processes revealed strong phylogenetic signals supporting major native grapevine groups proposed by previous morphological and

biogeographic studies. Coalescence-based species tree estimation suggests polytomies associated with uncertain relationships instead of incorrect patterns supported by concatenated data.

Introduction

Angiosperm phylogeny has benefited greatly from the analysis of molecular data. Using molecular data for phylogenetic inference can reduce the effects of convergence or parallelism, which are very common in plant morphological evolution (Scotland *et al.* 2003). Molecular data, especially DNA sequences from plant chloroplast genes, have become the primary source of phylogenetic data for plant systematics. However, chloroplasts are uni-parentally transmitted haploid organelles. The molecular evolution of chloroplasts can only reveal evolutionary history of one parental lineage. Flowering plants usually have diploid or polyploid genomes, which contain information on evolutionary history from both parental lineages. In addition, plant nuclear genomes with coding and non-coding regions are heterogeneous in terms of evolutionary constraint and substitution rate, which makes different genomic regions suitable for inferring phylogenies at different taxonomic levels. For example, coding regions of conserved, low-copy number nuclear genes are ideal for higher level phylogenetic analysis (stronger functional constraint on proteins encoded); while highly variable introns can be used at lower level such as genera and species. Variable nuclear regions are especially valuable at species boundaries, where allelic variations revealed by nuclear genes can help resolve the lineage sorting process during species formation (Sang 2002).

The genus *Vitis* L. (Vitaceae) is a widespread group of woody vines common in the temperate Northern Hemisphere. More than 50 *Vitis* species are recognized worldwide with significant diversity in Asia and North America. The common cultivated species,

Vitis vinifera, was domesticated around 7000 years ago (Zohary and Spiegel-Roy 1975) and is economically significant for the production of grapes, raisins, currants and wine. North America harbors a significant portion of *Vitis* biodiversity, including 15~20 native species genetically distinct from Old World *Vitis* species. However, because of the great morphological variability within the genus, taxonomists made constant and substantial efforts to produce treatments for *Vitis* in the past century. T.V. Munson's proposal of eight series as a start point of modern *Vitis* taxonomy (Munson 1909), are followed by subsequent treatments or descriptions of *Vitis*. A recent *Vitis* taxonomic system by B. L. Comeaux followed the same series but proposed an improved genus description and key to identification of wild grapevines in East North America (including *Vitis rotundifolia*, *Vitis labrusca*, *Vitis aestivalis*, *Vitis vulpina*, *Vitis cinerea*, Comeaux 1984). More recently, M. O. Moore focused specifically on the *Vitis* species in Southeast United States. He eliminated errors in *Vitis* specimen typifications and clarified many synonyms with extensive herbaria data analysis and fieldwork (Moore 1991). As all of the *Vitis* taxonomic studies are based on morphological characters, there is instability and difference among all of the systems due to high morphological variability of *Vitis*. As a result, vigorous testing of phylogenetic hypotheses are urgently needed, especially based on molecular markers, to incorporate or use systematics and evolutionary histories in the classifications.

In this study, *RPB2* gene is characterized for several native North America species in the genus *Vitis*, as well as additional taxa in the family. *RPB2* genes are emerging nuclear phylogenetic markers in both plants and fungi (Liu *et al.* 2006; Oxelman *et al.* 2004). At

higher taxonomic levels, the family Vitaceae has been included in the Rosids but is usually considered as the sister group to all Rosids (Wang *et al.* 2009). Thus, the family is at the boundary of a transition from early-branching eudicots and more advanced core eudicots and possibly a part of the initial radiation of eudicots. Vitaceae *RPB2*s are incorporated into a larger data set with angiosperm *RPB2* genes to test the various possible positions of Vitaceae in angiosperm phylogeny and its role in eudicots diversification. Within the genus *Vitis*, *RPB2* is also characterized as an additional nuclear molecular marker to investigate *Vitis* phylogeny. Selected variable regions of *RPB2* are surveyed in multiple species to resolve the phylogenetic relationship within the genus. Specifically, coalescence-based species tree methods are used to analyze multi-locus data to test prior phylogenetic hypotheses in *Vitis*.

Materials and methods

Molecular cloning and sequence analysis

DNA of *Vitis* species, including: *V. acerifolia*, *V. aestivalis*, *V. arizonica*, *V. californica*, *V. cinerea*, *V. girdiana*, *V. labrusca*, *V. mustangensis*, *V. monticola*, *V. palmata*, *V. riparia*, *V. shuttleworthii*, *V. vinifera* and *V. vulpina*, were extracted from fresh leaf tissue using a modified CTAB DNA extraction protocol (Doyle and Doyle 1987). Polymerase Chain Reaction (PCR) primers were designed in exon regions based on homology to the *Arabidopsis RPB2* gene. *Vitis* specific primers were designed based on the amplified fragments to recover the complete coding region of *RPB2*. Additional nuclear genes include *ADHI* and two additional nuclear markers (*FRIGIDA (FRI)* exon 1

and *TERMINAL FLOWER 1 A (TFL1A)*) (both are flowering time genes, see Chapter 4). Additional primers were prepared for each primary primer in case sequence divergence among native species prevented degenerative primary primers from working (Table 1). PCR were conducted using New England BioLab Taq polymerase (including standard 10X PCR buffer and dNTPs) and standard amplification program (denaturation at 72°C for 3 minutes, annealing of primers at 58°C for 0.5 minutes, elongation at 72°C for 1.5~2 minutes, 35 cycles of amplification). To obtain the 5' flanking regions beyond *RPB2* start codon, asymmetric PCR approach was used. Reverse primer in intron 1 was used in a regular PCR reaction with a modified program: step 1: 72°C denaturation for 3 minutes, annealing of primers at 58°C for 0.5 minutes, elongation at 72°C for 1 minute, 15 cycles of amplification, creating DNAs extending from exon 1 to 5' flanking regions; step 2: 72°C denaturation for 1 minute, 48°C for annealing, 72°C elongation for 1 minute, 35 cycles of amplification, creating random annealing of primers onto 5' flanking regions and complete the regular PCR reaction. All PCR products were cloned using Topo-TA cloning kit and harvested by PCR using plasmids as templates. The final products were cleaned and submitted for sequencing by Auburn University Genome Sequencing Laboratory. Sequence data were submitted to GenBank (accession numbers: *ADH1*, GU968573-GU968582; *FRI*, GU947839-GU947848; *RPB2*, GU947829-GU947838; *TFL1A*, GU947819-GU947828, *Cyphostemma RPB2-I*, GU947849, *RPB2-D*, GU947853; *Leea RPB2-I*, GU947850; *Parthenocissus RPB2-I*, GU947851, *RPB2-D*, GU947852; *Vitis RPB2-D*, GU947854). Additional *RPB2* data were collected from GenBank based on keyword and BLAST searches (Appendices Table 1).

Table 1 Primers used to characterize *Vitis RPB2*, *FRI* and *TFL1A*

Locus	Name	Region	Sequences
<i>ADHI</i>	5UTR3	5'-UTR	5'-GTCATCAAAATCCACTAGAC-3'
	3UTR2	3'-UTR	5'-CAATGCTACTCCAAAATACAC-3'
<i>RPB2</i>	5UF	5'-UTR	5'-GTTTCGTTACCAGGTTCTTTG-3'
	X1F	Exon 1	5'-CTTTGAAGAGAAGGGTTTGGTG-3'
	I6F	Intron 6	5'-TTTCATTCTTGTTCATTACCAGA-3'
	I7F	Intron 7	5'-TGAGCTTTGGATGTTGCTTG-3'
	I7R	Intron 7	5'-CAAGCAACATCCAAAGCTCA-3'
	I8F	Intron 8	5'-TACTAAGCTGTTTCTGTGCATGA-3'
	I8R	Intron 8	5'-TCATGCACAGAAACAGCTTAGTA-3'
	I17F	Intron 17	5'-GCGGACATGTTTCATATTTTCTGT-3'
	I17R	Intron 17	5'-ACAGAAAATATGAACATGTCCGC-3'
	3UR	3'-UTR	5'-TCTTTCTTCAGGGAGCAATG-3'
<i>FRI</i>	X1F	Exon 1	5'-CTGCCAAACTTGTACTTGAATGC-3'
	I1R	Intron 1	5'-GTTTAGCATCCCGGAAGGA-3'
<i>TFL1A</i>	5UF	5'-UTR	5'-GCCTTCCAAGAGACCAAGAGT-3'
	3UR	3'-UTR	5'-TGATCTTCCCGTTGGTTATTG-3'

Phylogenetic analyses

Sequence alignments were conducted using ClustalW (Thompson *et al.* 1994) with manual corrections where necessary. Phylogenetic analyses were conducted using PAUP* (version 4.0b11 Swofford 2002) with both parsimony and likelihood methods. Within the genus *Vitis*, multi-locus phylogenetic analysis was conducted by PAUP* and BEST (Liu and Pearl 2007) using four genes with the largest taxa sampling. *ADHI*, *RPB2* (fully characterized but only the region from exon 7 to 18 was included), *FRI* exon 1 and *TFL1A* complete coding regions from 11 species (*V. aestivalis*, *V. californica*, *V. cinerea*, *V. girdiana*, *V. labrusca*, *V. mustangensis*, *V. monticola*, *V. palmata*, *V. riparia*, *V. shuttleworthii* and *V. vulpina*, *V. vinifera*, as an descendent of Old World species, was used as an outgroup) were included in the multi-locus phylogenetic analyses. Multi-locus analyses with concatenated data were conducted with PAUP* using parsimony methods.

Likelihood-based phylogenetic inference was conducted with RAxML (Stamatakis 2006). First, four genes are designated as four partitions and the homogeneity of DNA evolution models among partitions was tested. The DNA evolution model for the multi-locus data set was selected by ModelTest (Posada and Crandall 1998) and used in the likelihood analyses. Heuristic searches for topology were conducted with 100 replications, sequences added randomly and TBR tree swapping. Supports to the resolved nodes were evaluated by 1000 bootstrap replicates. Species tree inference based on coalescence was conducted by BEST with independent DNA evolution models for each locus in a Bayesian framework. The default population genetic parameters were used as initial values followed by two 1 million-generation Markov Chain Monte Carlo (MCMC) runs. Tree topology and parameter estimates were sampled every 100 generations with burnin at 2500th generation. Bayesian posterior probabilities were summarized and indicated in the final tree. Likelihood-based tests of tree topology were conducted using both non-parametric and parametric methods. The non-parametric KH and SH tests (Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999) were conducted using PAUP* by comparing a priori specified competing tree topologies generated by BEST and RAxML. The parametric tests of tree topology were conducted following methods described by Goldman *et al.* (2000), using PAUP* and Seq-Gen (Rambaut and Grassly 1997). Tree topology suggested by BEST was considered the null hypothesis and compared with topology estimated by maximum likelihood (RAxML). First, likelihood scores were calculated (DNA evolution parameters were estimated) using data and BEST and RAxML topologies, respectively (as well as their difference). Second, the ML estimates of parameters and BEST tree (topology and branch length) were used to

generate 1000 simulated data sets. Third, each of the 1000 data sets went through the first step and likelihood scores based on BEST tree and RAxML tree (estimated for each data set) were calculated, which generate a null distribution of likelihood differences. Finally, the true likelihood difference (from actual data) was compared with the 95% rank of likelihood difference distribution so the significance can be determined. Ka, Ks values for Vitaceae *RPB2*-Ds were estimated using DNASP (Librado and Rozas 2009).

Identification of conserved non-coding regions in 5' flanking sequences of *RPB2* was conducted with FootPrinter (Blanchette and Tompa 2003) using default parameters.

Results

Characterization of *RPB2* orthologous and paralogous genes

The *Vitis RPB2* consists of 25 exons and 24 introns, covering a 7.7kb-long genomic region. The size of exons and introns are between from 50bp and 300bp, except for a 1kb-long intron one, significantly larger than other introns and exons. The entire coding region (exons) of *Vitis RPB2* is conserved compared to *Arabidopsis RPB2* gene, reflecting the strong functional constraint on the protein encoded by *RPB2*. An expanded investigation recovered the orthologous copy of *RPB2* in Vitaceae genera other than *Vitis*. *RPB2* genes were recovered in *Leea*, *Parthenocissus* and *Cyphostemma*. All of the *RPB2* showed strong conservation with *Vitis* and *Arabidopsis RPB2*, with no frameshift substitutions, suggesting that they are the fully functional *RPB2* copies in these genera.

Another copy of *RPB2* was discovered in multiple *Vitis* species when using degenerate

primers to characterize coding regions from exon 20 to exon 24. Further trials of recovering more duplicate *RPB2* failed, suggesting that this copy consists of a fragment of full length *RPB2* only. Because full-length functional *RPB2* were discovered and designated as *RPB2-I* in previous literature, the newly recovered *RPB2* fragment is named as *Vitis RPB2-D* and the full functional copy *RPB2-I*, following the nomenclature in other large-scale *RPB2* samplings. The *RPB2-D* fragments are also discovered in Vitaceae genera other than *Vitis*, including *Parthenocissus* and *Cyphostemma*. In the other two genera sampled in this study, *Leea* and *Cissus*, no trace of *RPB2-D* was recovered. All Vitaceae *RPB2-D* fragments share four frameshift mutations in exon 22 and exon 23 (three 1 bp insertions and one 2 bp deletion, Fig. 5, two not shown in alignment). Ratio of non-synonymous to synonymous nucleotide substitutions (K_a/K_s) in the coding regions from pairwise comparisons between Vitaceae *RPB2-D* fragments are very close to 1, indicating a lack of functional constraint or purifying selection (Table 2). Based on the high K_a/K_s ratio and partial existence of Vitaceae *RPB2-D* fragments, is shown that the Vitaceae *RPB2-D* fragments are degenerated pseudogenes originated from duplication events. Although none of them are complete and functional as *RPB2* genes, their presence in the family Vitaceae is of significance as potential source of phylogenetic signals.

Table 2 Ka/Ks of *RPB2*-Ds in Vitaceae

	K_s^a	K_a	K_a/K_s
<i>Vitis</i> vs <i>Cissus</i>	0.0233	0.0169	0.7253
<i>Vitis</i> vs <i>Parthenocissus</i>	0.0186	0.0112	0.6022
<i>Cissus</i> vs <i>Parthenocissus</i>	0.0093	0.0056	0.6022
	K_s^b	K_s	K_a/K_s
<i>Vitis</i> vs <i>Cissus</i>	0.0711	0.0169	0.2377
<i>Vitis</i> vs <i>Parthenocissus</i>	0.0348	0.0112	0.3218
<i>Cissus</i> vs <i>Parthenocissus</i>	0.0349	0.0056	0.1605

a. Silent substitutions in non-coding regions and synonymous substitutions

b. Synonymous substitutions only

RPB2 phylogeny and gene duplication in Angiosperms

All Vitaceae *RPB2* genes, including both *RPB2*-I genes and *RPB2*-D fragments, are incorporated into a large *RPB2* nucleotide sequence dataset covering major angiosperm groups. Both Maximum Parsimony (MP) and Maximum Likelihood (ML) trees clearly yield two major *RPB2* lineages, designated as I and D clade in previous studies, respectively (Fig. 6). As expected, all Vitaceae *RPB2*-I genes form a monophyletic group inside the I-clade and Vitaceae *RPB2*-D fragments are in D-clade. In a bootstrap analysis for the MP tree, the D and I clades are supported by 80 and 90, respectively, and Vitaceae *RPB2*-D groups are supported by 100. Although all Vitaceae *RPB2*-D fragments are pseudogenes accumulating substitutions completely without selective constraint, the phylogenetic signal indicating their origin as D-copy have been well preserved.

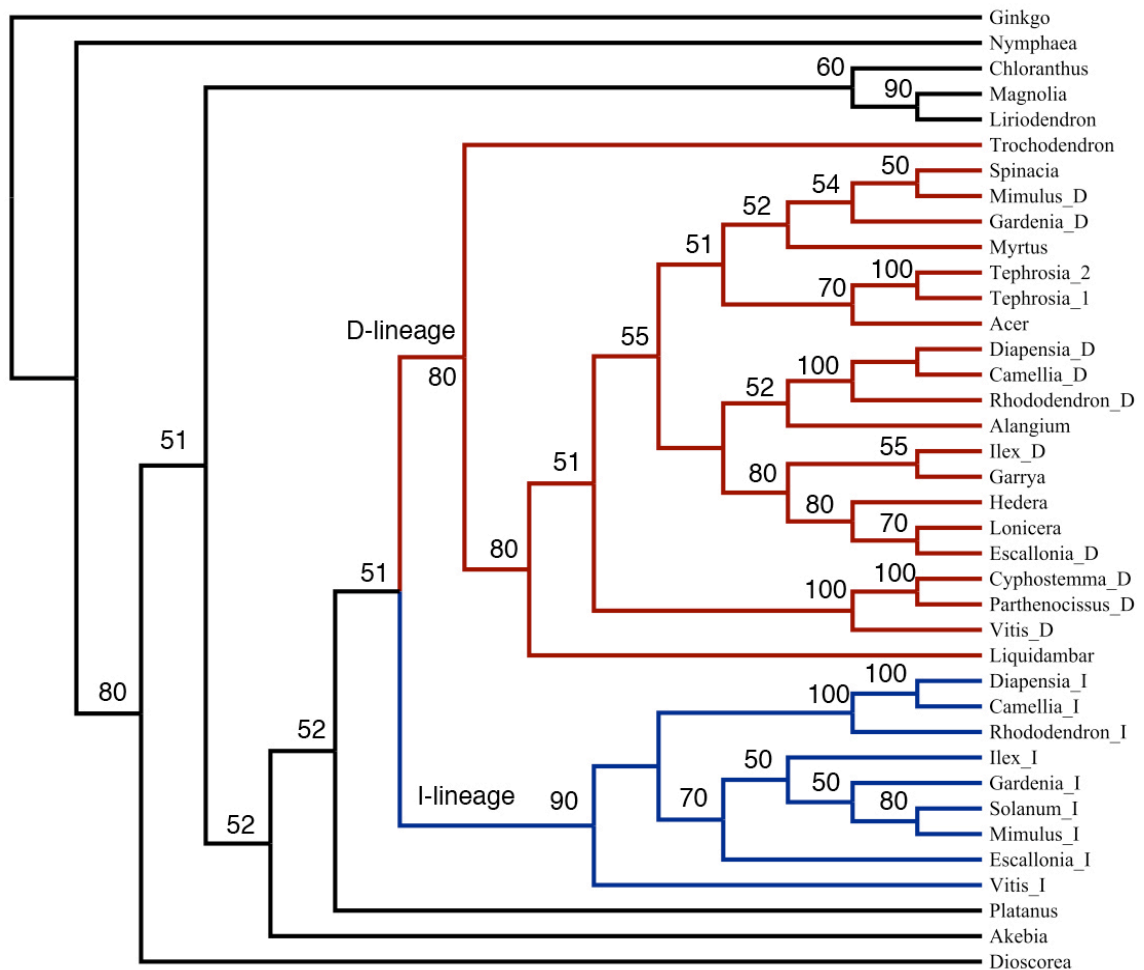


Fig. 6 D (red) and I (blue) lineages of *RPB2* in angiosperm, bootstrap (100 replicates) values are shown on the nodes

Combined phylogenetic analyses of *RPB2* and other nuclear markers

Heterogeneity of DNA evolution model was revealed by model testing processes. *FRI* exon 1 substitution patterns can be adequately explained by the Jukes-Cantor model, *TFL1A* by General Time Reversible (GTR) model, and both *ADH1* and *RPB2* by GTR model with proportions of invariants (detailed information about *FRI* and *TFL1A* in Chapter 4). Combined phylogenetic analyses based on concatenated multi-locus data and coalescence process of multiple genes show congruence and differences in tree topology

(Fig. 7). Several previously established groups such as Southeastern species (*V. aestivalis*, *V. mustangensis*, *V. labrusca* and *V. shuttleworthii*), and *V. vulpina*/*V. palmata* group were supported in both trees, suggesting strong phylogenetic signals supporting these groups independent of analytical methods. However, the coalescence-based method supported previous hypotheses that *V. californica* and *V. girdiana* do not belong to the 'normal' North American species, while in concatenated-data-based method, these species were incorrectly grouped with other North American species. When overall tree topologies estimated from two methods were compared, species tree estimated by BEST resulted more polytomies. Both non-parametric and parametric tests of tree topology differences suggested that there was significant differences between phylogenetic hypotheses generated by BEST and estimated by maximum likelihood (KH test $P = 0.008$, SH test $P = 0.005$, parametric test $P < 0.001$), suggesting species tree estimation based on coalescence provided a completely different set of phylogenetic hypotheses.

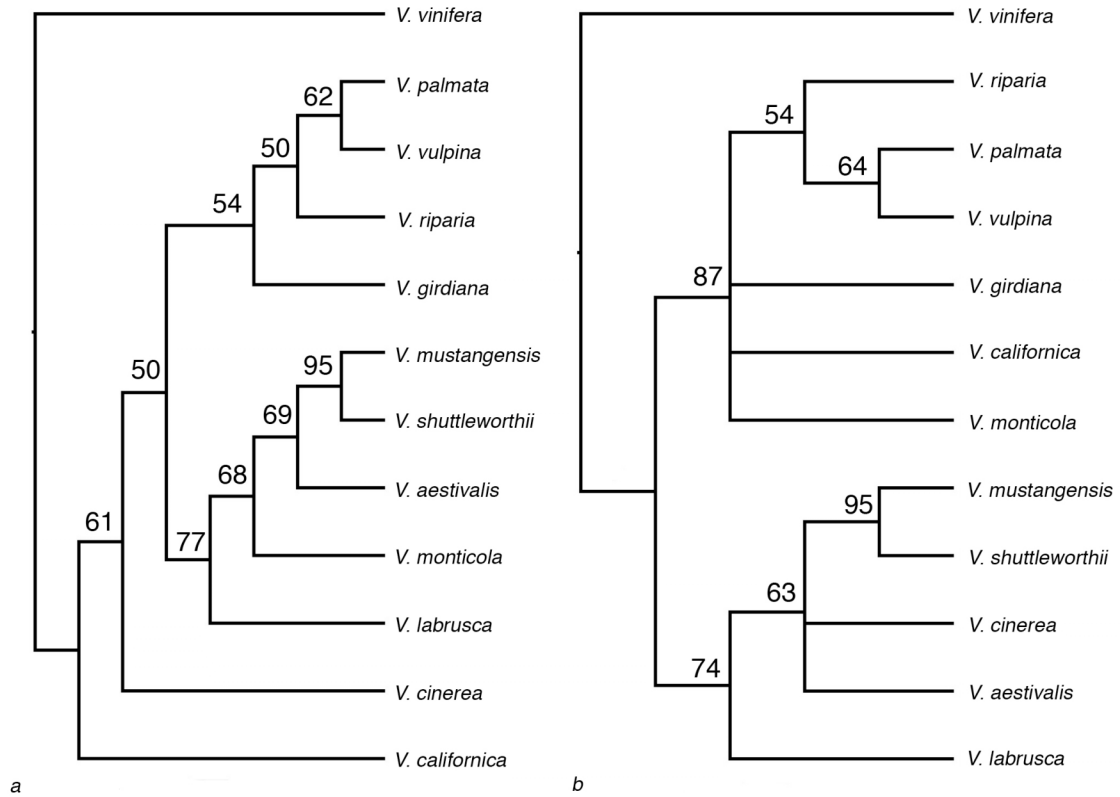


Fig. 7 Comparison of phylogenetic tree topologies within genus *Vitis* inferred from: a. concatenated multi-locus phylogenetic analyses, numbers on the nodes indicate bootstrap supports; and b. coalescence-based multi-locus phylogenetic analyses, numbers on the nodes indicate Bayesian posterior probabilities

Discussion

RPB2 duplication and angiosperm phylogeny

The *RPB2* duplication event is close to the divergence of core eudicots from early-branching eudicots. According to APG II tree (Angiosperm Phylogeny Group, A. P. G., 2003), the duplication happened before the divergence of Trochodendrace from Proteals, a time point suspected when one of the whole genome duplication events happened in

evolutionary history of flowering plants (Tang *et al.* 2008). Oxelmann *et al.* (2004) suggested that the duplication event which led to D and I paralogs happened before the origin of core eudicots. Based on the currently accepted angiosperm phylogeny (A. P. G. 2003), it was proposed that there had to be 7 losses of the I-lineage and 1 loss of the D-lineage. They also suggested that if they place Vitaceae, whose status in eudicots has not been well supported by other studies at the time, outside the eudicot group, the phylogeny would require only 5 losses of the I-lineage and no loss of the D-lineage. They reported two traces of pseudogene-like I copies in *Hypericum* and *Valeriana*, as well as one D pseudogene in *Luculia*, which are recent and taxon specific events. Here, all sampled species from Vitaceae group except *Leea* and *Cissus*, have recognizable traces of D-copy pseudogenes, which suggested a common ancestry of two copies and independent pseudogenization in this lineage. Although only four to six exons (dependent on species) of *RPB2* D-copy remained in Vitaceae, the phylogenetic signals provided enough evidence to group these fragments into the D-clade. Instead of a descendent of pre-duplicated ancestral *RPB2*, members of Vitaceae group do have two *RPB2* copies, one of which only left barely recognizable traces. This excludes the possibility that Vitaceae is a more basal group in angiosperm and confirms its status as a member of Rosids by placing Vitaceae after the duplication event, which is supported by other studies based on cpDNA. In addition, the loss of D-copy in Vitaceae lineage represents the only case in which a D-copy, not an I-copy, lost its function in a whole group (Oxelmann *et al.* 2004). Based on extensive expression analysis, Luo *et al.* (2007) suggested that *RPB2*-Ds were constitutively expressed in most taxa sampled, while *RPB2*-Is often showed tissue-specific expression. Because of this, independent losses of a D-copy are less likely unless

regulatory mutations can make the shift from a tissue-specific expressed gene to a constitutively expressed one. With a fully functional *RPB2-I* and *RPB2-D* pseudogene, loss of D-copy apparently occurred in Vitaceae. The most likely scenario is that Vitaceae diverged from other core eudicots at a very early stage after *RPB2* duplication, when differentiation of expression patterns between D and I copies had not been well-established. In that case, *RPB2-Is*, instead of D copies, in the common ancestor of Vitaceae plants may maintain the ability to be constitutively expressed and been preserved by selective constraints.

In order to find any possible *cis*-regulatory changes which account for the potential differentiated expression patterns between D and I copies, the 5' flanking sequences of *RPB2* in *Vitis*, the major genus in Vitaceae, was investigated by TAIL PCR approach. Several conserved non-coding sequences in this 5'-flanking region have been identified by comparing with available *RPB2* 5' upstream sequences of *Petunia*, *Antirrhinum*, *Rhododendron*, *Arabidopsis* D-copies, tomato, tobacco I-copies and unannotated genomic sequences near poplar *RPB2*. The limited availability of upstream sequences (usually no more than 200bp) reduces the power of *cis*-regulatory sequence identification by the phylogenetic footprinting approach. However, universal promoter sequences TATA-box are identified in all I-copies from tomato, tobacco, grape and unknown copy from poplar genome draft sequences, as well as *Arabidopsis* D-copy (this position of TATA-box in *Arabidopsis* is identified in another study). Because phylogenetic footprinting does not require *a priori* information on any known regulatory elements, the identification of universal promoters proved that this method is effective for discovering

expression-related DNA elements. In contrast, none of the D-copies included in this analysis showed any universal promoters within the few hundred base pairs. These D-copies do have conserved elements but without any homology to known plant regulatory elements. The reason of these ambiguous results of 5' upstream sequence analysis may be the lack of enough sequence information in most of taxa sampled. Alternatively, some of the 5' untranslated sequences come from cDNAs, which are retro-transcribed products from messenger RNA and thus do not contain a lot of expression control elements. Even with this limited analysis, discrepancies of *RPB2* 5'-regulatory elements between I-copies and D-copies have already been observed since all I-copies' TATA-box were detected by the analysis. The different compositions of conserved small elements upstream D- and I-copies may be the underlying mechanism of differentiated expression patterns between these two paralogs (Luo *et al.* 2007), and the reason why certain copies had been preserved in different plants.

Multi-locus phylogenetic analyses based on concatenation and coalescence

With the advancement of molecular technology to characterize and sequence specific genomic regions, molecular markers (mainly genes but also other non-genic regions) have been used to reconstruct phylogeny at almost every taxonomic level (Chase *et al.* 1993; A. P. G. 2003). Usually, single-copy genes are used because phylogenetic relationship among orthologous copies accurately reflect speciation events in evolutionary history. However, gene trees may not always reflect species trees (Nichols 2001), so multiple genes are sampled to gather 'total evidence' and accommodate genome heterogeneity of evolutionary dynamics. Traditionally, 'total evidence' was achieved by

simple concatenation of multiple loci as if they were parts of a continuous stretch of DNA molecule. This approach provides an excellent approximation to the actual population genetic changes if speciation events are very old. As a result, concatenation serves the purpose of phylogenetics and systematics at higher taxonomic levels, at which lineage sorting is complete, ancestral polymorphisms are usually absent, and genes in different taxa form different monophyletic groups. However, at lower taxonomic levels, especially at species boundaries, incomplete lineage sorting is very common, as a result of speciation with genetic exchanges and genome heterogeneity in terms of gene flow and divergence (Via 2009). In these cases, concatenation of multiple loci oversimplifies the complex genetic and phylogenetic processes underlying speciation and may yield less reliable phylogenetic inference. With the abilities to sequence more taxa and more markers simultaneously, new coalescence-based multi-locus phylogenetic analysis approaches have been developed to take advantages of population level sequence data (Edwards 2009; Liu *et al.* 2009). The innovative aspect of such approaches lies in the modeling of coalescence, the way phylogenetic signal from multiple markers is summarized. The major difference between these new approaches and simple concatenation is that coalescence-based methods model gene evolution independently, in which congruence and conflicts of phylogenetic signal from different markers can be assessed. In this study, a coalescence-based phylogenetic analysis method BEST was applied to our species-level phylogenetic analyses of North American native grapevines. In stead of analyzing four nuclear markers as a linear stretch of DNA molecule, BEST estimate gene trees of four markers independently and reconcile their phylogenetic signal to estimate a species tree. The simplest, and the most important improvement over

concatenation is that different DNA evolution models are considered for each locus, which can not be achieved in concatenation-based analyses, as shown in this study (four loci, three different models). The other difference is how the conflicts of phylogenetic signal are handled. In the concatenation-based method, 'weaker' signals in certain loci were covered by 'stronger' signals from other loci, which both ignores the 'weak' signals and lowers supports for 'strong' signals. However, coalescence-based methods like BEST force species tree to be compatible with gene trees estimated from each loci. As a result, species tree method will lead to more polytomies, as shown in this study. Although polytomies are usually not desired by systematists, lack of phylogenetic resolution is better than potential misleading phylogenetic hypotheses suggested by concatenated data. The comparison of tree topologies of North American grapevines estimated from two methods showed consistency between two methods as major well-established groups were revealed by both approaches, suggesting that strong phylogenetic signals are independent to analytical methods. However, coalescence-based approach showed slight improvements because of its lack of support to poorly-understood phylogenetic relationships in native grapevines in West North America, as a result of the tendency of species tree methods to generate more polytomies. In contrast, concatenated-based method strongly supported the relationships contradict to prior knowledge, with the potential to reinforce incorrect conclusions.

Chapter 3

Ecological niche partitioning and divergence of flowering phenology of North American native grapevines (*Vitis spp.*) revealed by herbarium specimens

Abstract

The geographic distribution and flowering phenology of plants are increasingly studied by large-scale analyses of herbarium specimens. Grapevines (*Vitis*, Vitaceae), are widespread woody vines with significant biodiversity in the Southeastern U. S. Detailed information on the geographic distribution, habitat preference and phenology of native species is needed to better understand their evolutionary history. Here, ranges, ecological niches and flowering time of several native grapevine species were investigated based on herbarium records. Ideas that flowering time divergence is a major component of reproductive isolation were investigated in the context of phylogeny and ecological niches. Geographic distributions of native grapevine species were estimated from locality data extracted from herbarium specimens. Flowering phenology was analyzed in association with ecological niches specified by environmental variables. Analyses of herbarium records generated significant amount of data on grapevine ranges, niches and flowering times. Estimated species ranges from georeferenced specimens were largely consistent with results from smaller or more localized efforts but offered unprecedented detail and scope. Significant ecological niche differentiations were revealed in several geographic groups. Unique flowering phenology was observed in association with niche partitioning, suggesting possible adaptive evolution in divergence of certain grapevine

species. Analysis of herbarium records provided large-scale and realistic estimates of ranges for North American native grapevines, which facilitate further niche and habitat studies. Species-specific adaptations are reflected by association between flowering phenology divergence and ecological niche partitioning. Flowering time difference is a potential reproductive isolation mechanism or a by-product of adaptations to unique habitats in grapevines.

Introduction

Phenology is the study of seasonal life history events in plants and animals and their interactions with climate and environments. Because phenological phenomena are sensitive and strongly associated with temporal and spatial climate variations, phenological records were used as a proxy to investigate historical climate changes (Fitter and Fitter 2002). Phenological records can also be utilized to study evolution of plants and animals as they reflect organisms' phenotypical changes in response to environments and adaptations to specific ecological niches. Here, a study of flowering time, an important component of plant phenology, is conducted in a specific plant group, North American native grapevines. In this study, phenological records were extracted from plant specimen collections of herbarium databases, and used to investigate contemporary flowering time variation and its association with ecological niches.

Plant specimen collection has been a standard scientific practice for over three hundred years, and an unparalleled source of information on plant distribution and diversity. Unfortunately, before the widespread use of database management and the rise of Internet, valuable information was scattered around the world in thousands of herbaria with very limited sharing and accessibility. Only recently have herbaria started to put information associated with their collections into databases and make them publicly available. This transition has allowed many opportunities to re-examine and analyze the collections on an unprecedented scale and synthetic level. For example, Bradley *et al.* (1999) examined climate change as reflected by phenological changes in a small flora in

southern Wisconsin, United States. With advanced Geographic Information System (GIS) technology, Leimbeck *et al.* (2004) analyzed herbarium specimens with landscape and climate variables to investigate endemism of Araceae in Ecuador. Plant specimens can provide information such as reproductive status and geographic localities in addition to their taxonomic values. With climate and environmental information associated with localities, the features of specific habitats (ecological niches) can be specified for plant species. The ecological divergence and niche partitioning can be assessed across species, as well as compared to flowering phenology divergence. Hence, herbarium collections are ideal to address ecological and phenological problems in large scale, e.g. Lavoie and Lachance 2006; Miller-Rushing *et al.* 2006; Gallaher and Leishman 2009.

To demonstrate the power of large-scale plant specimen data in evolutionary studies, I focused on a specific group: North American native grapevines (*Vitis* spp). There are approximately 50 distinct taxa in the genus with significant clusters of diversity in Asia and the Southeastern U.S (Moore 1991). To better understand the natural history of North American grapevines, much work is needed on their geographic distributions, habitats and phenology, especially the temporal and spatial aspects of flowering phenotypes. As an important horticultural crop, grapevine and its wild relatives have received consistent attention from collectors in North America. Tens of thousands of grapevine specimens are preserved in herbaria around North America, and many of them are being digitized as parts of biodiversity informatics projects. This provides us with the opportunity to integrate information collected by several generations of botanists working in numerous places. Flowering time, a key reproductive trait, is at center stage of phenological studies

of flowering plants. Flowering time difference is a common prezygotic reproductive isolation mechanism in flowering plants (Antonovics *et al.* 2006; Silvertown *et al.* 2006). In grapevines, flowering time divergence has long been hypothesized as primary candidate for reproductive isolation among sympatric grapevine species (Comeaux 1984). Because grapevine flowering periods are assumed to be relatively short (about two weeks), a slight temporal shift in flowering time can effectively prevent gene flow, hybridization and maintain species boundaries.

In this study, herbarium data methods developed for biodiversity informatics and ecological niche modeling are used to investigate the geographic distribution, flowering phenology and phenology-habitat associations of North American grapevines. Large scale, yet detailed analysis of *Vitis* ranges, ecological niches and phenology were conducted with the aid of advanced statistical and GIS tools. Our specific objectives are: use specimen data to test the hypotheses about ranges of North American *Vitis* spp based on the more limited and local field work; test the hypothesis that flowering time is strongly correlated with habitats in *Vitis* spp; test the hypothesis that significant flowering time differences are present in closely related *Vitis* spp. Results from this study will facilitate further ecological and genetic investigations of grapevine evolution, speciation and the flowering process. The innovative biodiversity informatics approaches and analytical processes can also be extended to evolutionary studies of other organisms.

Materials and Methods

Data collection

Specimen data of native North American grapevines were obtained primarily from online plant databases found at individual herbarium websites. A list, originally compiled by Alan Prather (Michigan State University, pers. comm.), is maintained and updated at http://www.cals.ncsu.edu/plantbiology/ncsc/type_links.htm (Holmgren PK and Holmgren NH), is used. Collections from both U. S. and non-U. S. institution are searched to exhaust the specimens collected within U. S. Additional online herbarium databases were identified, focusing on areas and states in the country where grapevines naturally occur in abundance including California, Oregon, the Southwestern U. S. and all of the Eastern U. S. Some online databases represented consortia, or data clearinghouses for several herbaria (*e. g.* Wisconsin Botanical Information System and the Consortium of California Herbaria, for a complete list of about 70 herbaria see Appendices Table 2). Most records were directly downloaded through search queries at herbarium websites or databases. Numerous data were obtained through e-mail requests directly to the herbaria. In several cases, a standard loan of plant material was requested and the label data were extracted.

The total 8137 *Vitis* records obtained were standardized with the fields shown in Appendices Table 3. These records were then sorted to include only specimens collected in the contiguous 48 states of the U.S., and excluded poorly documented records (without standard label information such collection date and location) that were deemed uninformative. After standardizing and sorting, the database contained 7891 complete

records of 15 *Vitis* species.

Georeferencing process

Specimens with GPS coordinates were directly incorporated into the data set. Specimen records with only locality descriptions were imported into GeoLocate program (Bart and Rios 2007) to infer the coordinates. GeoLocate uses a large gazette (including cities, towns, roads, water bodies and so on) as reference system to link the locality description to the actual geographic marks and estimate the coordinates. All samples were mapped to a preliminary map to identify the 'geographic outliers', which were specimens in the ocean or outside U. S. All of the final georeferenced data were entered, stored and manipulated in the GIS software ArcGIS (ESRI 2009) for estimations of species ranges.

Range estimations

Computational approaches were used to predict the ranges based on locality data. Because of plants' limited mobility and strong association with habitat, specimen density is an important factor in range estimation and should be taken into consideration. Here, the neighbor-based adaptive LoCoH method was used to estimate the range of each species (Getz and Wilmers 2004). LoCoH creates a series of polygons around 10% of the points, 20%, 30% and so on until all samples are covered. In this way, not only the extent of ranges is estimated, but also the core areas for each species, where the density is the highest, are identified. By selecting polygons enclosing 90% or 100% of all samples, relatively realistic ranges with density information can be extracted and analyzed. LoCoH was installed as an ArcGIS tool and the analysis was done within the ArcGIS software.

Identification of flowering specimens

By inspecting a sample of available grapevine specimens, it is shown that specimens in different reproductive stages can be clearly distinguished. Although individual grapevine flowers are inconspicuous, the clusters of flowers can be identified, indicating that the rate of correct identification of flowering specimens is very high. As a result, 4 out of 15 species were excluded from the following analysis due to small sample size (< 20 specimens) of flowering specimens. The final data set contains 1324 specimens in flower (with geographic coordinates, grouped on a species basis) from 11 species for the analyses of ecological niches and flowering times. Collection dates were converted into Julian day.

Climate data and ecological niche analysis

The climate data provided by DIVA-GIS were used to generate environmental variables at the specimen localities (Hijmans *et al.* 2001). The data layers were generated by interpolating climate data from weather stations on a 30 arc-second grid resolution. In the interpolation process, values of environmental variables for localities without weather stations were predicted by a spatial algorithm based on data from nearby stations and altitude, latitude and longitude (Hijmans *et al.* 2005). Environmental variables included three basic ones: annual average of minimum temperature, maximum temperature and precipitation, averaged from 1955 to 2000, as well as 19 bioclimatic variables. The bioclimatic variables are derived from the basic climate variables. They are indicators of annual trends, seasonality and limiting environmental factors, commonly used in ecological modeling process (Hijmans *et al.* 2005, details see Appendices Table 4). The

averaging process across years reduced the noise in the data by partially eliminating effects of abnormal climate conditions in certain individual years. For the 11 species as a whole, one-way ANOVA was conducted for the three basic climate variables. Additional one-way contrasts were also conducted among major species groups. A Principle Component Analysis (PCA) was conducted for the 19 bioclimatic variables to find the independent principle components of climate conditions. The first four principle components (totally explained over 99% variations in data) were subjected to a discriminate function analysis (DFA) to identify the major factors, LDs (the Linear Discriminators, linear combinations of the principle components) that discriminate each species' specific ecological niche. LD1, the first linear discriminator, was used as the primary indicator of ecological niche for each species. All statistical analyses were conducted using the statistical package R (R Development Core Team 2009).

Flowering time analysis

Box plots of flowering times were made to show variation across species. Outliers were removed after inspections of the unusual records. Most of the outliers were flowering specimens with a date record in November and December (not likely a natural phenomenon). Sample sizes, means, medians, earliest and latest flowering dates were calculated on a species-by-species basis. One-way ANOVA analysis was conducted for flowering time using taxa as the factor. Flowering time was used as the response variable to regress against environmental variables, and the best model was selected using Akaike Information Criterion (AIC). Specifically, collection year was included in the model selection process to identify the relative contribution of yearly fluctuation to flowering

time variation. To identify flowering time differences without effects of environmental variables, ANCOVA (Analysis of Covariates) was conducted for all specimens and major groups using taxa as factor and temperature and rainfall as covariates. Then contrasts and pairwise comparisons of flowering times were conducted based on the adjusted means.

Results

Range estimations

Estimated ranges revealed several sympatric and allopatric species pairs. In the Western group, *V. arizonica*, *V. californica* and *V. girdiana* are exclusively allopatric, occupying northern California, southern California and Arizona, respectively (Fig. 8). In Eastern North America, species density was considerably higher. The three closely related species, *V. labrusca*, *V. mustangensis* and *V. shuttleworthii*, are exclusively allopatric, found in the northeastern U. S., Texas and Florida, respectively. *V. aestivalis*, the sister species to the previous three, has a much broader range from the Northeastern U. S. to the Gulf Coast, and is sympatric with all three (Fig. 1). At the same time, the range of *V. aestivalis* largely overlaps with another two widely distributed but unrelated species: *V. cinerea* and *V. vulpina* (Fig. 8). Both range widely from the Northeastern U. S. to the Gulf Coast. The phylogenetic relationships among them as well as with *V. aestivalis* are not clearly resolved or not well-supported, so it was difficult to infer the evolutionary histories of these species (Fig. 1 in Chapter 1). Comeaux (1984) suggested that *V. vulpina* and *V. palmata* may be sister species and they are sympatric. Preliminary phylogenetic analyses (see Fig. 1 in Chapter 1) also supported the sister species status of this pair. In

this study, the estimated ranges for these two species are indeed overlapped, with *V. palmata* completely distributed within *V. vulpina*'s range (Fig. 8c). Finally, *V. rotundifolia* and *V. riparia* are predominately distributed in Southern and Northern North America, respectively, and are almost allopatric with a narrow contact strip around 37 degrees north latitude. Overall, the estimated ranges are largely consistent with the ranges estimated by Comeaux (1984) based on data from field work. Native North American *Vitis* spp are generally divided into two major geographic groups, Eastern species and Western species, by the Rocky Mountains. Combining the estimated ranges and our preliminary phylogeny (Fig. 1 in Chapter 1), the analyses of flowering periods were based on a geographic and phylogenetic grouping: the Southeastern group, consisting of, *V. aestivalis*, *V. labrusca* and *V. mustangensis* (Fig. 8a); the Western group with *V. californica*, *V. arizonica* and *V. girdiana* (Fig. 8b); the Southern group, containing only one species, *V. rotundifolia* (Fig. 8d); and the Northern group, also containing only one species, *V. riparia* (Fig. 8e). The remainder (*V. cinerea*, *V. monticola* and *V. vulpina*) was denoted as the 'others' (Fig. 8f).

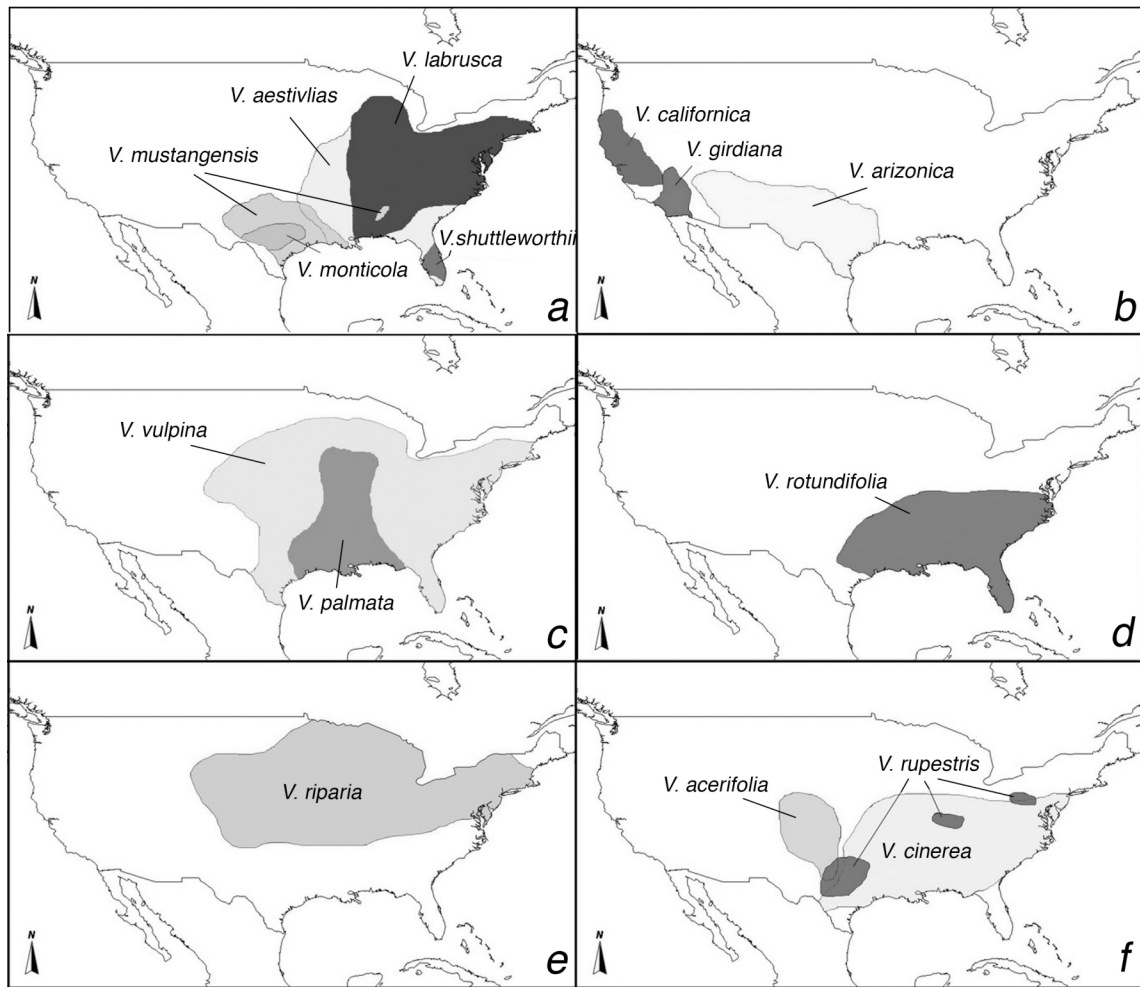


Fig. 8 Estimated ranges of *Vitis* spp, species are grouped according to geography: a. Southeastern group, including *Vitis aestivalis*, *Vitis labrusca*, *Vitis monticola*, *Vitis mustangensis* and *Vitis shuttleworthii*; b. Western group, including *Vitis arizonica*, *Vitis girdiana* and *Vitis californica*; c. *Vitis palmata* and *Vitis vulpina*; d. Southern species *Vitis rotundifolia*; e. Northern species *Vitis riparia* and f. other species, including *Vitis acerifolia*, *Vitis cinerea* and *Vitis rupestris*

Ecological niche partitioning

One-way ANOVA analysis of environmental variables across all species suggested a

strong difference of climate conditions among taxa ($P < 2.2\text{e-}16$ for all three major variables). Contrasts among major groups also showed significantly different climate conditions (Table 3). This suggested a possible strong niche partitioning between the Southeastern group and other species. PCA results showed that within both Southeastern and Western group, the ecological niches were partitioned among species (Fig. 9a and b), indicating fine-scale niche partitioning within geographic groups. Within the Southeastern group, *V. mustangensis*'s LD1, the major discriminator of ecological niches, showed significant shift compared to the other two species, suggesting ecological niche partitioning between this species and others (Fig. 11a).

Table 3 Contrasts of environmental variables among groups

Contrasts	Maximum temperature	Minimum temperature	Rainfall
Southern vs All others	<0.001**	<1e-09**	<1e-05**
Northern vs All others	<0.001**	<1e-09**	<1e-05**
Southeastern vs All others	0.0193*	<1e-09**	<1e-05**
Western vs All others	<0.001**	5.03e-09**	<1e-05**
Southeastern vs Western	0.4440	<1e-09**	<1e-05**
Southern vs Southeastern	<0.001**	3.43e-09**	<1e-05**
Southern vs Western	<0.001**	<1e-09**	<1e-05**
Northern vs Southern	<0.001**	<1e-09**	<1e-05**
Northern vs Southeastern	<0.001**	<1e-09**	<1e-05**
Northern vs Western	<0.001**	<1e-09**	<1e-05**

* Significant at $\alpha = 0.05$ level; ** Significant at $\alpha = 0.01$ level.

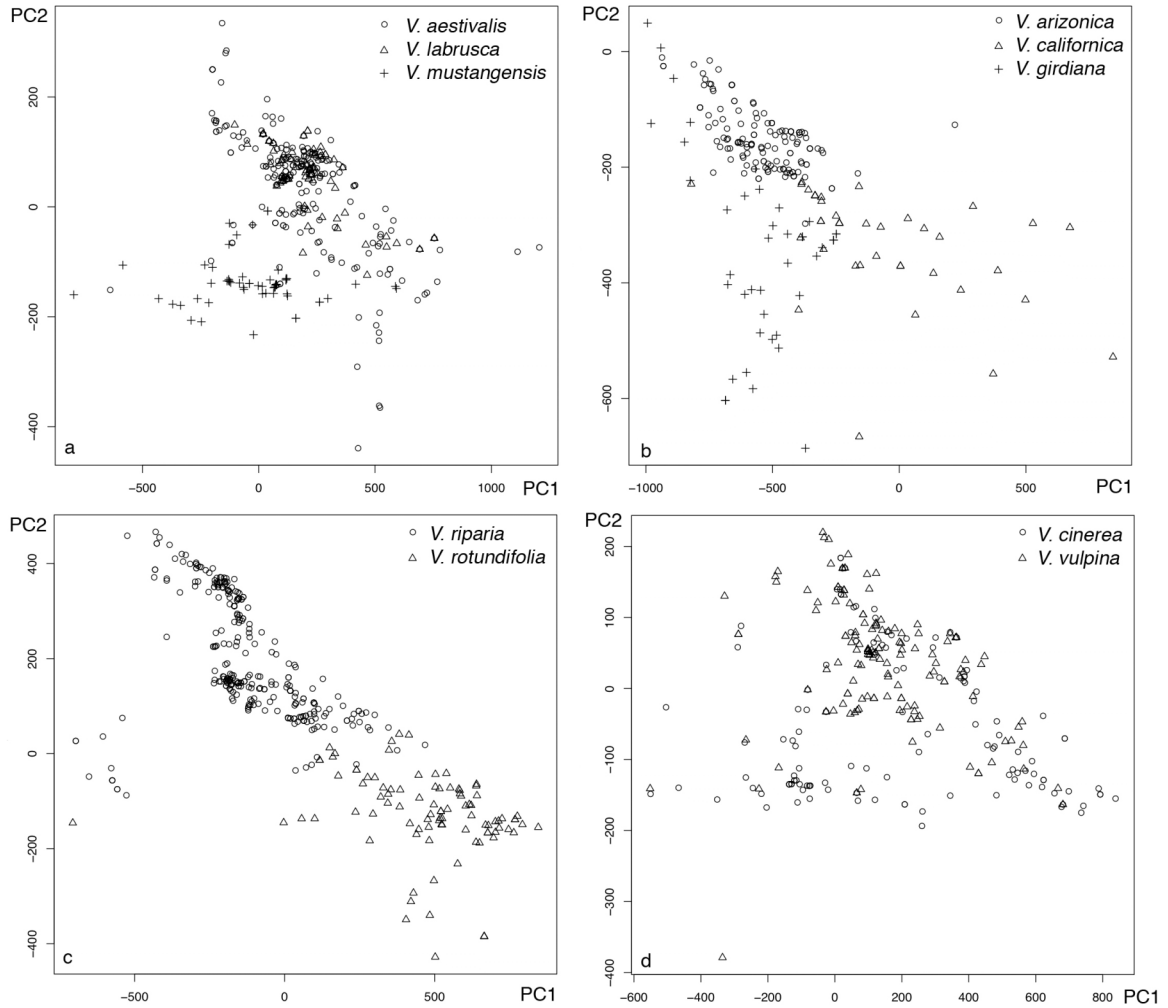


Fig. 9 Results of PCA (Plot of the first and second principle components, PC1 and PC2, explained over 95% variance in all groups) of bioclimatic variables within a. Southeastern group; b. Western group; c. Northern and Southern group and d. other species, different species are indicated by points of different shapes (legends shown in the up-right corner)

Flowering time difference

Without any correction for bias due to climate conditions, the data showed that most of the species flowered from late May to early June on average (Table 4, the 140th to 160th

day of the year) and with moderate variations. *V. girdiana* and *V. mustangensis* were exceptions: they flower significantly later (late June) and earlier (early April), respectively. *V. mustangensis* also had the smallest variation. The box plots of flowering periods for 11 species are shown in Fig. 10. Taxa were grouped according to geography. One-way ANOVA analysis showed that flowering times were different across species (F-value 26.7, P-value < 2.2e-26). According to analyses of the genus-wide data set, a linear model including all variables and their interactions best described flowering time. Specifically, model without ‘year of collection’ variable can explain flowering time variation equally well compared to model with this variable (P = 0.34), suggesting yearly fluctuation contributes small fraction of total flowering time variations. Biologically, significant interaction terms suggested that direct comparisons of flowering time among species in the overall data set is impossible since different taxa do not share parallel linear relationships between flowering time and environmental variables. However, for the Southeastern group, the best-fit model did not include interactions between taxa and other variables, which made overall pairwise comparisons of flowering time between species possible. The environment-adjusted mean flowering time of every species were compared for each pair, which provided flowering time difference corrected for climate conditions. Pairwise comparisons of flowering time involving *V. mustangensis* within Southeastern group showed significant difference (Table 5). Clearly, the early flowering time of *V. mustangensis* is strongly associated with its specific ecological niche (Fig. 11).

Table 4 Flowering time descriptive statistics.

taxa	group	sample size	earliest	latest	mean	median
<i>V. aestivalis</i>	Southeastern	250	103	208	156	159
<i>V. labrusca</i>	Southeastern	82	116	190	156	153
<i>V. mustangensis</i>	Southeastern	64	65	125	96	96
<i>V. arizonica</i>	Western	145	83	192	140	140
<i>V. californica</i>	Western	40	120	212	156	158
<i>V. girdiana</i>	Western	44	99	256	161	150
<i>V. rotundifolia</i>	Southern	93	91	188	139	140
<i>V. riparia</i>	Northern	313	114	191	157	157
<i>V. cinerea</i>	Others	125	105	179	143	145
<i>V. monticola</i>	Others	20	92	171	120	114
<i>V. vulpina</i>	Others	148	105	199	149	150

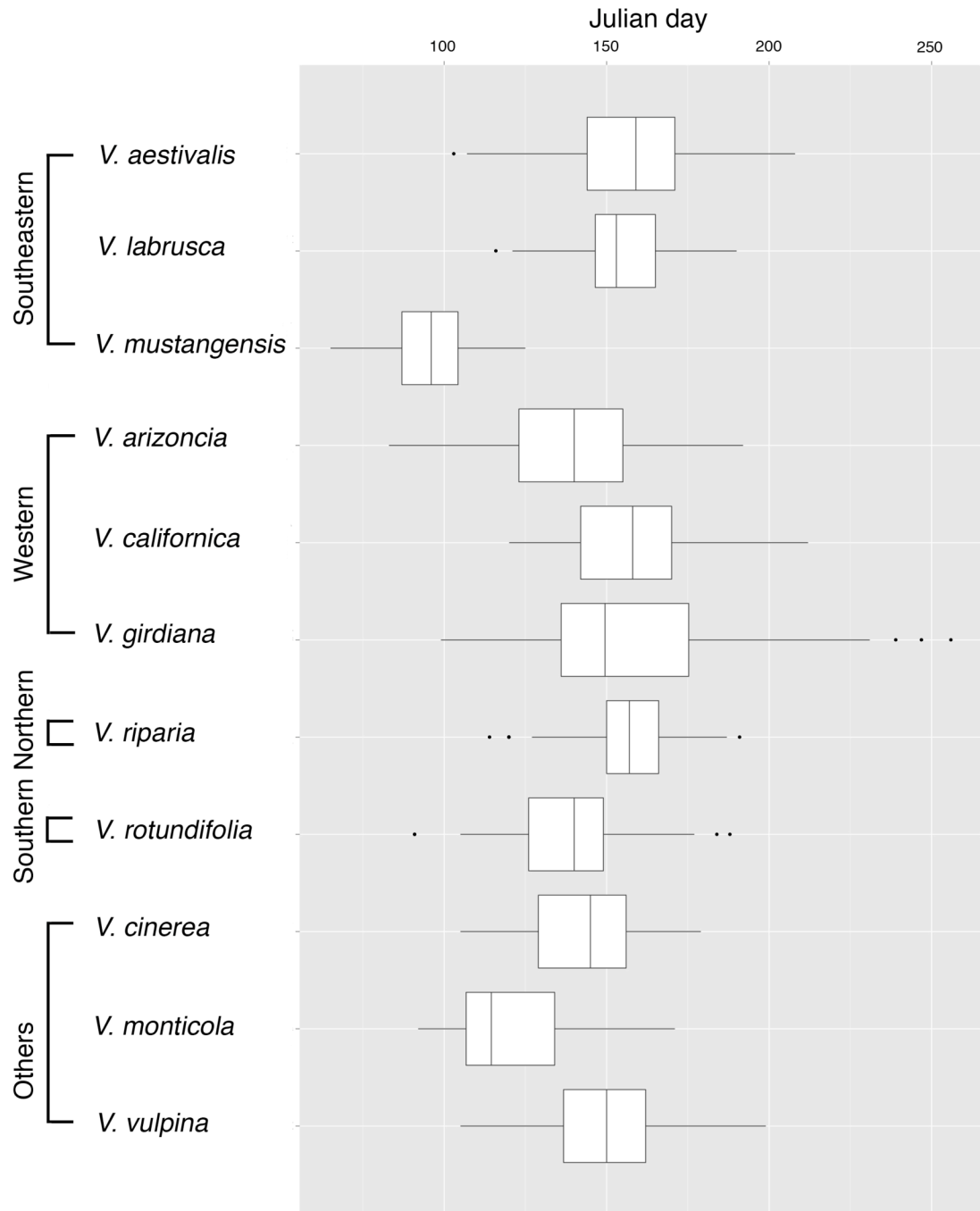


Fig. 10 Box plots of flowering time (Julian day) distribution for *Vitis* spp, species grouped by geography, vertical lines indicate median flowering time

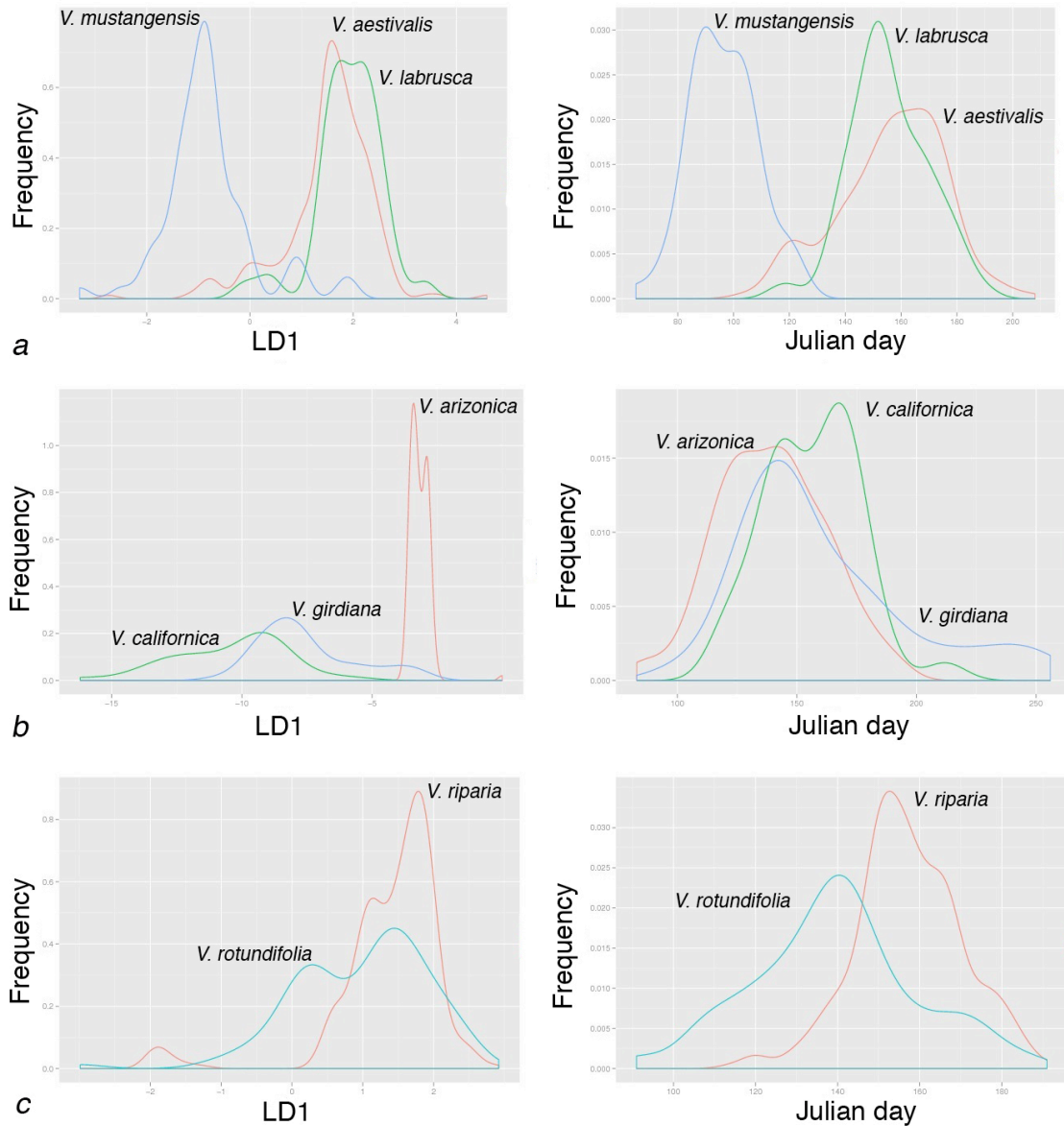


Fig. 11 Comparison of ecological niches (frequency plots of LD1) and flowering time (frequency plots in Julian day) in geographic groups, a. Southeastern; b. Western; c. Northern/Southern. LD1, the first linear discriminator, a linear combination of principle components (also environmental variables) specifying characteristics of ecological niches

Table 5 Pairwise comparisons of flowering time and climate within Southeastern group.

Comparisons	Flowering time (ANCOVA)	Maximum temperatur e	Minimum temperature	Rainfall
<i>V. labrusca</i> vs <i>V. aestivalis</i>	0.0375	0.1348	0.5144	0.4582
<i>V. mustangensis</i> vs <i>V. aestivalis</i>	<0.001**	<0.001**	<0.001**	<0.001**
<i>V. mustangensis</i> vs <i>V. labrusca</i>	<0.001**	<0.001**	<0.001**	<0.001**

* Significant at $\alpha = 0.05$ level; ** Significant at $\alpha = 0.01$ level.

Discussion

Range estimation

Large quantities of herbarium specimens with associated geographic data can be used to estimate ranges and other aspects of species distribution. The accuracy of georeferencing depends on the quality and sources of geographic information. Geographic coordinates predicted by GeoLocate are estimates so the quality of georeferencing done in this way is not as reliable as direct GPS points. However, in this continent-scale study, the inaccuracies generated by GeoLocate did not affect the final results and conclusions because the largest error margins do not exceed county boundaries. Therefore, specimens with GeoLocate generated coordinates were included and not distinguished from other source of coordinates. The analysis of geographic distributions of plant specimens was usually associated with assessment of biodiversity and endemism (Lienert *et al.* 2002; Gimaret-Carpentier *et al.* 2003). Estimation of ranges based on georeferenced specimens also allows more complicated techniques than, for example, drawing simple polygons around all samples. In this study, a progressive estimation method was employed to account for the specimen density, which ensured that realistic estimates are made.

In this study, estimated ranges based on *Vitis* specimens are highly consistent with the ranges estimated by Comeaux (1984) based on regional data from field work. Native North American *Vitis* spp are generally divided into two major geographic groups, Eastern species and Western species, by the Rocky Mountains. Interestingly, the phylogeny inferred from nuclear genes (Fig.1 in Chapter 1) suggested a closer relationship between *V. californica* and Asian species. Because of this, if evolution of this species is to be considered, Asian species should also be included in the taxon sampling for a broadened phylogenetic analyses. However, the relationship to Asian species is restricted into this single Western grapevine species. *Vitis*, as a genus, have a pan-Beringial distribution also observed in other organisms. But unlike the organisms showing distribution of closely related species in Asia and Southeast North America, Native *Vitis* in Southeastern United States and China both seem to represent two distantly related and independently evolving lineages resulting from ancient vicariance.

Environmental variables and ecological niches

The data set in this study included not only phenological and geographic information, but also sources of climate information obtained based on geographic coordinates. Elevation, temperature and rainfall data were extracted from climate models (Hijmans *et al.* 2001) based on several decades of observations from thousands of weather stations, which provided us the means to analyze the relationship between flowering time phenology and ecological niches specified by the environmental variables (Crimmins *et al.* 2008). Temporal variables of specimens, for example, year of collecting, were not included in our analysis because of three reasons: 1. Climate variables already

accommodated temporal variations in climate conditions and reduced random error caused by large deviation of climate in individual abnormal years; 2. The response variable, flowering time, closely resembled a normal distribution, indicating that our samples are fairly good representatives of plants in nature and yearly fluctuations had no significant effects; 3. Yearly flowering time variations explain a much smaller proportion of overall variations than environmental variables (see results), suggesting that yearly fluctuation contributes neglectable variations of flowering time than environmental conditions.

Gimaret-Carpentier *et al.* 2003; Leimbeck *et al.* 2004 and Scheldeman *et al.* 2007 employed climate variables to analyze landscape, biodiversity and endemism. Recent literature has greatly developed the use of climate data in phenological analyses e.g. Primack *et al.* 2004; Bustamante and Búrquez 2008. Lavoie and Lachance (2006) created a new method to reconstruct the phenological patterns across a large area. Specifically, they correlated phenology with a unique reference climate variable (date of snow disappearance) to facilitate direct comparisons of historical flowering time. In this study, the data analysis focused on contemporary species-level phenology and ecological niches. ANOVA revealed genus-wide variations in climate conditions, while more sophisticated PCA and DFA of the bioclimatic variables identified the factors that best describe the specific habitats for each species. In the Southeastern group of grapevine species, fine scale niche partitioning was demonstrated by the shift of major indicator of specific habitats. Ecological niche modeling using integrated information can be used not only for predicting ranges for species of interest, but also to reveal ‘cryptic allopatric’

species that are separated by factors other than simple geography (Kearney and Porter 2009).

Flowering time differences

In this study, flowering time information were extracted from herbarium data, a non-traditional source of phenological records. Even though the data were temporarily and spatially diverse, significant biological signal, flowering time differences, were detected among species. In some species, the sample size was limited to 50 ~ 100 specimens, which may compromise the statistical power and mask biological signals. However, flowering times were shown to be normally distributed both in overall data and individual species, suggesting that with current size, our samples are fairly good representatives of the natural populations. To detect flowering time differences, data sets with smaller sample size may even provide more statistical power than larger ones. For variables like flowering time, significant P-values (<0.05 or 0.01) can always be achieved by increasing sample sizes. A small difference will be statistically significant with large enough sample size, but without any biological meaning.

Flowering time differences could be found mainly related to three groups: Southern, Northern and Southeastern. Significant early and late flowering in *V. rotundifolia* (139th day) and *V. riparia* (157th day), were discovered in contrasts of flowering time ($P < 0.01$, Fig. 4). Based on our estimates, the only species in the Southern group, *V. rotundifolia* distributes mainly Southern and Southeastern U. S., with a northern limit in Virginia. In contrast (Fig. 8d), the estimated range of *V. riparia* covers the northern U. S., extending

into Canada (Fig. 8e). The Southern and Northern biased distributions may contribute to the flowering time difference in these two species as a result of adaptations to climate conditions (Fig. 11c). In the preliminary phylogeny (Fig. 1 in Chapter 1), *V. rotundifolia* was set as an outgroup because *V. rotundifolia* has a different chromosome number than the other species, which is a strong reproductive barrier. So even though this species' flowering time is different from others, flowering time alone cannot be the major reproductive isolation mechanism between this species and others.

The Southeastern group was a well-supported monophyletic group in the preliminary phylogeny (Fig. 1 in Chapter 1). Within the group, *V. mustangensis* has very early flowering times and all pairwise comparisons involving it were significant even after accommodating climate conditions (Table 5). The significant shift of flowering time is coincident with this species' specific ecological niche, as shown in Fig. 11a. Hence, the unusual flowering time may be a result of the species' unique habitat. This is consistent with the earlier observations that flowering time divergence can be a by-product of adaptations to differentiated niches. More importantly, the shift of flowering time may be the first sign of prezygotic reproductive isolation between populations adapted to different niches (Antonovics 2006). Devaux and Lande (2008) conducted a simulation study to assess the possibility of allochronic speciation due to flowering time shifts. They attributed incipient allochronic speciation to limited population size, which is somewhat consistent with the situation of *V. mustangensis*. This species has a very limited distribution with a widely distributed closely related species *V. aestivalis*. It is possible that the divergence of *V. mustangensis* from its closely related congeners involved a

dramatic reduction of population size as a result of adaptive selection or change in population structure, and early flowering evolved as the major reproductive isolation.

Implications for research on plant phenology and evolution

Traditional field-based methods in biodiversity and biogeographical studies are time and effort consuming hence usually limited to small taxonomic groups and geographical regions. Biodiversity informatics approaches, such as online database mining and herbarium surveys, have been utilized in recent efforts to collect, index and share highly valuable biodiversity information and to study, for example: phenological responses to climate change (Walther *et al.* 2002; Wolfe *et al.* 2005), invasive plant spread (Delisle *et al.* 2003; Godoy *et al.* 2009), biodiversity (Singh and Kushwaha 2006; Scheldeman *et al.* 2007) and conservation strategies (MacDougall *et al.* 1998; Lienert *et al.* 2002). In this study, herbarium data resources are used in innovative ways: phenological data were collected from plant specimen collections, an unconventional source for such records; phenological data were used to identify contemporary flowering time variations and analyze interactions between flowering time and climate in a group of closely related species. This approach proved to be very useful in exploratory studies to identify interesting phenological traits, which may be followed by finely-controlled experiments.

The large-scale data collection provided enough data to conduct sophisticated statistical analysis of flowering times across a vast geographic area. The collation of herbarium data also provided enough samples for estimating relatively realistic ranges for the taxa under investigation. Such combined analysis of geographic distribution, phenology and

ecological niches in plants has rarely been performed on this scale. Furthermore, analyses were conducted with well-supported biogeographic information and preliminary phylogenetic resolution. The synthetic nature of this study, with the aid of a large data set, provided a fairly comprehensive starting point to explore future research questions. By combining herbarium specimens, biogeography and phylogenetics, important questions may be addressed in other organisms in terms of natural history and species formation. Clearly, while the *in silico* approach is capable of generating statistically relevant results, field studies are still highly desirable and needed to verify situations like micro-habitats partitioning or the presence of hybrids. A more complete picture of plants and their environments can be obtained by combining the macro-scale (such as herbarium database mining) and micro-scale (field work) research, laying a solid foundation for further study on evolution of phenology and species formation.

Chapter 4

Characterization and analyses of grapevine genes controlling flowering time

Abstract

Flowering time difference is among the first-appearing prezygotic isolation mechanisms after divergence of plant populations. In native North American grapevines, flowering time shifts were found in closely related species, usually associated with adaptations to specific habitats. In *Arabidopsis*, as well as other plants, the regulatory networks and key signal integrators of flowering time control were characterized. A candidate gene approach is adopted here to characterize orthologous copies of flowering time genes in several native grapevine species. Five critical flowering time control genes, including *CONSTANS* (*CO*), *FCA*, *FRIGIDA* (*FRI*), *FLOWERING LOCUS T* (*FT*) and *TERMINAL FLOWER 1* (*TFL1A*), are characterized in native grapevines. In *TFL1A* genes, a microsatellite region with species-specific length is identified in 5'-flanking regions in front of the start codon, with potential to influence gene expression patterns in different species. Alternative transcription of grapevine *FCA* is revealed by analysis of gene expression, suggesting a conserved *FCA* auto-regulation process between grapevine and *Arabidopsis*. Analysis of site-specific and lineage-specific selection was conducted on *TFL1A* and *FRI*. No signature of natural selection at molecular level was revealed in these loci, suggesting that evolution of these candidate genes may not be strongly associated with phenological evolution in native grapevines. In addition, it is suggested that flowering time, as a trait with large plasticity, may be determined by large number of

small effect loci, and no dramatic genetic change underlies the observed flowering time difference in certain grapevine species. Flowering time divergence may be a by-product of other phenotypic and genomic changes associated with adaptations to specific ecological niches.

Introduction

Several previous studies found that flowering time divergence is among the first isolation mechanisms appearing when plant populations begin to diverge (Antonovics 2006, Silvertown *et al.* 2006). However, the molecular basis of flowering time, and molecular evolution accompanying flowering time divergence, is completely unknown in these cases. Generally, the genetic basis and evolution of reproductive isolation is the least studied area in plant speciation. In Wu and Ting's review (2004) on speciation genes that determine hybrid incompatibilities, all were identified in animals. A more recent review (Presgraves 2010) included few genes directly affecting reproductive isolation, serving as speciation genes, in plant systems (Bomblies *et al.* 2007). The lack of genes causing intrinsic incompatibilities is consistent with Grant's (1981) theory that plant species are largely compatible intrinsically, but separated by external factors like ecology and phenology. Therefore, speciation genes also include gene responsible for ecological niche preference and local adaptations, which may not have direct connections with reproduction. Flowering time difference is the major component of prezygotic isolation in plants and may evolved as a by-product of adaptations to specific environments (Antonovics 2006, Silvertown *et al.* 2006). As a result, genes controlling and affecting flowering time are candidates of speciation genes. In *Arabidopsis*, large-scale QTL mapping and mutagenesis identified several key genes affecting flowering time variations. In the aid of fully sequenced genome, the molecular basis of the key genes as well as the pathways underlying flowering time control are largely deciphered (Simpson and Dean 2002).

Flowering time in *Arabidopsis* is controlled by four pathways (autonomous, Gibberellins, vernalization and light-dependent). The signals from these pathways are integrated into expression control of several genes, a process sometimes called the integration pathway. For every pathway, key genes are recovered by screening flowering time mutants (Koornneef *et al.* 1991). For example, florigen, the mysterious flowering signal (floral hormone) generated in leaves and transported into flower meristems, was found to be protein encoded by *FT*, and transported through phloem (Jaeger and Wigge 2007). *FT* lies in a central position of flowering time pathways as an integrator of all signal inputs from four pathways. Interestingly, the same gene, *FT*, played the same role in flowering in other plants such as Poplar. Similarly, FT-like proteins were revealed affecting flowering time in both rice and wheat, suggesting conservation of *FT* function across plant taxa (Kojima *et al.* 2002 and Yan *et al.* 2006). The protein encoded by *TFL1* is also a mobile signal controlling meristem identity throughout plant life cycle (Conti and Bradley 2007). Both genes belong to the same gene family, as all proteins of the family contains a phosphatidylethanolamine-binding domain. Interestingly, the two genes showed antagonistic functions. *FT* usually acts as flowering promoter while *TFL1A* acts as flowering repressor (Kobayashi *et al.* 1999). Hanzawa *et al.* (2005) suggested that as few as a single amino acid change can reverse functions of *FT* and *TFL1A*. *CO* gene integrates light-dependent signals and controls the photoperiod pathway, and directly affects *FT* expression (Putterill *et al.* 1995). In vernalization pathway, *FLC* is the signal integrator, which is not only influenced by upstream vernalization genes such as *FRI* but also autonomous 'pathway' gene such as *FCA* (He 2009). In natural *Arabidopsis* populations, Le Corre revealed that the frequency of null alleles, disrupted by indels in

exon 1 of *FRI*, was related to flowering time. The higher frequency of null alleles in a population, the later average flowering time this population has. Several geographical isolated natural populations already diverged in terms of null allele frequencies and as a consequence, they showed significantly different flowering times (Le Corre 2005). In addition, *FCA* is a regulator of flowering promoter *FLC*, working in a self-regulatory way by alternative transcription in *Arabidopsis* (Macknight *et al.* 1997).

The nature of mutations causing natural flowering time variations varies from alternation of protein to changed expression levels (Alonso-Blanco *et al.* 2005). Because of the prevalence of TEs in plant genomes (Bennetzen 2005), plant phenotypes are affected by TE insertions into functional important genes, and flowering time genes are not exceptions. TE insertion in *Arabidopsis FLC* first intron caused misexpression of this gene and altered flowering time (Michaels *et al.* 2003). In addition, *FRI* and *FLC* have epistatic effect on flowering time and form a latitudinal cline of flowering time variation in natural *Arabidopsis* populations (Caicedo *et al.* 2004). Furthermore, substitutions in coding regions consisted more than half of the cases in which flowering times are changed by changes in flowering genes (Alonso-Blanco *et al.* 2005). All of these showed that the changes in flowering time could be traced down to molecular level. Along with the accumulation of plant genomic data (Tuskan *et al.* 2006, Jaillon *et al.* 2007), the search-for-molecular-basis process will be shifting from time consuming phenotype-genotype QTL experiments to association analyses between molecular features and phenotypes (Templeton 1994). As the second part of an association study to identify the molecular basis of flowering time variation in native North American grapevines,

grapevine orthologous genes related to flowering time, including *CO*, *FCA*, *FT*, *FRI* and *TFL1A*, are identified and analyzed for their molecular evolution in a phylogenetic framework (reconstructed by an additional gene *RPB2* and other nuclear markers (see Chapter 2)). Hypothesis that molecular evolution of flowering time genes is correlated to flowering time divergence is tested based on comparative and associative studies of *Vitis* flowering time genes.

Materials and methods

Molecular cloning and sequence analysis

All DNAs of *Vitis* species, including 14 species: *V. acerifolia*, *V. aestivalis*, *V. arizonica*, *V. californica*, *V. cinerea*, *V. girdiana*, *V. labrusca*, *V. mustangensis*, *V. monticola*, *V. palmata*, *V. riparia*, *V. shuttleworthii*, *V. vinifera* and *V. vulpina*, were extracted from fresh leaf tissue using a modified CTAB DNA extraction protocol (Doyle and Doyle 1987). Genomic regions of *V. vinifera* *CO*, *FCA*, *FT*, *FRI* and *TFL1A* were identified by tBLASTn search against the grapevine genome using *Arabidopsis* orthologous proteins as queries. *V. vinifera* Expression Sequence Tag (EST) sequences were identified by BLASTn search against the grapevine EST database using genomic sequence as queries (Altschul *et al.* 1997, Jaillon *et al.* 2007). PCR primers were designed in both 5' and 3' UTR regions as well as first and last exons according to the genomic sequences (Table 6). Additional primers were prepared for each primary primer in case sequence divergence among wild species prevents degenerative primary primers working (Appendices Table 5). PCR were conducted using New England Biolab Taq

polymerase (including standard 10X PCR buffer and dNTPs) and standard amplification program (denaturation at 72°C for 3 minutes, annealing of primers at 58°C for 0.5 minutes, elongation at 72°C for 1.5~2 minutes, 35 cycles of amplification). The *V. vinifera* Pinot Meunier (used in *FCA* transcript characterization) was kindly provided by Peter S. Cousins of USDA Grape Genetics Research. Total RNAs were extracted from fresh leaves. cDNA synthesis was conducted in two steps: 1. denaturation of total RNAs at 65°C for 5 minutes with *FCA* X13R primer; 2. cDNA synthesis with DTT, RNAase inhibitor and AMV Reverse-Transcriptase at 50°C for 30 minutes followed by 85c for 5 minutes. The *FCA* cDNA was used as templates for following PCR with primer *FCA* X1F and *FCA* X13R (denaturation at 72°C for 3 minutes, annealing at 54°C for 0.5 minutes, elongation at 72°C for 1.5 minutes, 35 cycles). All PCR products were cloned using Topo-TA cloning kit and harvested by PCR using plasmids as templates. The final products were cleaned and submitted for sequencing by Auburn University Genome Sequencing Laboratory. The *FCA* cDNA sequence was submitted to GenBank (accession number GU300763).

Table 6 Primers for characterization of flowering time genes in *Vitis* spp.

Locus		Primers
<i>CO</i>	5UF	5'-ATGACATGCACATTAAATGATTCCA-3'
	3UR	5'-GGAATGAGTAATGAGAAGCTGAGTTTCAA-3'
<i>FCA</i>	X1F	5'-CTACGGCAACAACCCTGACT-3'
	X13R	5'-GGATCAGCAAACCGAACAGT-3'
<i>FT</i>	5UF	5'-GCAGATAGCACCGGACTAGTTAT-3'
	X1F	5'-GAGAAGTAGCAAATGGCTGTGA-3'
	X4R	5'-GGATTATGCTTCACCCCATAGA-3'
<i>FRI</i>	X1F	5'-CTGCCAAACTTGTACTTGAATGC-3'
	IIR	5'-GTTTAGCATCCCGGAAGGA-3'
<i>TFL1A</i>	5UF	5'-GCCTTCCAAGAGACCAAGAGT-3'
	3UR	5'-TGATCTTCCCGTTGGTTATTG-3'

Molecular evolutionary analyses

Tests of positive selection were conducted by estimating codon-specific dN/dS ratio using MrBayes with gene-specific DNA evolution model coupled with MG94 codon model, 2 MCMC runs with 100,000 generations, sampling every 10 generations, burnin set at 250th generation (Huelsenbeck and Ronquist 2001, Ronquist and Huelsenbeck 2003). Lineage-specific tests of positive selection were conducted using HyPhy using the same codon model and phylogenetic tree reconstructed in Chapter 2. *A priori* hypothesis that *V. mustangensis* has different dN/dS ratio was tested in a maximum likelihood framework (Pond *et al.* 2005).

Results

Characterization of flowering time genes (*CO*, *FCA*, *FT*, *FRI* and *TFL1A*)

Based on sequence similarities to *Arabidopsis* flowering genes and *V. vinifera* genomic sequences, several grapevine flowering genes were characterized, including *FCA*, (first exon) and *FLC* in autonomous pathway; *CO* in light dependent pathway; and *TFL*/*FT* family in integration pathway (Fig. 12). *TFL1A* and *FT*, which belong to the same gene family, were characterized in several native grapevine species. In *TFL1A*, a microsatellite region adjacent to the start codon was discovered in every sampled grapevine species. Microsatellite loci are usually fast evolving with large variations in length (repeat number) even between generations, but the lengths of the repeats near *TFL1A* contain less variations within species than across species (Fig. 13).

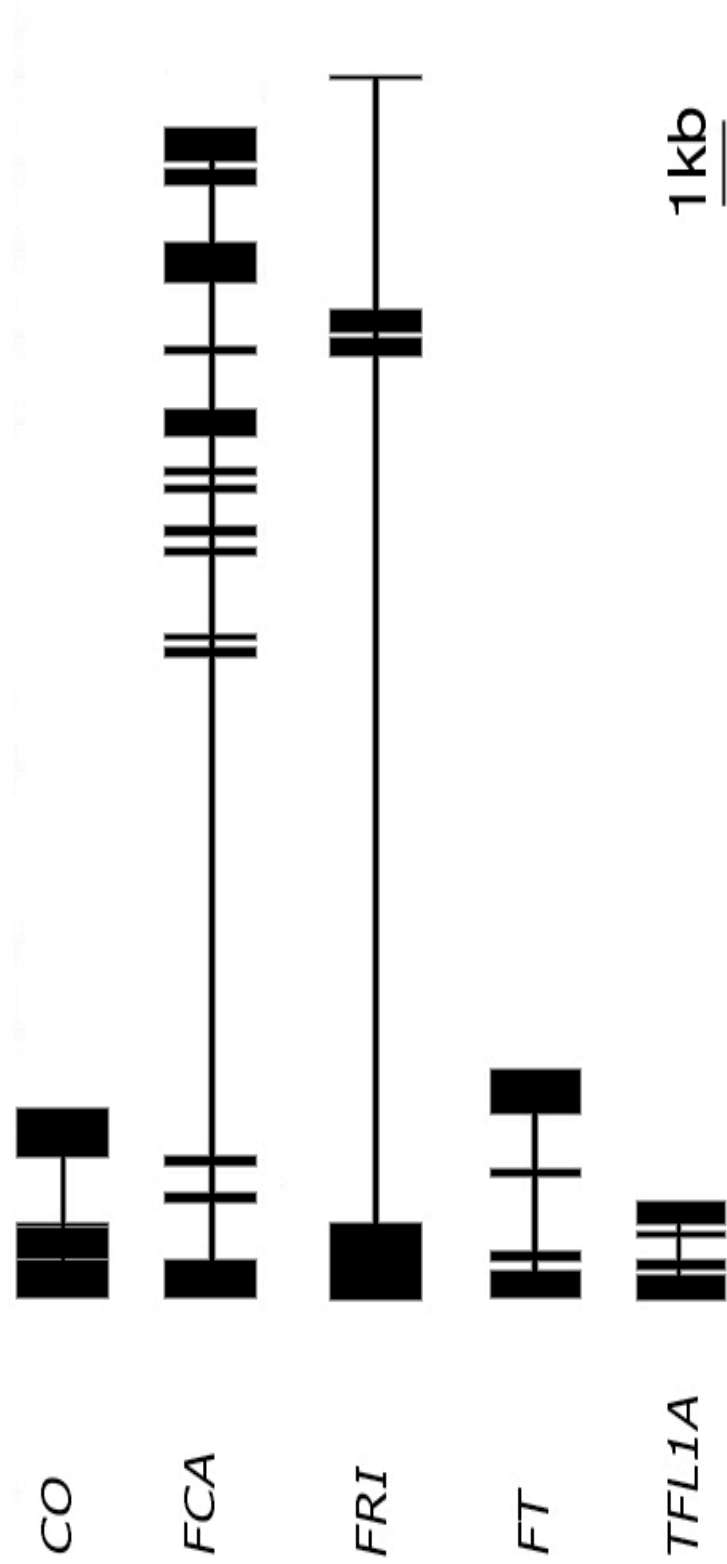


Fig. 12 Gene models of characterized flowering time genes. Boxes represent exons and lines represent introns

Alternative transcription of *FCA* in grapevine

12 EST sequences of *FCA* were identified in *V. vinifera* by BLAST search. When mapped to genomic regions, 10 ESTs are composed of the first three exons, with various portions of first exon and third intron. Most of these ESTs have poly-A tails within intron 3, suggesting that they come from *FCA* transcript β . Two remaining ESTs (at 3' end) contain the last five exons and only the last exon, respectively. They apparently come from *FCA* transcript α or γ if alternative transcription is conserved between *Vitis* and *Arabidopsis*. A new transcript was recovered in cultivar Pinot Meunier from a cDNA library. This transcript spans from exon 1 to exon 9, with canonical splice sites (GT/AG) observable at every intron/exon junction. Intron 3 was correctly spliced out, indicating that this transcript does not belong to the α or β form. If alternative transcription of *FCA* is highly conserved between *Vitis* and *Arabidopsis*, this transcript can be γ or δ . The relative abundance of the two forms in *Arabidopsis* young seedlings are ~35% and ~10%, respectively. Therefore, the newly identified transcript in *Vitis* leaves is very likely transcript δ . An improved annotation of *FCA* was suggested based on ESTs and the new transcript. The 3'-neighboring gene of *FCA* of current annotation was included in newly annotated *FCA*. This was supported by the presence of two 3' ESTs, and protein similarity between grapevine and *Arabidopsis* at this end. In addition, canonical intron splicing sites are present after connecting the last exon of *FCA* and the first exon of downstream gene when introducing a new intron (Fig. 14).

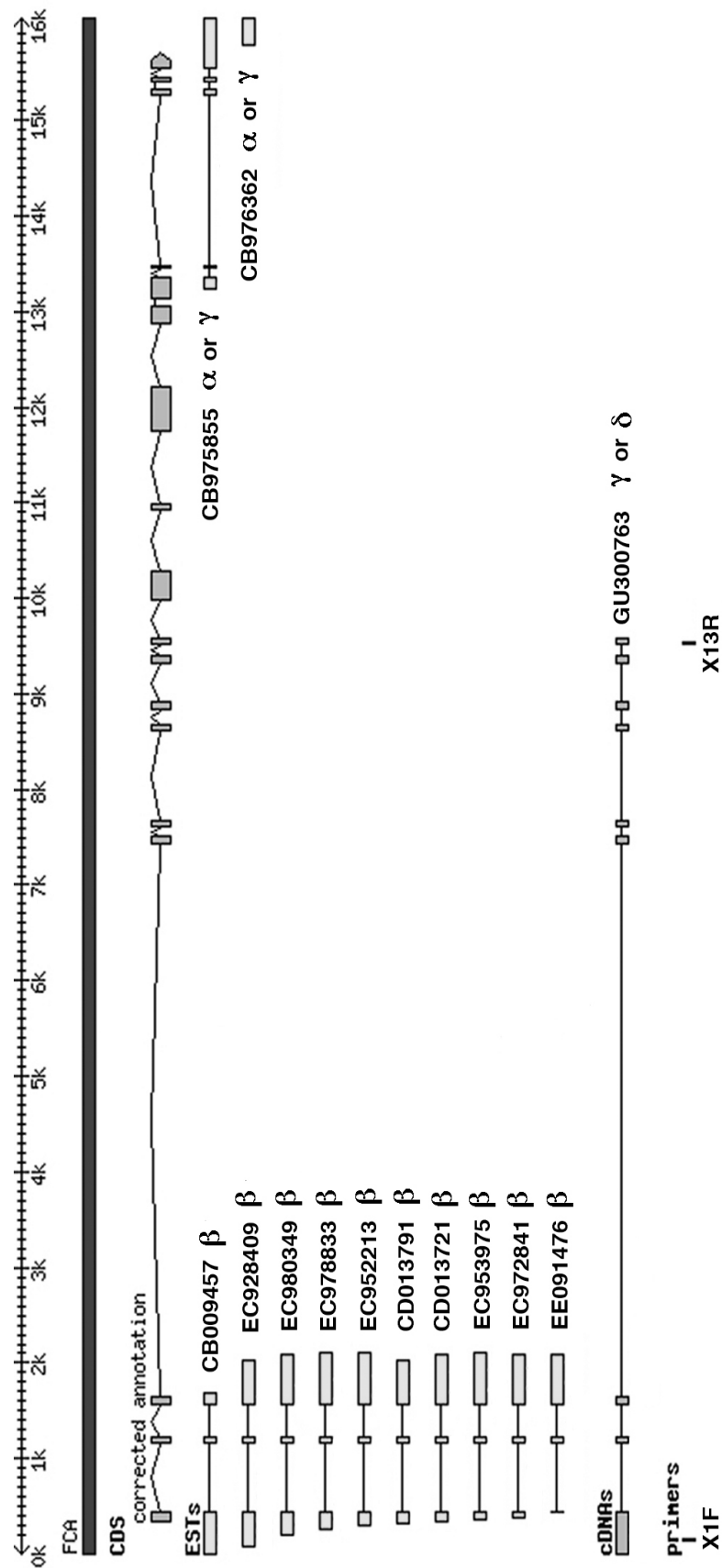


Fig. 14 Diagram of annotation of *Vitis vinifera* FCA gene. Dashed bar: genomic coordinates (kbp); dark bar: genomic sequences; shade boxes: exons; lines: introns; filled boxes: primers. ESTs and cDNAs annotated with GenBank accession numbers and putative transcript form (β , γ or δ)

Analysis of natural selection on flowering time genes

Site and lineage-specific analysis of positive selection on *FRI* exon 1 and *TFL1A* revealed neither temporal (certain parts of tree) or spatial (within coding regions) evidence of positive selection. The average dN/dS ratio in *FRI* exon 1 stays around 0.4 in the entire region while in *TFL1A*, the estimates stay at 0.25, suggesting strong purifying selection on the coding regions of these two genes (Fig. 15). Specifically, early flowering species *V. mustangensis* does not show accelerated or slowed amino acid substitutions in *FRI* exon 1 and *TFL1A* coding regions.

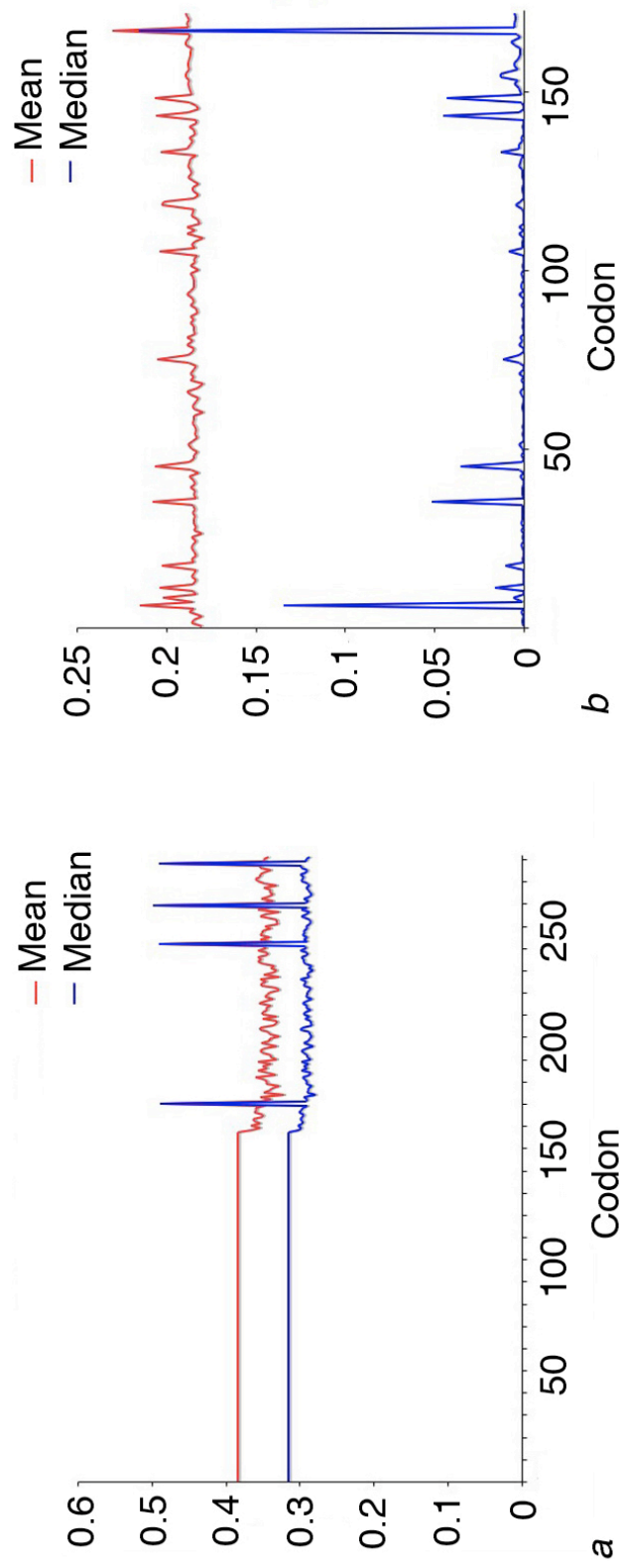


Fig. 15 Site-specific dN/dS ratio estimates based on codon evolution models, a. *FRI* exon 1; b. *TFL1A* complete coding regions, blue and red lines (connected data points) indicate estimated median and mean values of dN/dS ratio for individual amino acid site

Discussion

FT/TFL1A family in grapevines

The protein encoded by plant *FT* gene has been recently identified as the long-sought mobile signal, florigen, which moves from leaves to floral meristem to promote flowering (Mathieu *et al.* 2007). Members of *FT/TFL1* gene family have been characterized in domesticated grapevine, *V. vinifera*, with both grapevine *FT* and *TFL1* orthologous genes correctly identified. Furthermore, their expression patterns and transgenic effects in *Arabidopsis* were investigated, which indicated conserved functions of *FT/TFL1* between grapevine and *Arabidopsis* (Carmona *et al.* 2006). In this study, both genes were characterized in multiple native grapevine species. Both showed very low sequence diversity within the genus *Vitis*, suggesting very strong purifying selection imposed on them. This is consistent with the observation that both *TFL1* and *FT* have very low nucleotide diversities within *Arabidopsis* (Flowers *et al.* 2009). Molecular evolutionary analyses of both genes at larger phylogenetic scale (all flowering plants) also revealed very low sequence divergence among distantly related flowering plants (Chapter 5). In summary, all evidence point to a very conserved evolutionary history of *FT* and *TFL1* with very slow nucleotide and amino acid substitution rates at all taxonomic levels.

Interestingly, a microsatellite region very close to the start codon of *TFL1* showed variable lengths in different species. The temporal and spatial expression pattern of *TFL1* gene is finely-controlled as a result of its strong pleiotropic effects on plant development. The distance between *cis*-regulatory elements at 5'-flanking region and the start codon

may be an optimum stabilized by natural selection. As a result, variable length of microsatellite close to the 5' end of *TFL1* may have significant effects on its expression. The rapid evolution of 5'-flanking region of *TFL1* have also been observed within *Arabidopsis* (Olsen *et al.* 2002). In contrast to the ultra-conserved coding regions, 5'-flanking sequences of *TFL1* may be fast evolving as a result of molecular adaptations of gene expression in response to environmental fluctuations at low taxonomic level such as species and population level.

FCA alternative transcription

An improved annotation of *V. vinifera* Pinot Noir *FCA* gene was proposed based on EST data and sequence homology. The incorrect annotation split *FCA* exons into separate genes due to ambiguous intron position and splicing site identifications. This phenomenon is fairly common in annotation of *V. vinifera* genome as many genes contain extraordinary long introns (Chapter 6), which presents a major difficulty to automated genome annotation efforts (Wang *et al.* 2003). Here, it is suggested that genomic annotations should combine evidence from genomic sequences, homology and expression data. As more EST and cDNA data are available in both model and non-model systems, large-scale verification of gene annotation by comparing with expression data is both practical and necessary to improve annotation quality (Coyne *et al.* 2008).

The similarity between alternative transcripts of *FCA* in *Vitis* and *Arabidopsis* suggests conservation of the *FCA* autoregulation mechanism in grapevine. This is perhaps not surprising since the functional conservation can also be found between *Arabidopsis* and

rice indicating that *FCA* functional components were present in the common ancestor of monocots and eudicots (Lee *et al.* 2005). The homology of RRM domains can also be traced to a *Plasmodium* protein, indicating an extremely long history of *FCA* and its RNA-recognizing abilities.

Selective constraints on flowering time genes

This study is the second part of an association study of flowering time differences and flowering time gene evolution. Previous studies showed that *V. mustangensis* has a significant early flowering time compared to other species. The molecular evolutionary analyses of flowering time genes focused on finding evidence of molecular adaptations related to early flowering of *V. mustangensis*. Site and lineage (*V. mustangensis*) specific analyses of dN/dS ratio showed no significant shift of selective constraints associated with *V. mustangensis*, suggesting that the genetic basis of early flowering time in *V. mustangensis* may not be associated with the coding regions of flowering time genes characterized and analyzed in this study. It is possible the flowering time divergence in closely related species is determined by more subtle expression pattern changes, which may be affected by genetic changes located outside coding regions. Alternatively, flowering time, as a quantitative trait, may be influenced by large number of small effect QTLs as a result of extensive outcrossing in wild species. Actually, outcrossing was suggested as the major reason that large number of small effect QTLs are responsible for the flowering time variations in maize (Buckler *et al.* 2009).

Chapter 5

Variations in evolutionary rates of Angiosperm flowering time genes

Abstract

As a key trait determining reproductive success in Angiosperms, flowering time is finely controlled by four pathways and several signal integrators that are relatively conserved in all flowering plants. However, variations in evolutionary rates and patterns were observed in genes of plant metabolic pathways, suggesting correlation between weaker pleiotropic effects and faster evolution in downstream genes of these pathways. Here the orthologous genes of key flowering time controllers in Angiosperms are identified and investigated for their molecular evolutionary patterns. The potential orthologous copies of flowering time genes are identified by integrating sequence similarity, gene-specific molecular evolution and angiosperm phylogeny. To accommodate unequal taxon sampling among different loci, mitochondrial gene *rbcL* is used to generate genetic distances and identify comparable phylogenetic depth for each locus. Globally, flowering time controlling genes show different evolutionary rates across comparable genetic distances in angiosperm phylogeny. Higher evolutionary rate of both synonymous and non-synonymous substitutions are revealed in *CONSTANS* (*CO*), while accelerated amino acid substitution is discovered in *FLOWERING LOCUS C* (*FLC*). Analysis of site-specific and lineage-specific natural selection in flowering time genes suggested strong selective constraints on functional domains of most genes, indicating conserved evolution of flowering time pathways in Angiosperms. However, consistent

with the overall high non-synonymous substitution rate, positive selection is revealed in functional domain (MADS-box) of the vernalization signal integrator, *FLC*, indicating potential adaptive evolution of this locus in Angiosperms. In addition, lineage-specific shifts of selective constraints are observed in genes investigated. All genes except *FLOWERING LOCUS T (FT)/TERMINAL FLOWER 1 (TFL1)* family showed elevated level of amino acid substitutions in the branch leading to eudicots or within eudicot group, suggesting an association between changing selective constraints in flowering time genes and the rapid diversification of eudicots.

Introduction

Flower development in Angiosperms represents a major innovation and complex process in seed plants. The correct timing and execution of flowering process is controlled by highly sophisticated gene regulatory networks and fine-tuned coordination of expression of key genes. Major pathways involved in flowering time control include the autonomous pathway, photoperiod pathway, Gibberellins pathway, and in some plants, vernalization pathway (Putterill *et al.* 2004). Each pathway consists of a set of key genes to response the endogenous and environmental cues and integrates the signals. *CO* gene is the signal integrator of photoperiod pathway, in which environmental cues related to day length and photoperiod are responded and integrated as input to downstream flowering process (Putterill *et al.* 1995). *FLC* is the signal integrator of vernalization pathway, in which plants' responses to temperature and vernalization process are integrated and converged to expression of *FLC* gene (He 2009). All pathways converge into the *FT* with different signal inputs via activating or repressing interactions. *FT* acts as a mobile signal and interacts with floral identity genes like *AP1* and *TFL1* to fine-tune the timing of flowering process (Kobayashi *et al.* 1999, Fig. 16). The *FT* protein was recently identified by several groups as the long-sought floral hormone, florigen (Jaeger and Wigge 2007). *FT* and *TFL1* belong to the same gene family and are antagonistic in function and one amino acid change can reverse their effects on flower development (Hanzawa *et al.* 2005).

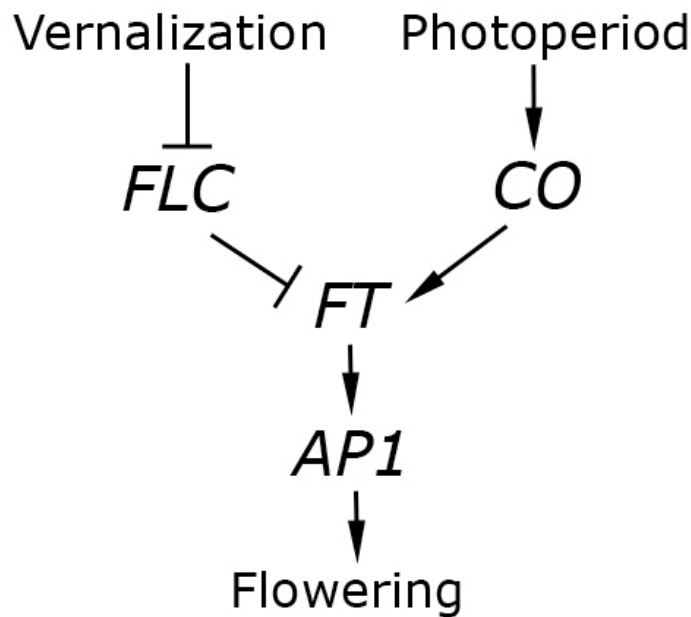


Fig. 16 Simplified diagram of major pathways (vernalization and photoperiod) controlling flowering time; arrows indicate activations, blunted lines indicate repressions; *AP1*, *APETALA1*; *CO*, *COSTANS*; *FLC*, *FLOWERING LOCUS C*; *FT*, *FLOWERING LOCUS T*

Plant genomes are heterogeneous in terms of pattern and rate of evolutionary dynamics (Via 2009). In particular, evolution rates of protein coding genes are largely shaped by functional constraints imposed on them. Based on a few case studies on specific biochemical pathways in both large and small taxonomic groups of plants, evolutionary rates are generally higher in downstream genes due to relaxed functional constraint as a result of lower regulation complexity and weaker pleiotropic effects. For example, Rausher *et al.* (1999) studied rate variations of genes in anthocyanin biosynthetic pathway with orthologous genes from three plant species (maize, snapdragon and

Ipomoea) and revealed more rapid evolution of genes downstream in the pathway. At a lower taxonomic level, Lu and Rausher (2003) analyzed evolution rate variations of genes in anthocyanin pigment pathway within genus *Ipomoea* and confirmed the same trend that upstream genes evolve slower than downstream genes in the pathway. In addition, Rausher *et al.* (2008) analyzed the effects of strengths of functional constraints, as a result of complexity of gene regulation, on the evolutionary rates of anthocyanin pathway genes, and detected positive correlation between reduced connectivity in regulatory network and relaxed selective constraints. Recently, studies of evolutionary rates in other pathways showed general the same results. Ramsay *et al.* (2009) studied correlation between evolution rates and pathway positions of genes in plant terpenoid biosynthesis pathway from five plant species (*Oryza*, *Vitis*, *Arabidopsis*, *Populus*, and *Ricinus*). They found the consistent trend of faster evolution in downstream genes. In addition, they introduced a new measure of pathway position with information on pleiotropic effects as a better indicator of pleiotropic effects and evolution rates. Livingstone and Anderson (2009) characterized genes in carotenoid biosynthetic pathway in a number of higher plants (mainly eudicots) and showed accelerated evolution of downstream genes. At a lower taxonomic level, Yang *et al.* (2009) analyzed the evolution rates of Gibberellin pathway genes within genus *Oryza* and demonstrated the same general trend of faster evolution of downstream genes.

Molecular evolution and selective constraints of genes in flower development pathway have been analyzed within *Arabidopsis* with population genetic approaches (Olsen *et al.* 2002, Flowers *et al.* 2009). However, such studies have not been carried out at a larger

phylogenetic scale. In this study, based on the functions of flowering time genes and possible different selective constraints, the following hypotheses are to be tested: 1) there are substantial variations of evolutionary rates among flowering time genes in Angiosperm evolution; 2) evolutionary rate variations may be observed among functional domains within genes and in specific lineages of flowering plants in association with phenotypic evolution. To investigate variations in evolutionary rates among these genes, *API*, *CO*, *FLC*, *FT* and *TFL1A* orthologs are characterized in Angiosperms ranging from early flowering plants to most major orders of core eudicots. A 'control' gene, *RPB2*, the gene encodes the second largest subunit of RNA polymerase II, was also included as a slow evolving 'control' gene (Chapter 2). Comprehensive molecular evolutionary analyses were conducted on these genes using currently accepted angiosperm phylogeny (A. P. G. 2003) as the phylogenetic framework, with plastid gene *rbcL* divergence as a measure of genetic distances. Analyses of molecular clock, rate variations among genes, as well as selective pressure within genes and in phylogenetic lineages were conducted to reveal the temporal and spatial patterns of evolution rate changes in genes control flower development and flowering time.

Materials and methods

Taxon sampling and data collection

Five genes, including *API*, *CO*, *FLC*, *FT* and *TFL1*, all related to flowering and flowering time control, were analyzed in this study. The gene encodes the second largest subunit of RNA polymerase II, *RPB2*, was also included as a slow evolving 'control' gene

(Chapter 2). Due to the different availabilities of genes in different specific species, our goal was to identify orthologous genes in as many as orders in angiosperm phylogeny with one representative from each order (may be in different families). The sampling scope should be comparable among genes in order to compare evolutionary rates across relatively similar genetic distances.

Protein sequences for these genes from *Arabidopsis*, *Populus* and *Vitis* were identified from GenBank and aligned using ClustalW (Thompson *et al.* 1994) with manual corrections when necessary, to generate a consensus, which was used as a query to conduct a BLAST search against GenBank nucleotide and EST databases (Altschul *et al.* 1997). tBLASTn (using protein sequences as queries to search against translated nucleotide databases) was used to generate the list of significant hits in Angiosperms. The outputs were processed with BLAST output processing program blast2table.pl (<http://www.genome.ou.edu/informatics.html>) and spreadsheet processing program GNU awk to generate hits IDs for batch retrieval from GenBank (accession numbers see Appendices Table 6). Complete Coding Sequences (CDSs) were extracted from retrieved nucleotide sequences by FeatureExtract while coding regions were manually annotated for retrieved Expressed Sequence Tag (EST) sequences (Wernersson 2005). The collected data sets of all similar sequences were aligned using transAlign (EMBOSS, Rice *et al.* 2000) with manual corrections for further analyses.

Orthologous gene identification

Phylogenetic trees were constructed from aligned data sets by PAUP* (version 4.0b11,

Swofford 2002). Well-supported monophyletic groups of orthologous genes were identified. Ideally, for each gene, all orthologous genes should form a monophyletic group and the phylogenetic relationships within the group should reflect angiosperm phylogeny. Based on these criteria, orthologous and paralogous gene copies were sorted, extracted and aligned for following analyses of evolutionary rates using ClustalW.

Analyses of evolution rates

Molecular clock analysis

Tests for molecular clock were conducted for each gene in a maximum likelihood framework. Likelihood ratio tests were conducted between evolution processes (gene-specific DNA evolution models selected by ModelTest (Posada and Crandall 1998)) with or without enforcing a molecular clock along the current accepted angiosperm tree.

Genetic distances

To obtain an independent measure of genetic distances across the sampled angiosperm phylogenetic space, plastid genes *atpB*, *rbcL* and *matK* were collected from the species sampled in this study. As *rbcL* was found in every sampled taxon, only this gene was used to generate genetic distances independent to nuclear genomes. The genetic distances were maximum likelihood estimates by PAUP* based on *rbcL*-specific DNA evolution model selected by ModelTest. Genes share similar distributions of pairwise *rbcL* genetic distances have comparable sampling of taxonomic diversity.

Evolutionary rates

General evolution rates were estimated based on genetic distances (maximum likelihood estimates by PAUP*, gene-specific model selected by ModelTest).

Synonymous (dS) and non-synonymous (dN) rates were estimated by maximum likelihood methods by Nielsen and Yang implemented into PAML (Yang 2007). Average rates for each gene were compared using standard ANOVA to recover significant differences and variations among genes. Linear models were built and compared to identify the factors (genetic distance and gene identity) influencing dS and dN.

Interactions between genetic distances and gene identities in predicting dS and/or dN indicate different evolution patterns of dS and/or dN among genes. All statistical analyses and plots were conducted using the statistical package R (R Development Core Team 2009).

Site-specific analyses of positive selection were conducted for each gene using HyPhy (Pond *et al.* 2005). Site-specific dN and dS values were estimated with gene-specific DNA evolution model selected by ModelTest, MG94 codon model (Muse and Gaut 1994), angiosperm phylogeny and Single Ancestor Counting method, without any lineage-specific dN, dS rate variations. The results were analyzed and plotted with statistical package R.

Lineage-specific analyses of positive selection were also conducted for each gene using HyPhy. Lineage-specific dN/dS ratio were estimated with gene-specific DNA evolution model selected by ModelTest, MG94 codon model and angiosperm phylogeny, without

any site-by-site dN/dS ratio variations. The branch leading to eudicots was *a priori* selected to compare five competing hypotheses using Likelihood Ratio Tests (LRTs). The five hypotheses are: the null hypothesis H0: one uniform dN/dS ratio across the entire tree and alternative hypotheses: H1: different dN/dS ratio between the branch and early Angiosperms and eudicots combined; H2: different dN/dS ratio between eudicots, branch combined and early Angiosperms; H3: different dN/dS ratio between early Angiosperms, branch combined and eudicots; and H4: different dN/dS ratio among early Angiosperms, eudicots and branch. The likelihoods of alternative hypotheses were compared to null hypothesis, as well as between alternative hypotheses. LRTs were conducted to detect significant supports to specific hypothesis using χ^2 tests with Bonferroni corrections for multiple comparisons.

Results

Taxonomic sampling

Taxon sampling of genes are shown in Appendices Table 6 with GenBank accession numbers, including representatives from major groups of early Angiosperms and eudicots. Based on the genetic distance of *rbcL* (a plastid gene used in phylogenetic analyses of Angiosperms, Fig. 17), the phylogenetic depth of taxon sampling is comparable among the genes under investigation. The similar phylogenetic depth surveyed for different genes facilitated the direct comparisons of DNA and protein evolutionary rates in a range of similar genetic distances, without biases of uneven distribution of sequence differences across taxa.

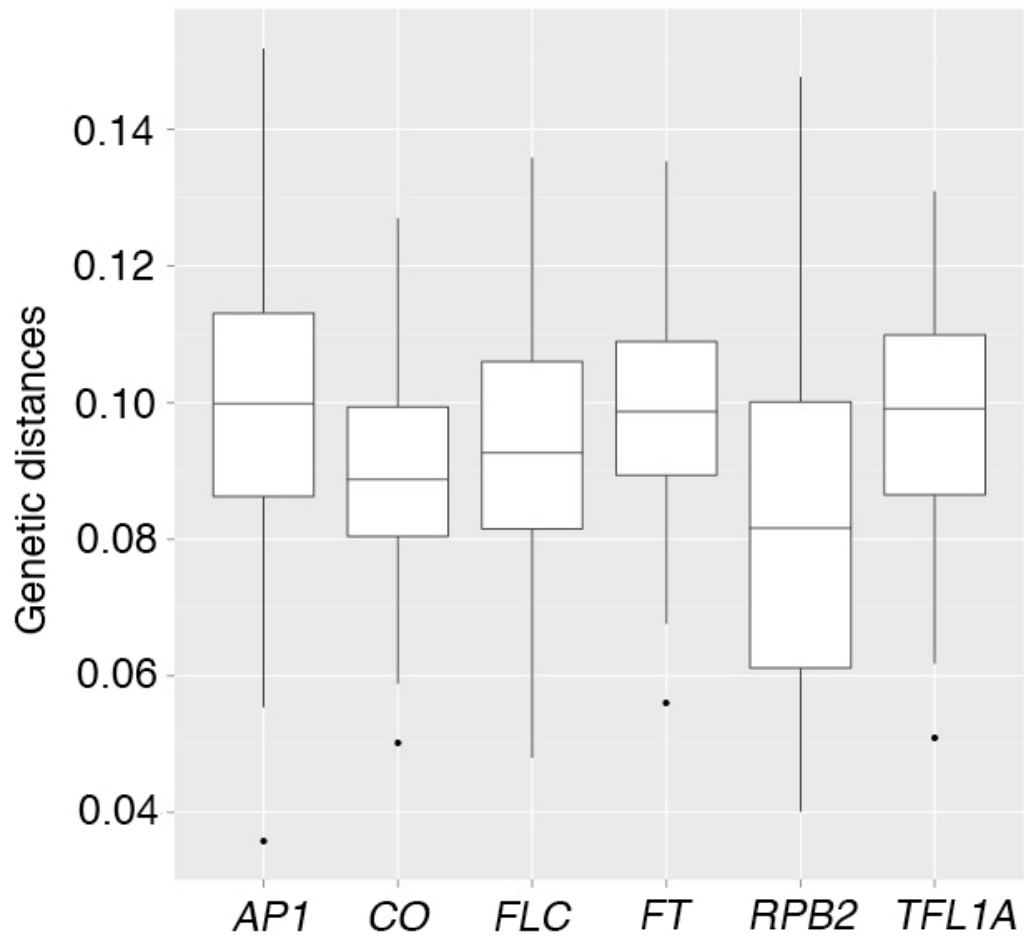


Fig. 17 Distribution of genetic distances estimated from the plastid gene *rbcL*

Ortholog identification

There are several rounds of whole genome duplications in Angiosperm evolutionary history, leading to extensive gene duplications and selective retentions of duplicated copies. Only comparisons of orthologous genes, which are homologs defined by speciation events, can provide accurate estimates of evolutionary rates. The genes under investigation in this study have large number of homologous copies representing both large and small scale genome duplications, which require detailed analyses of their evolutionary histories to correctly identify the true orthologs.

API and *FLC*

Both *API* and *FLC* belong to the large plant transcription factor family, MADS-box genes, so identification of orthologous copies were combined for these two genes. The combined phylogenetic tree constructed by all homologues of *API* and *FLC* resulted in two monophyletic groups consisting of only *API* and *FLC* genes, respectively (Fig. 18a). Interestingly, only eudicot copies are present in the monophyletic groups, indicating a not clear-cut differentiation of *API* and *FLC* in early flowering plants. As a result, following analyses of evolutionary rates only included eudicot copies for *API* and *FLC*.

However, the structural distinctions between *FLC* and *API* beyond the MADS-box domain were recognizable without phylogenetic analyses. As a result, the hypothesis testing of positive selection (lineage and site specific dN/dS rate variations) were conducted based on *API* genes of not only eudicots but also other flowering plants.

CONSTANS family

As an important transcription factor, *CO* belongs to a gene family with multiple homologs. However, all duplication events seem to be older than the diversification of Angiosperms, so *CO* orthologs can be clearly identified as a well-supported clade in the phylogenetic tree of all homologues, and the phylogenetic relationship among *CO* orthologs closely reflect angiosperm phylogeny.

FT and *TFL1A* family

FT and *TFL1A* belong to the same gene family so identification of orthologous copies were combined in one phylogenetic analyses. As grapevine has three paralogs of *TFL1* (*TFL1A*, *TFL1B* and *TFL1C*, Carmona *et al.* 2006), only the true orthologous copy *TFL1A* was used. The phylogenetic tree constructed with all homologues clearly revealed two distinct and well-supported groups, representing *FT* and *TFL1A* lineages, respectively (Fig. 18b). Interestingly, there is not enough phylogenetic signal in the coding regions to resolve relationships within the two groups, suggesting that both genes are very conserved in flowering plants.

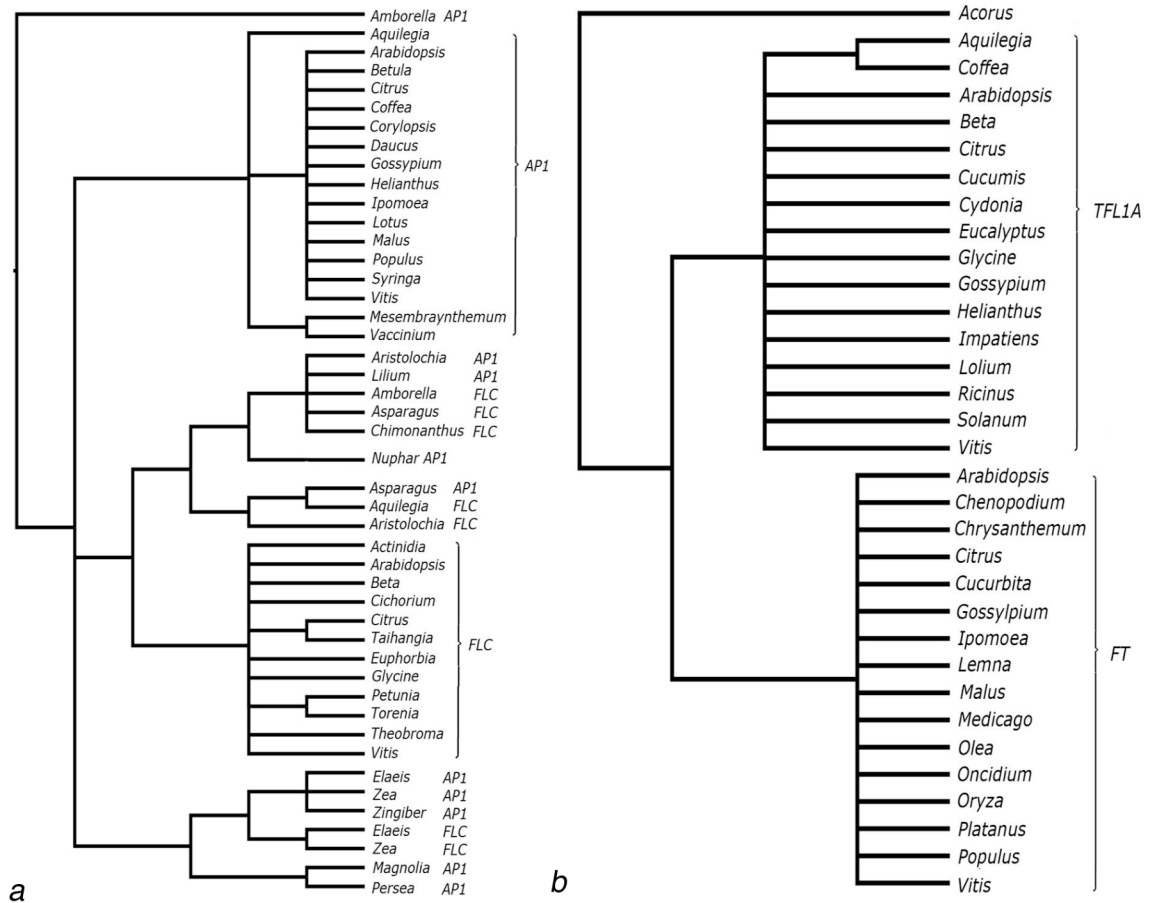


Fig. 18 Gene trees of a. *AP1* (*APETALA1*) and *FLC* (*FLOWERING LOCUS C*); b. *FT* (*FLOWERING LOCUS T*) and *TFL1A* (*TERMINAL FLOWER 1*) reconstructed by identified homologues, monophyletic groups of orthologous genes are indicated by brackets with gene names (only taxa within the monophyletic groups are included in the molecular evolutionary analyses)

Evolutionary rate

Molecular clock

The hypothesis that the genes under investigation are evolving under a molecular clock was tested using the DNA sequences of coding regions in the context of currently

accepted angiosperm phylogeny. All genes show significant rate variations across the entire angiosperm tree, indicating that none of them evolves in a clock fashion (Table 7).

Table 7 Tests of molecular clock

Locus	model	-lnL clock	-lnL no clock	D statistic	df	LRT P-value
<i>API</i>	GTR+I+G	14430.16	14241.61	377.1	25	<0.001
<i>CO</i>	GTR+I+G	17531.31	17269.77	523.08	18	<0.001
<i>FLC</i>	GTR+I+G	11806.7	11429.78	753.84	18	<0.001
<i>FT</i>	GTR+I+G	5547.881	5479.466	136.83	14	<0.001
<i>RPB2</i>	GTR+I+G	21545.19	21231.24	627.9	18	<0.001
<i>TFL1A</i>	GTR+I+G	4567.336	4518.438	97.796	11	<0.001

Variations in evolutionary rates

The evolutionary rates reflected by average genetic distances across the entire phylogeny show significant differences among the gene investigated (ANOVA $P < 0.001$), with higher rates of *CO* and *FLC* genes and lower rate of *RPB2* gene. The separated analyses of synonymous (dS) and non-synonymous substitution rates (dN) revealed differences of rate variations among genes between the two substitution types.

Synonymous rates are comparable among all genes with only *CO* showing slightly elevated substitution rate (Fig. 19a, ANOVA $P < 0.001$ with *CO*, $P = 0.08$ without *CO*).

In contrast, Non-synonymous substitution (amino acid substitution) rates among these genes are significantly different (Fig. 19b, ANOVA $P < 0.001$) with *FLC* showing the highest rate. Model selection processes used to detect slope (dS/distance, dN/distance) differences among the genes suggested no strong interactions between distance and gene to predict dS rate ($P = 0.013$ without *CO* and $P = 0.008$ with *CO*, Fig. 20a) but strong interactions between distance and gene to predict dN ($P < 0.001$, Fig. 20b). In the analysis of dS, the intercept is also not significant. In other words, except *CO*, which has

a distinct synonymous rate, all genes have similar relationships between genetic distances and synonymous substitution rates. In contrast, across the similar genetic distances, amino acid substitution rates are variable among genes.

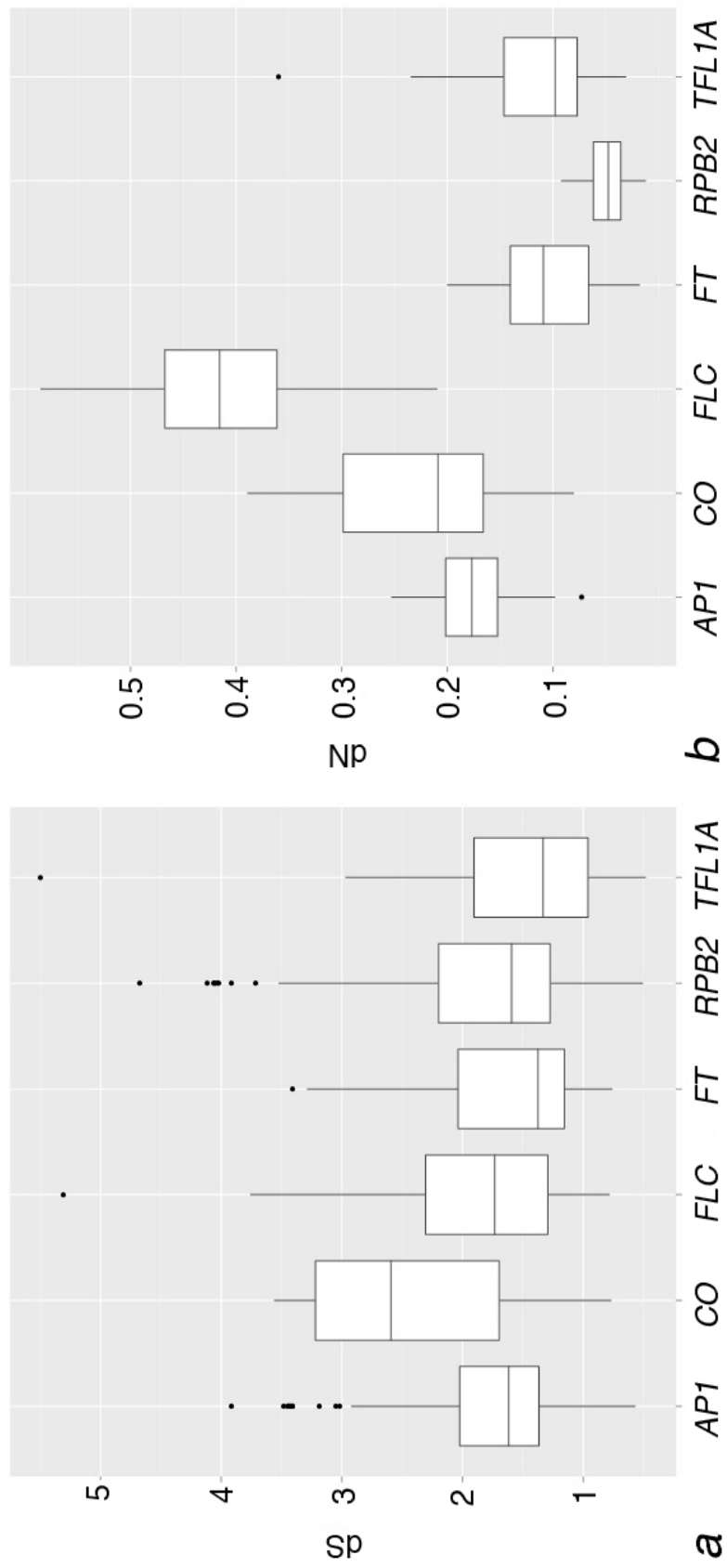


Fig. 19 Distribution of dS and dN across the genes investigated, a. dS rates; b. dN rates

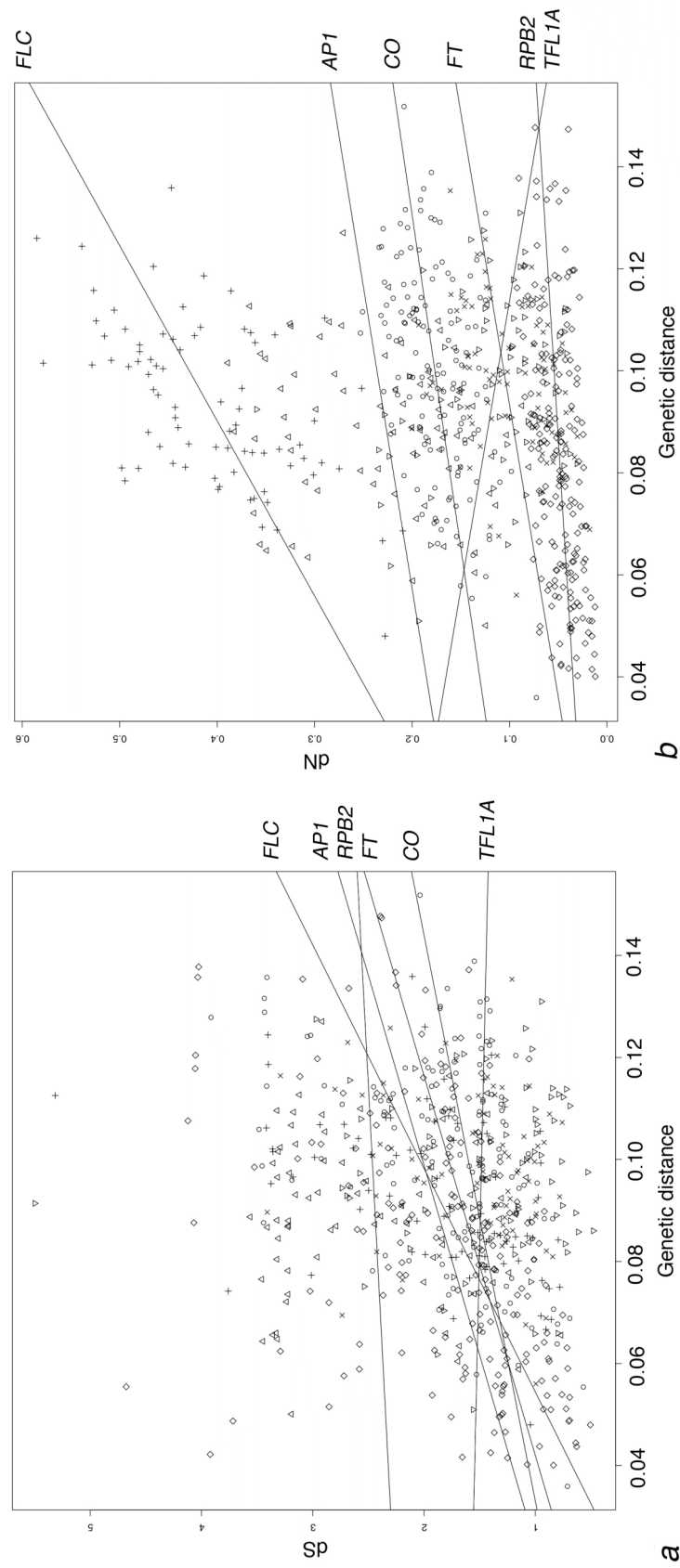


Fig. 20 dS and dN plots against genetic distances (*rbcL*), showing locus-specific regression lines between evolution rate and genetic distance, a. dS; b. dN

Selection analysis

FT and *TFL1A* generally have very low evolution rates and few non-synonymous substitutions, and preliminary analysis suggested that there is no evidence of positive selection on these two genes. As a result, the analyses of site-specific dN/dS rates in these two genes showed no excess of amino acid changes. However, larger variations of non-synonymous substitutions across coding regions were observed in *API*, *CO* and *FLC* genes (Fig. 21). In *CO*, higher substitution rates (both synonymous and non-synonymous) were revealed by site-by-site estimates of substitution rate, consistent with the previous observations of higher average rates of this gene (see previous sections). A substantial proportion of codons in *CO* (15%) have dN/dS ratio larger than one, suggesting relatively common positive selection on this gene. In addition, the codons under positive selection are evenly distributed within coding regions, indicting no specific domains under particular intensive selective pressure. In the MADS-box genes *API* and *FLC*, different types of selective schemes across coding regions were observed. In *API*, numbers of non-synonymous substitutions are comparable with synonymous changes outside the functional SRF and K-box domains, suggesting strong selective constraint on functional domains and relaxed selective pressure on other regions. However, there are a few codons with dN/dS ratio larger than one in the domain with transcription factor activity, suggesting potential positive selection for innovative functions. In contrast, significant excess of non-synonymous changes are observed in the K-box domain in *FLC*, suggesting more amino acid substitutions were fixed along the angiosperm phylogeny within the active domain of *FLC*. By comparing the site-specific estimates and overall average of non-synonymous changes in *FLC*, it is revealed that a significant amount of

amino acid substitutions were contributed by the changes within the K-box domain (Fig. 21).

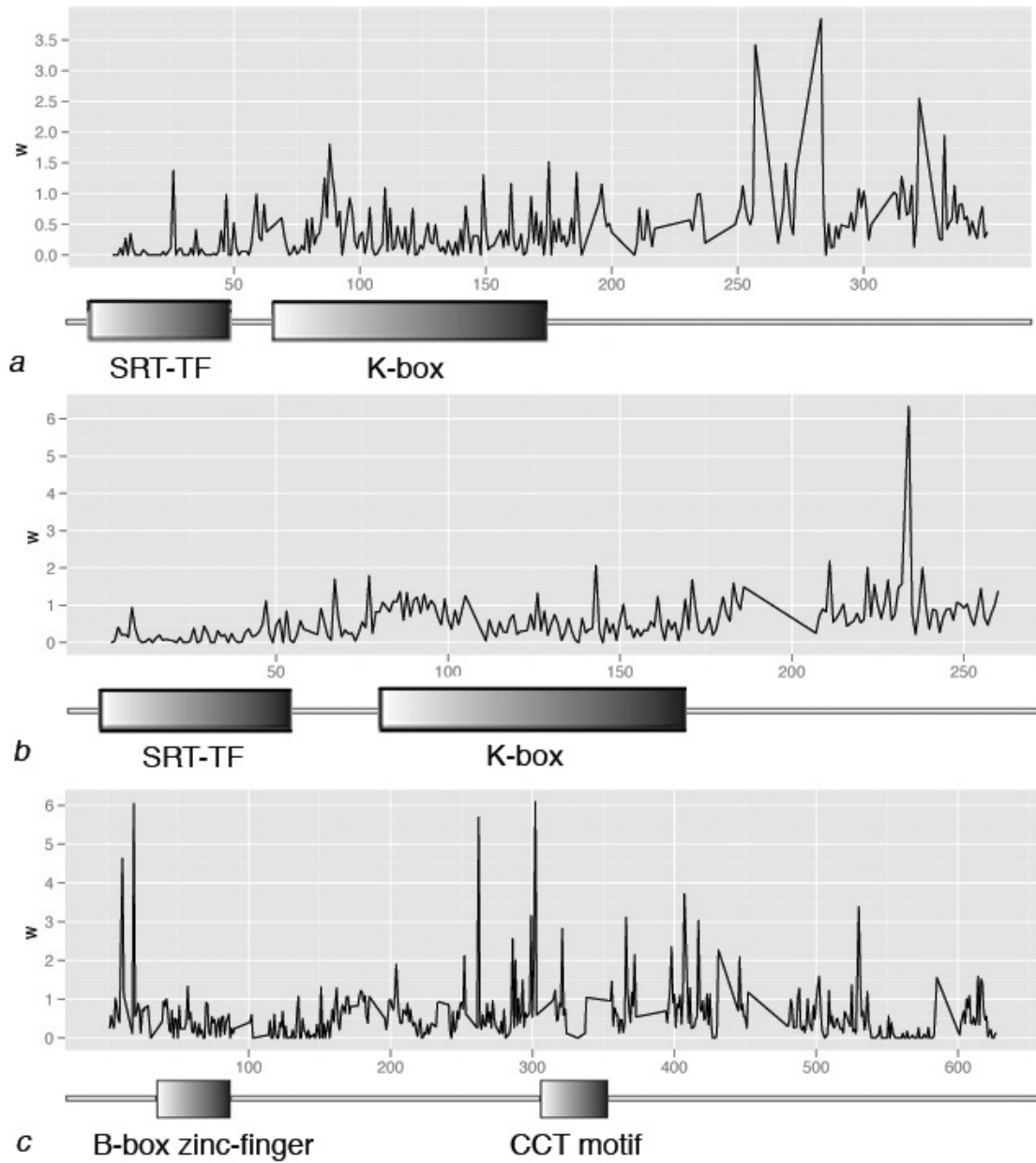


Fig. 21 Site-specific dN/dS ratio (w) in a. *API* (*APETALA1*); b. *FLC* (*FLOWERING LOCUS C*) and c. *CO* (*CONSTANS*), x-axis shows the amino acid sites with functional protein domains indicated by shaded boxes

Hypothesis testing of various scenarios of lineage-specific dN/dS rate variations were conducted in a maximum likelihood framework. The only branch of interest is the branch leading to eudicots (as only eudicot *TFL1* were identified, this analysis could not be conducted in *TFL1*). One of the five competing hypotheses were supported dependent on different genes. In *API*, the three rate class model was supported, indicating that there are specific dN/dS rates for early Angiosperms (0.16), the branch leading to eudicots (0.58) and the eudicot group (0.22). This suggested a change of selective constraint on *API* coincident with diversification of eudicots. In *CO*, the hypothesis that both early Angiosperm and eudicots have the same dN/dS rate while the branch leading to eudicots has a different rate, was supported. Although the test was significant with very small P-value (0), the actual rate difference is very small (0.20 vs 0.18), which suggest a very slight shift of selective pressure. In *FLC*, a eudicot specific dN/dS rate was supported, suggesting a higher level of positive selection (0.3 vs 0.1) within eudicots. In *FT*, a uniform low dN/dS rate (0.09) was supported, suggesting that this gene is very conserved and there is no any evidence of selection on this gene across angiosperm phylogeny (Table 8).

Table 8 LRT results of eudicot specific dN/dS ratio.

	Locus	<i>API</i>	<i>CO</i>	<i>FLC</i>	<i>FT</i>
H0 ^a	-lnL	28201.12616	>50000	22564.42743	10406.51625 ^f
	LRT P-value	1	1	1	1
H1 ^b	-lnL	28192.96474	34416.51375	22560.36447	10407.95421
	LRT P-value	0.0014341	0*	0.0138046	0.453441
H2 ^c	-lnL	28195.0999	34411.16594 ^f	22492.81612	10407.62643
	LRT P-value	0.0046104	0*	0*	0.345527
H3 ^d	-lnL	28190.73401	34411.38872	22482.88909 ^f	10408.03697
	LRT P-value	0.000431144*	0*	0*	0.488749
H4 ^e	-lnL	28187.10954 ^f	34412.92898	22483.43809	10407.78529
	LRT P-value	0.000122388*	0*	0*	0.255258

a H0: One uniform dN/dS ratio across the entire tree; *b* H1: different dN/dS ratio between

the branch and early Angiosperms and eudicots combined; *c* H2: different dN/dS ratio

between eudicots, branch combined and early Angiosperms; *d* H3: different dN/dS ratio

between early Angiosperms, branch combined and eudicots; *e* H4: different dN/dS ratio

among early Angiosperms, eudicots and branch; *f* Hypotheses supported for individual

flowering time gene; * P values statistical significant after Bonferroni corrections for

multiple comparisons

Discussion

Rate variations relative to positions in pathways

Genome evolutionary dynamics is characterized by heterogeneous evolution rates in different genomic regions. The rate variations are caused and shaped by region-specific mutation rates and different functional constraints. The rates of evolution in protein coding regions are largely determined by multiple factors such protein folding physics, complexity of regulation, expression level and evolutionary age (Konin

2009, Vinogradov 2010). However, these determinants were identified in molecular evolutionary analyses based on large number of genes (genome-wide) among small number of taxa. As a result, these studies identify the major causes of evolutionary rate variations with data sets of great genome depth but limited phylogenetic depth. The rate distribution generated represent rate variations in short evolutionary distances as our capabilities to identify correct orthologous genes decrease with increase of genetic distances. Distribution of protein evolution rates suggested a neutral scenario of protein evolution within relative small evolutionary time scale. On the other hand, samplings of selected plastid genes in large number of taxa were conducted for reconstructing angiosperm phylogeny (Soltis *et al.* 1999). Substantial rate heterogeneity was recovered in different phylogenetic lineages with accelerated rates often associated with 'extreme' habitats, indicating that molecular evolution rates may have changed parallel to morphological evolution rates (Bateman and DiMichele 1994). The parallel alternations of molecular and morphological evolution rates suggested functional constraints imposed by environments can be a major determinant of gene evolution rates.

In this study, only small numbers of genes of known functions were surveyed in great phylogenetic depth. The observed patterns do not contradict with the general neutral trend observed by genome-wide studies but represent specific samples of rates in the rate distribution (Lobkovsky *et al.* 2010). As these genes were identified from mutants and their functions were well-established, evolution of these genes is influenced by functional constraints, which differentiate genes with variations in number of accumulated synonymous and non-synonymous substitutions over large evolutionary time scale. In

metabolic pathways, functional constraints on different genes are strongly correlated with the degree of pleiotropy. Previous molecular evolutionary studies on specific metabolic pathways within relatively small taxonomic groups all suggested the general trend that upstream genes evolve slower due to their high degrees of pleiotropy (Ramsay *et al.* 2009).

This study represents the first systematic analysis of evolutionary rates of Angiosperm genes controlling flowering time over large phylogenetic scale, *i. e.* entire angiosperm group). Consistent with previous studies, variation of evolutionary rates at different levels are present among these genes. Not only average evolutionary rates vary among flowering genes across entire angiosperm phylogeny, but within each gene, a general molecular clock was rejected as a result of lineage specific evolution rates. In addition, evolution rates are heterogeneous within genes as a result of different functional constraints on protein domains coded by different regions. In summary, temporal and spatial rate variations are common in flowering genes, which is consistent with the differences of functional constraints imposed on these genes.

However, it is shown that there is no clear correlation between evolution rates and genes' specific positions in flowering control pathways. The lack of this correlation in flowering genes is due to the fundamental structural differences between flowering time control pathways and metabolic pathways. Metabolic pathways are generally linear with decreasing regulation complexity towards downstream genes. Decreasing regulation complexity results in decreasing pleiotropy effects and consequently relaxed selective

constraints. However, flowering time is controlled by networks of cross-talking pathways with higher complexities than metabolic pathways. As shown in Fig. 2 (Chapter 1), different pathways integrate together to control flowering time, and downstream genes like *FT* and *API* do not show decreased pleiotropy effects as they are involved in other flower development process as well. The presence of strong pleiotropy effects of all the genes without decreasing gene regulation complexity suggested that it was not likely to observe increasing evolutionary rates along the pathways towards downstream. In addition, as the genes sampled in this study were identified in a simplified flowering control pathway, 'upstream' genes like *FLC* and *CO* actually are signal integrators for their upstream factors. For example, *FLC* is the signal integrator and the most downstream factor of vernalization pathway, although it shows at an 'upstream' position in the simplified flowering pathway (He 2009). In this sense, the observed higher amino acid substitution rate in *FLC* is consistent with the trend that downstream genes have higher evolution rates, although the upstream genes of *FLC* need to be analyzed to verify the conclusion.

Site-specific and lineage-specific changes of selective constraints

In flowering gene analyzed in this study, temporal and spatial variations of evolution rates are observed, specifically, variations of selective constraints indicted by dN/dS ratio, suggesting dynamic shifts of strength and duration of functional constraints on these genes. Domain specific accelerated amino acid changes were observed in *API* and *FLC*, both are MADS-box transcription factors. Interestingly, in *API*, elevated amino acid substitutions are found mostly outside the functional domain (MADS-box), while in

FLC, elevated amino acid substitutions are found right inside the K-box domain, a major functional part with transcription factor activities. Potential positive selection within MADS-box functional domains of *API* and *FLC* suggested that amino acid substitutions and subsequent subtle changes of transcription factor activities may have significant phenotypic consequences due to the strong pleiotropic effects of these two genes in flower development.

The temporal patterns of selective schemes alternation are dependent on specific evolutionary histories of each gene. For example, over entire angiosperm phylogeny, *FT* gene showed very slow evolution and much lower amino acid substitution rate compared to other genes, indicating a very strong purifying selection on this gene in Angiosperm evolutionary history. This is consistent with the role of *FT* protein as the mobile signal transported from leaves to meristems to promote flowering. This function is very conserved among flowering plants, because: first, *FT* is always single copy in all the flowering plants surveyed; second, similar functions as mobile signal of *FT* were demonstrated in multiple flowering plants; and third, *FT* seems to have function beyond flowering time control, with strong pleiotropy effects. Slow evolution of *FT* gene is expected as a result of strong functional constraints. In contrast, lineage-specific analyses of dN/dS ratio suggested that *API*, *CO* and *FLC* showed various degrees of ratio shifts associated with diversification of eudicots. In *API*, the branch leading to eudicots showed elevated dN/dS ratio; In *CO*, the branch and eudicot group combined showed slightly elevated dN/dS ratio; and In *FLC*, dN/dS ratio increased within eudicots. The lineage-specific shifts of selective constraints are common in other MADS-box genes (Shan *et al.*

2009). Evolution of eudicots, especially core eudicots, is considered representing a major rapid radiation of lineages in flowering plant evolutionary history (Soltis *et al.* 2005). The evolutionary changes associated with this 'deep' radiation of eudicots are still unknown, but are suggested to be radical with significant genotypic and phenotypic effects (Bateman and DiMichele 1994). Based on our observations of significant shifts of selective schemes on strongly pleiotropic flowering genes associated with the lineage leading to eudicots and eudicots themselves, it is suggested that accelerated sequence evolution of plant flowering genes may play an important role in the rapid generation and diversification of eudicot lineages. To further test this hypothesis, not only more sequence data are needed for the taxa close to the origin of eudicots, but also more information about their functional changes are required to decipher the connections between alternation of gene sequences, phenotypic changes and driving forces behind rapid radiation.

Chapter 6

Genome-wide intron size expansion in domesticated grapevine

Abstract

Splicesomal introns are important components of eukaryotic genes. However, their size evolution is rarely studied. To investigate intron size dynamics in flowering plants, in particular domesticated grapevines, a complete survey of intron size and contents of wine grape (*V. vinifera* Pinot Noir) genes was conducted in this study. The transcriptome of *V. vinifera* genes assembled from ESTs was produced and mapped to genomic regions to characterize and analyze splicesomal introns. An unusual expansion of splicesomal intron size was revealed in this plant, inconsistent with overall genome size dynamics. Intron size expansion is related to reduced gene expression but not gene functions. However, in many MADS-box genes, a major class of plant transcription factors, extensive size expansion was found in the first and second introns. The intron size expansion only occurred at specific lineages of MADS-box genes, indicating functional constraint on intron sizes. The composition of expanded introns in genes with conserved structure in flowering plants indicated extensive transposable element (TE) activities within intronic regions. TEs cover about 80% of the intron space in these genes and recent LTR retrotransposon insertions are enriched in intronic regions. Detailed analysis of selected intronic regions in *V. vinifera* cultivars and native species revealed that some TE activities were associated with grape domestication, even to specific cultivars. The transcriptome approach not only recovered the intron size expansion but also improved

annotations of the *V. vinifera* genome. Intron size may still be under functional constraint as they appear to affect gene expression. In particular, lineage-specific size expansion of first introns of MADS-box genes could reflect a balance between selection for shorter introns associated with higher expression levels, and longer introns housing more regulatory elements. In addition, intron size expansion associated with recent TE insertions may be related to grapevine domestication. Most modern grapevine cultivars undergo vegetative propagation and receive intensive human care to maintain their desired properties, which simultaneously promote TE proliferation and repress TE removal mechanisms such as recombination. As a result, the tolerance to introns of extreme sizes may be a by-product of relaxed natural selection and asexual reproduction.

Introduction

Eukaryotic genes contain spliceosomal introns that are post-transcriptionally removed by the spliceosome, an RNA-protein complex (Chow *et al.* 1977). The length, position and phase of spliceosomal introns are important components to the evolution of genome architecture (Irimia and Roy 2008). However, most intronic regions are often considered 'junk' DNAs similar to intergenic regions or other non-coding sequences. Increasing scrutiny of genomic data revealed many conserved non-coding sequences (CNS) in the non-coding DNA of both plant and animal genomes (Elgar 2009). CNS may conduct important functions and experience selective constraints (Keightley and Gaffney 2003). One major function of CNS is regulating gene expression via interactions between small DNA motifs and transcription machinery (transcription factors and RNA polymerase) (Woolfe *et al.* 2005). The regulatory motifs are enriched in 5' promoter regions and the 5'-introns of plant and animal genes (Majewski and Ott 2002).

Plant and animal genomes have diverged greatly in the average size of introns: human and reptile genomes are composed of relatively huge genes, spread across hundreds of kbp by the presence of large size introns (Smith *et al.* 2009). Plant genomes in contrast have more compact genes due to small introns (Hong, Scofield and Lynch 2006).

Arabidopsis, the plant model system first completely sequenced, has an average gene size 2000 bp and average intron size 180 bp (*Arabidopsis* Genome Initiative 2000). The difference in intron size reflects introns' capacity of holding regulatory elements and selective constraints on introns in animal and plant genomes' evolutionary histories.

Intron size is hypothesized to be constrained by energy use in transcription. Large introns may require more energy to be transcribed and spliced so there is selection against intron of excessive size (Castillo-Davis *et al.* 2002). However, in certain genes, selection against intron size may be counteracted by selective preference of bigger introns for more regulatory elements and finer control of gene expression (Marais *et al.* 2005).

Transposable elements (TEs) are known to be major components of plant genomes (Kumar and Bennetzen 1999). The origin and proliferation of TEs in plant genomes shaped plant genome dynamics and genetic diversity. Unlike human genome in which TEs are predominantly in introns (Wong, Passey and Yu 2001), plant TEs are usually found in intergenic regions, possibly a result of strong purifying selection against TE insertions in exons and introns (Bennetzen 2000). The observed genome size expansions of cereals are thought to be driven by whole-genome duplications and TE invasions in non-coding regions (Messing and Bennetzen 2008). However, in a few cases, TE insertions in introns altered temporal and spatial expression patterns of specific genes. The insertions caused significant genetic and phenotypic changes and seem to be preserved by natural selection (Lempe *et al.* 2005). During plant domestication, key traits selected by human may be the result of genetic diversity generated by TE insertions (This *et al.* 2007).

To investigate intron size's evolution in plant genomes, a novel approach is adopted to identify and analyze introns of large size in a domesticated grapevine cultivar (*Vitis vinifera* Pinot Noir). The specific hypothesis to be tested in this study is that *V. vinifera*

experienced a unique intron expansion process. In addition, hypotheses that intron size affects gene expression, intron size expansion is associated with specific gene families, and intron size expansion is primarily driven by insertions of repetitive elements, are also tested by comparisons of ESTs, genomic sequences and annotations, and analysis of intron sequence properties. Generally, it is shown that introns of extraordinary sizes are widespread in cultivated grapevines, an unusual phenomenon not observed in other plant genomes.

Materials and Methods

EST data collection, processing and assembly

All currently available *V. vinifera* Expressed Sequence Tags (ESTs) were obtained from NCBI GenBank using the PartiGene package (Parkinson, Guiliano and Blaxter 2004). Vector and poly-A sequences were trimmed by the built-in functions of PartiGene. Cleaned EST sequences were clustered using both CLOBB (the default clustering method of PartiGene, Parkinson *et al.* 2002) and TGICL (Perteau *et al.* 2003). Both methods employ megaBLAST (Altschul *et al.* 1997) searches to produce clusters based on sequence similarity. For both clustering methods, a similarity cutoff over 99% were employed in addition to the prerequisite of having over 100bp of overlapped region between two ESTs. Both approaches yielded similar number of EST clusters and the clustering results of PartiGene were adopted for further analyses. The clusters were assembled and the consensus sequences for each cluster were predicted by PHRAP (de la Bastide and McCombie 2007).

Identification of large genes and introns

The consensus sequences were mapped to genomic sequences of *V. vinifera* Pinot Noir (Jaillon *et al.* 2007) to identify genomic locations and exon/intron structures of the predicted genes. The mapping process was performed with BLAT, by which ESTs or cDNAs were aligned with genomic regions (Kent 2002). Most alignment gaps in BLAT results represent intronic regions. Genes with potential large introns were selected based on three criteria: 1. The similarity of BLAT search is higher than 99%; 2. The mapped genomic region (excluding introns) covered 95% of the EST consensus sequence; 3. The predicted genomic region is larger than 10 kbp. The first two criteria ensure that predicted gene sequences were mapped to the correct genomic positions, the final one identifies gene candidates with possible large introns. The process was automated with a script for parsing BLAT results, in which the gap sizes were calculated and gap sequences (and coding regions associated with them) were extracted. Extracted gap regions were manually inspected for splicing signals. The same process was also conducted on *Arabidopsis* and *Populus* ESTs and genomes to identify large introns in those two plants.

Genes with potential large introns were submitted to the following analysis: 1. The coding regions were predicted and translated to protein sequences by ESTScan (Iseli, Jongeneel and Bucher 1999); 2. Genes were annotated based on Genoscope annotations and manual corrections with additional information from ESTs; 3. BLAST searches were conducted using the predicted coding regions as queries against *V. vinifera* EST database. The number of high similarity (cutoff E-value = 0.001) hits is positively correlated with

expression level; 4. BLAST searches using predicted protein sequences were conducted against the *Arabidopsis* and *Populus* genomic databases to identify the orthologous genes in the two plant genomes as an indicator of possible functions. All statistical analyses were conducted with the statistical package R (R Development Core Team 2009).

Analysis of large intron contents and selected individual introns

To ensure the candidate genes are normally expressed, only candidates that also had EST sequences from one other *Vitis* (*V. shuttleworthii*) were included to the analysis of intron size and content. Introns larger than 3 kbp were identified (3 kbp is larger than 98% of introns in *Arabidopsis* and poplar) and subjected to the following analysis: 1. The intronic sequences were screened by RepeatMasker (Smit, Hubley and Green, <http://repeatmasker.org>) to identify repetitive elements using *Arabidopsis* repetitive element library; 2. The introns without any repetitive elements detectable by RepeatMasker were subjected to BLAST searches against *V. vinifera* genome. Presence of large number of high score hits (cutoff $e=0.001$) suggests grapevine-specific repetitive elements. 3. Full length LTR retrotransposons in intronic regions were identified by LTR_FINDER (Xu and Wang 2007). The similarities between two LTR regions were used to estimate the age of LTR retrotransposon insertions using the methods described by SanMiguel *et al.* (SanMiguel *et al.* 1998). Jukes and Cantor substitution model was used to correct multiple substitutions in LTR regions (Jukes and Cantor 1990).

Intronic regions were chosen for further biology analysis when they meet the following criteria: 1. There is only one repetitive element in the intron, indicating a relatively recent

expansion; 2. The total intron size without repetitive elements does not exceed 3 kb (to facilitate PCR amplifications). Primers were designed based on the sequences of both exons and conserved regions within the repetitive elements to test the presence and absence of elements in introns (primer sequences see Appendices Table 7). PCR amplifications of selected intronic regions were conducted for six native grapevine species with one individual for each: *V. rotundifolia*, *V. californica*, *V. girdiana*, *V. aestivalis* and *V. labrusca*, *V. Jacquemontii*, and seven different *V. vinifera* cultivars: Cabernert sauvignon, Chardonnay, Dolcetto, Pinot Noir, Riesling, Sangiovese, Zinfandel. Alleles with or without TE insertions were scored by sizes of PCR amplicons.

Identification and molecular evolutionary analyses of MADS-box genes

The MADS-box domain of *Arabidopsis AP1* was used as a query for a tBLASTn search against *V. vinifera* genome to recover all MADS-box domains (cutoff $e=0.1$). The annotation of genes containing MADS-box domains were obtained from Genoscope and manual inspection. Large introns were identified from gene annotations and TEs in introns were identified using the same methods for large introns from genome-wide data. The MADS-box domains were aligned using ClustalW (Thompson, Higgins and Gibson 1994) followed by manual corrections. Phylogenetic analyses were conducted with PAUP* (Swofford 2002).

Results

EST collection, processing and assembly

Totally 353748 publicly available EST sequences of *V. vinifera* were clustered into 50819 clusters, 32896 of them contain only one EST member, and remaining (17923) clusters contain more than one EST sequences. The largest cluster is composed of 1998 EST sequences. The distribution of cluster size suggested that most of *V. vinifera* genes are not expressed at very high level (65% singletons) and only a few (3 clusters contain more than 1000 ESTs) genes are highly expressed in the representative cDNA libraries.

Identification and properties of large genes and introns

After mapping the assembled consensus sequences to *V. vinifera* genomic regions, 996 genes exceeding 10 kb in length were identified. The size and sequence of exons are conserved among grapevine, *Arabidopsis* and *Populus*, so large size introns are the only contributors to difference in gene size. The expression level represented by EST numbers varies across the 996 genes. By comparing these large genes to a sample of 996 clusters randomly selected from *Vitis* ESTs, the distribution of gene expression level is significantly skewed towards lower values in large genes (Fig. 22a). Genes of very large size are more likely singletons or represented by only 2 ESTs (χ^2 test, $p < 0.001$). However, the function or gene ontology of the 996 genes does not show bias to any particular gene family, indicating that gene size (and intron size) is not correlated with gene function.

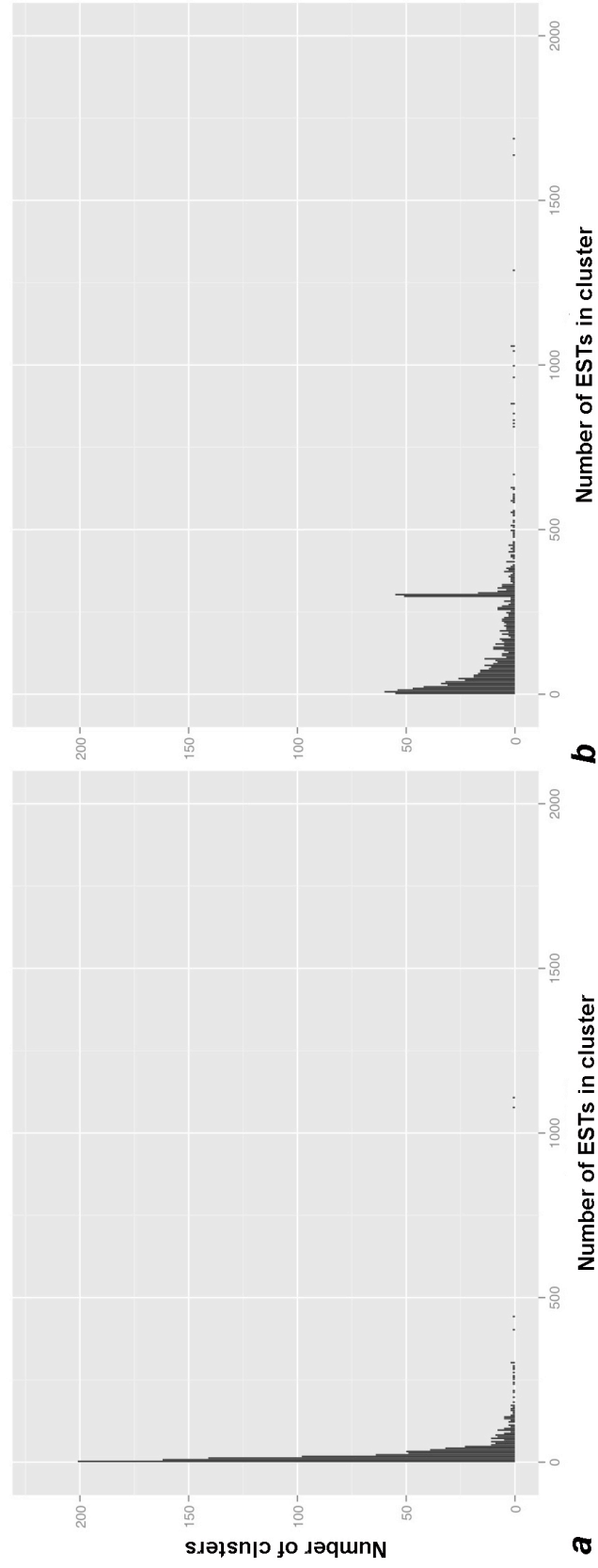


Fig. 22 Expression level represented by EST number in clusters. Bars represent number of clusters with certain number of ESTs: a. 996 large genes identified in *Vitis vinifera*; b. 996 genes randomly selected from all assembled clusters

In the 996 genes, 2370 gaps larger than 3 kb between coding segments were identified (20 in *Arabidopsis* and 54 in *Populus*). Manual inspection of splicing signals indicated that over 90% of automated extracted gaps in BLAT results were true introns with canonical or semi-canonical splicing sites. Out of 996 genes, 177 genes that have at least one *V. shuttleworthii* EST sequence were selected for detailed analysis of intron size and contents. A total of 234 large introns (>3 kb, all with canonical splicing signals GT/AG) were identified, with 43 introns newly annotated by correcting Genoscope annotations using EST data (28 genes incorrectly annotated). The size distribution of these introns revealed an excessive number of 10 kbp-long introns (Fig. 23a). Large introns are predominantly distributed in intron position 1 and 2 (40% of total, Fig. 23b). In the 177 genes, 34 genes have protein location information included in Genoscope annotations (10 ER, 8 mitochondrial, 10 nuclear, 16 plastid). Considering the lack of positioning information of the remaining genes, it is not clear that whether gene size is related to their ultimate locality.

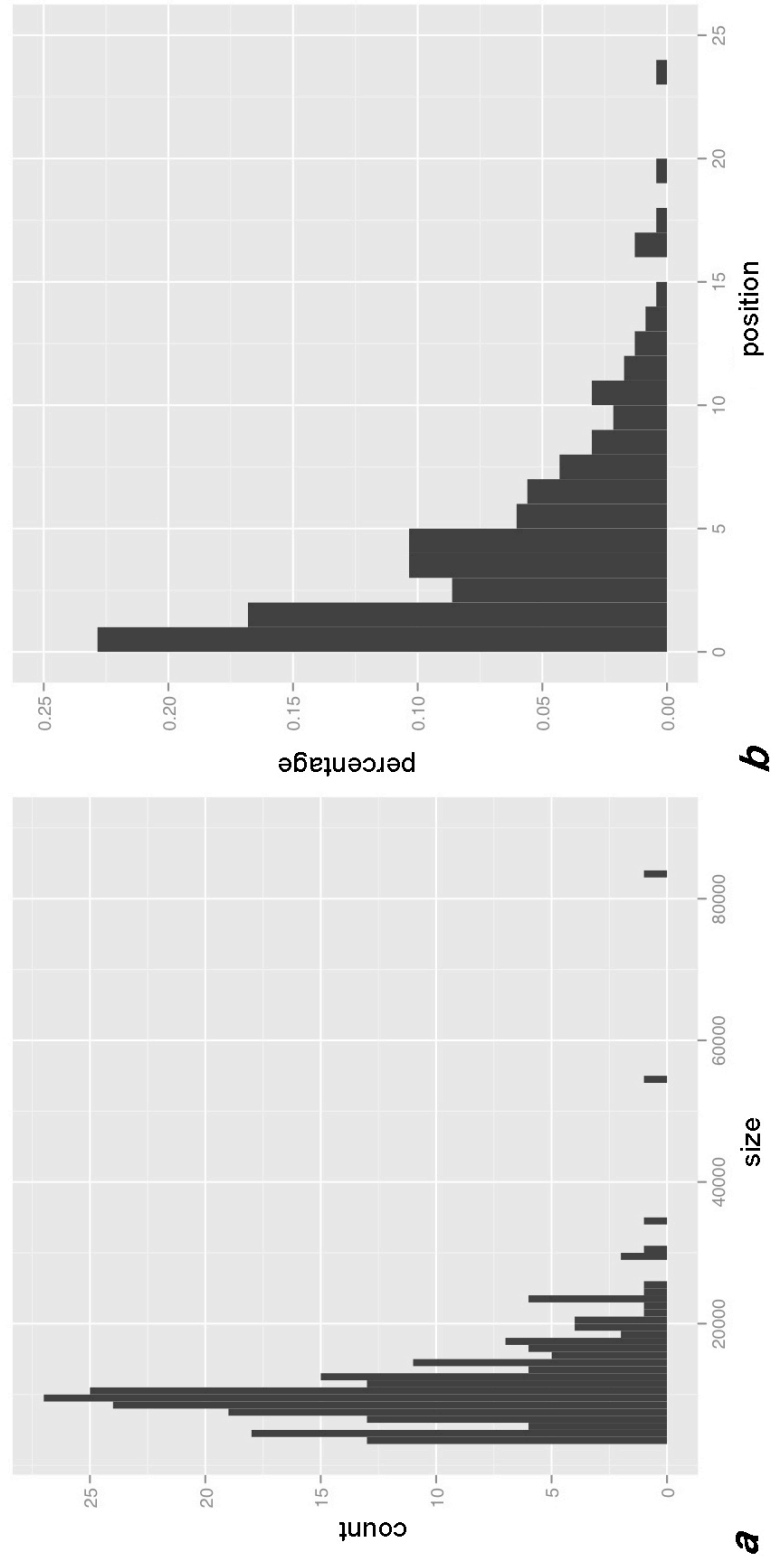


Fig. 23 a. Size distribution of large introns identified in *Vitis vinifera* (>3000 bp). Bars represent number of introns of certain size (bp); b. Distribution of large introns within genes. Bars represent proportions of intron positions within genes

In total, 134 genes have orthologs in *Arabidopsis*, 81 genes have orthologs in *Populus* and 79 genes have orthologs in both *Arabidopsis* and *Populus*, and 39 genes have orthologs in *Arabidopsis* and *Populus* with exon/intron structures conserved among all three, which makes comparison of intron sizes possible. In these genes, 74 large introns were identified. The one-to-one comparison of intron sizes revealed that *V. vinifera* genome experienced a 4-fold expansion compared to *Arabidopsis*, with a 12-fold expansion of intron size. *V. vinifera*'s genome size barely changed but experienced a 5-fold expansion of intron size when compared to *Populus* (Fig. 24b). The increase of intron sizes are generally the same in each individual gene (Fig. 24a).

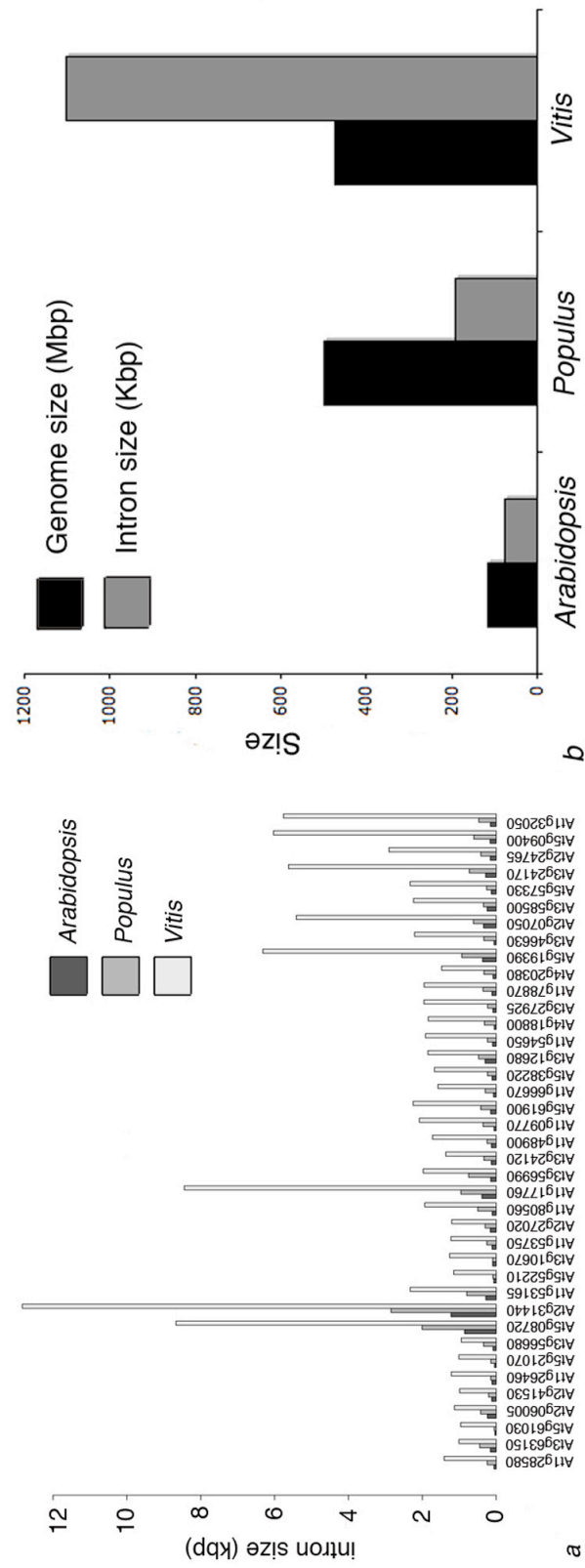


Fig. 24 Intron and genome size comparisons among *Arabidopsis*, *Populus* and *Vitis*: a. Intron size variation at gene-by-gene base; b. Overall intron size (kbp) and genome size (mbp) comparison

The largest gene (with the largest intron) identified in this study is *VvFLC* gene, a *Vitis* homolog to the *Arabidopsis* flowering time gene *FLOWERING LOCUS C (FLC)*. Although *V. vinifera FLC* contains an 83 kbp-long first intron and 16 kbp-long fourth intron, there are over 10 known EST sequences for this gene, indicating that it is normally expressed in *V. vinifera* (Fig. 25).

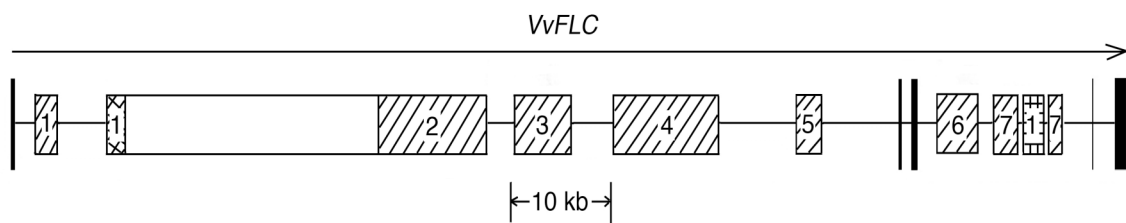


Fig. 25 Annotation of *Vitis vinifera FLC*. Filled bars: exons; lines: introns; open bars: unknown sequences (Ns); hatched bars: LTR retrotransposons; cross-hatched bars: LINEs; square-hatched bars: DNA transposons; numbers indicate the orders of TEs annotated within introns (DNA transposon number 1 inserted into LTR retrotransposon number 7 so LTR retrotransposon number 7 is divided into two parts)

Contents of large introns and TEs

RepeatMasker screening and BLAST searches of the 234 introns suggested that over 80% of the intron contents were repetitive elements. The dominant types of repetitive elements are low complexity sequences, simple repeats, LTR retrotransposons and LINEs. Number of *Copia*-type LTR retrotransposons is about 5 times that of the *Gypsy*-type, in contrast to the genome of another domesticated plant, Medicago (Wang and Liu 2008). The similarities between LTR regions of the 45 complete LTR retrotransposons

are shown in Fig. 26a. When compared to the similarities between LTR regions identified in *V. vinifera* chromosome 1 (used as a random portion of the whole genome, Fig. 26b), LTR regions in introns contain more highly similar pairs (about 50% of the pairs with identity higher than 99%, χ^2 , $P < 0.01$), suggesting younger LTR retrotransposon insertions and recent LTR retrotransposition activities in introns.

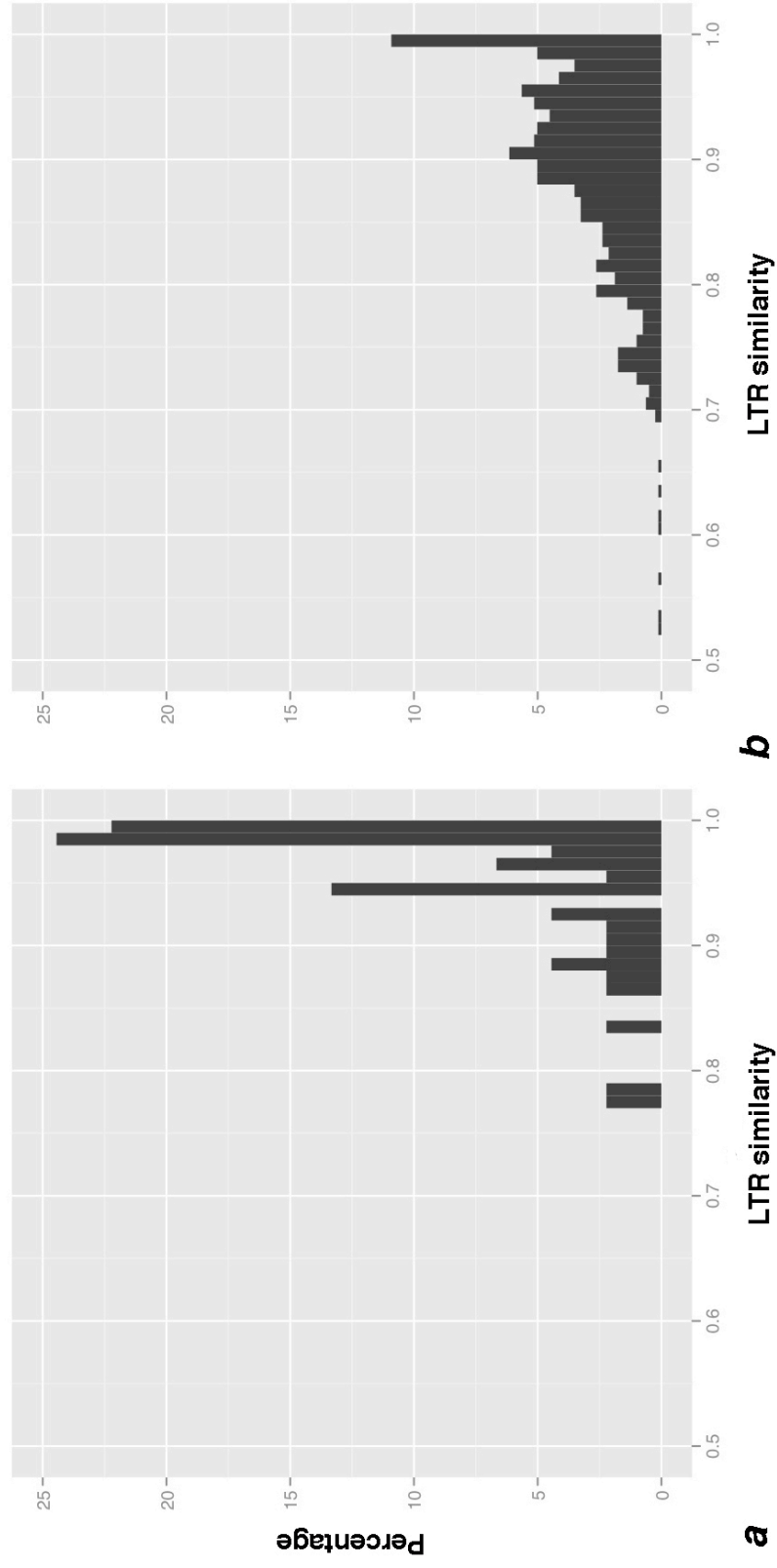
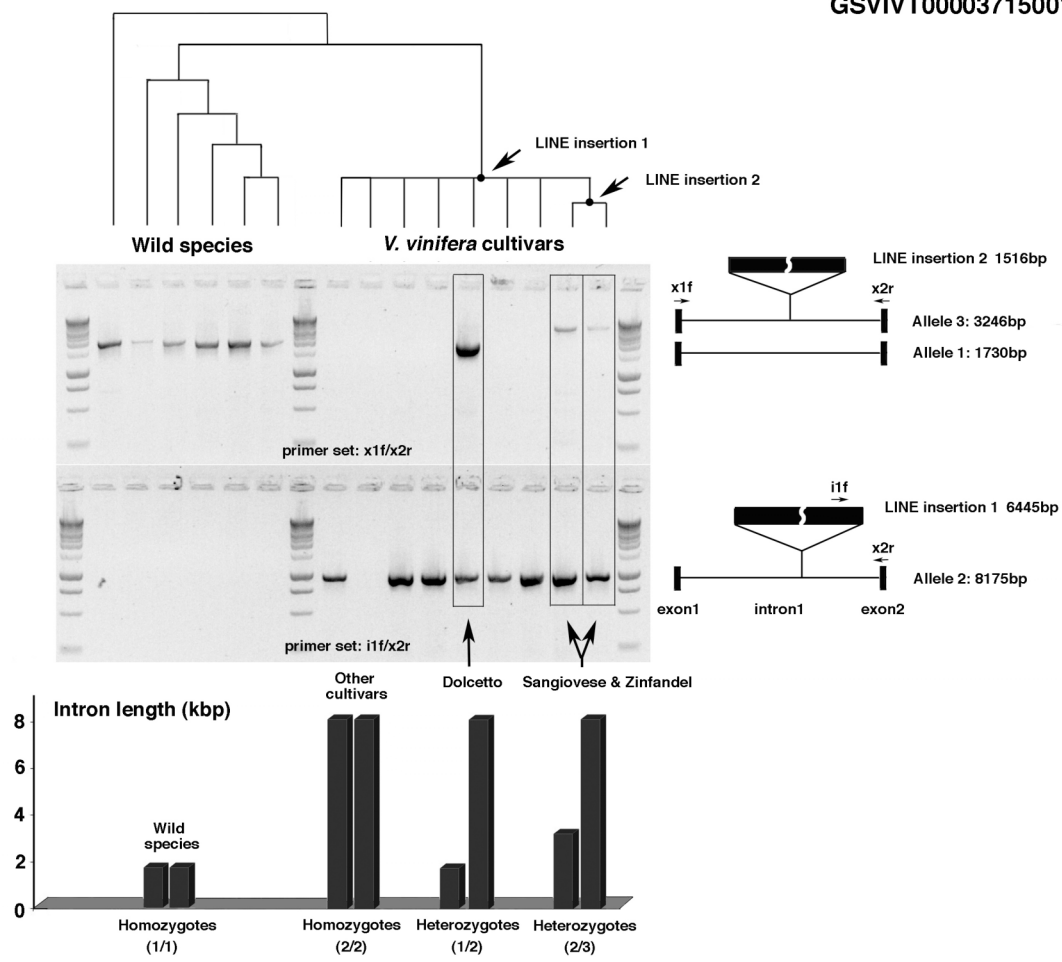


Fig. 26 Distribution of LTR retrotransposon age represented by similarity between LTR regions (the higher similarity, the younger the elements) in a. large introns; b. *Vitis vinifera* chromosome 1

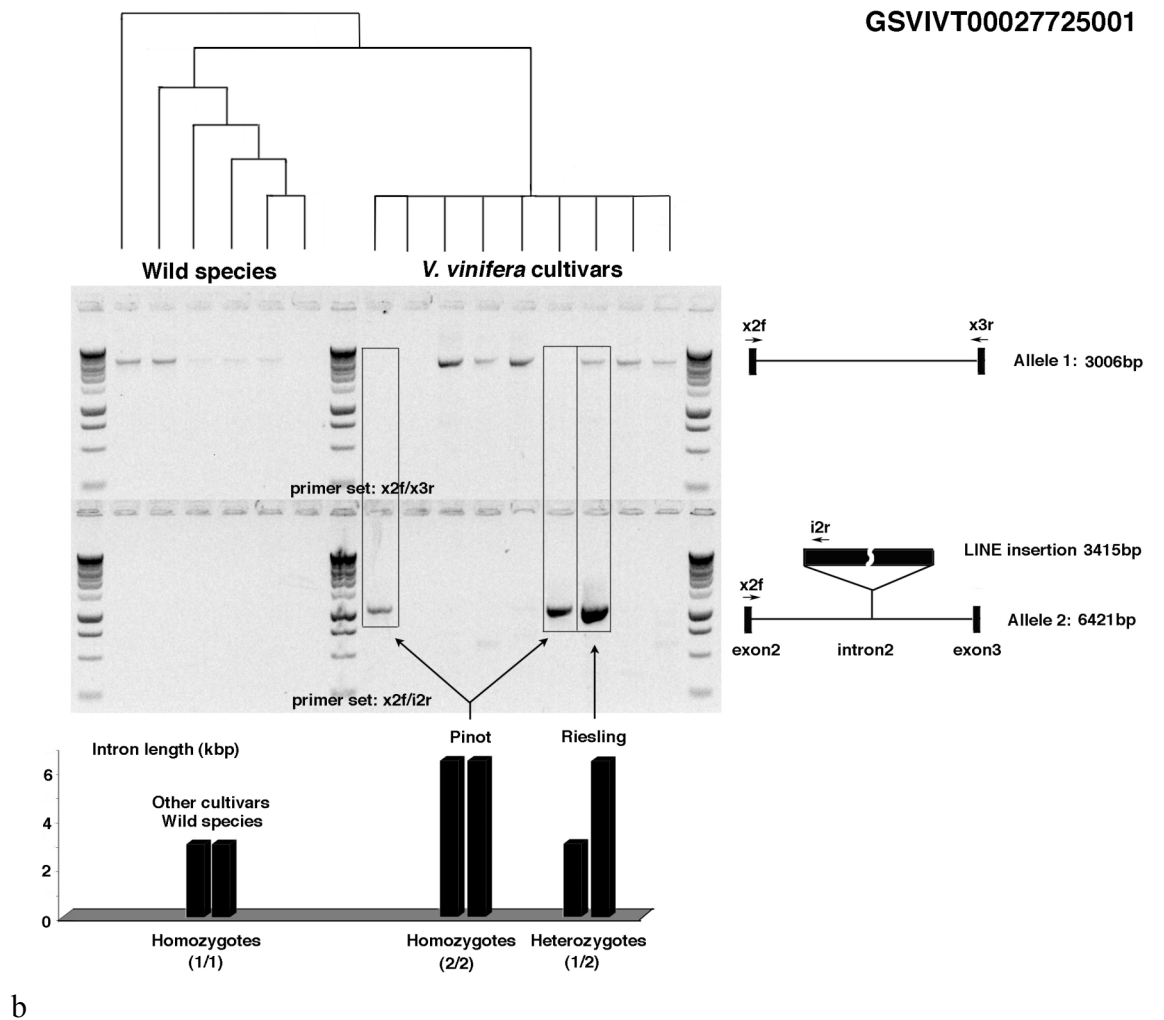
In the 39 genes with conservative exon/intron structures, repetitive elements composed roughly 90% of the 74 intronic regions, in which 10 complete LTR retrotransposons were identified. Among the 74 introns, 1 intron contains only a LINE; 1 intron contains only LTR elements; 15 introns contain LINEs and other repetitive elements; 15 introns contain LTR elements and other repetitive elements; 5 introns contain LTR, LINEs and other repetitive elements; 34 introns contain only other repetitive elements; 3 introns do not contain repetitive elements. According to a BLAST search against *V. vinifera* genome, 'other repetitive elements' were highly repetitive within *V. vinifera* and likely represent un-classified or unnamed grapevine-specific TEs. The three introns without any identified repetitive elements all contain LINE-like ORFs otherwise highly repetitive in the *V. vinifera* genome. Further analyses were conducted on four introns selected from four genes.

Characterizations of 4 selected introns (all containing LINE insertions) in *Vitis* species and varieties revealed intron length variations across these species and varieties (Table 1). None of the native *Vitis* species have LINE insertions in introns. For the first two introns, the size expansion were not found in any other *Vitis* species or varieties, including Pinot Noir, suggesting that the LINE insertion is only present in the specific individual or lineage that was used for genome sequencing. The third intron contain identical LINE insertions in both alleles in Pinot Noir (homozygous), but heterozygous in Dolcetto (Fig. 27a). The fourth intron contain LINE insertions in all varieties of domesticated grapevines. The intron size expansion is homozygous in Pinot Noir, consistent with genomic sequences, as well as Cabernet Sauvignon, Chardonnay and

Riesling. However, Dolcetto is heterozygous with one native-grapevine allele and one Pinot Noir allele. In addition, Sangiovese and Zinfandel contain an additional LINE insertion in one allele. Together with an Pinot Noir allele, both varieties are heterozygous for the intron length and carrying additional TE insertions not revealed by genomic sequences (Fig. 27b). Detailed information on each intron is provided in Table 9.



a



b

Fig. 27 Intron length variations due to LINE insertions in *Vitis* species and *Vitis vinifera* varieties. Phylogenetic tree shows relationships among taxa sampled. Gel picture shows amplification results of alleles shown on the right with primer names LINE insertion size and total allele sizes. Bar graph shows intron size variations among *Vitis* species and *Vitis vinifera* varieties including zygotic status. a. GSVIVT0003715001 intron 2 b.

GSVIVT00027725001 intron 1

Table 9 Characterization of selected introns in wild grapevines and *V. vinifera* cultivars

Gene name	GSVIVT00033984001	GSVIVT00022278001	GSVIVT00027725001	GSVIVT0003715001
Gene functions	nucleotide binding	signal transduction	ATP binding	hydrolase activity
Intron	4	12	2	1
<i>V. rotundifolia</i>	-/-	-/-	-/-	-/-
<i>V. californica</i>	-/-	-/-	-/-	-/-
Wild	-/-	-/-	-/-	-/-
<i>V. girdiana</i>	-/-	-/-	-/-	-/-
species	-/-	-/-	-/-	-/-
<i>V. aestivalis</i>	-/-	-/-	-/-	-/-
<i>V. labrusca</i>	-/-	-/-	-/-	-/-
<i>V. jacquemontii</i>	-/-	-/-	-/-	-/-
Cabernet Sauvignon	-/-	-/-	-/-	P2/P2
Chardonnay	-/-	-/-	-/-	P2/P2
Dolcetto	-/-	-/-	P1/-	P2/-
<i>V. vinifera</i>	-/- ^a	-/- ^a	P1/P1 ^b	P2/P2 ^c
Pinot Noir	-/-	-/-	-/-	P2/P2
Riesling	-/-	-/-	-/-	P2/SZ ^d
Sangiovese	-/-	-/-	-/-	P2/SZ
Zinfandel	-/-	-/-	-/-	P2/SZ

Gene names and functions follow Genoscope terminology and annotations (Jaillon et al. 2007). Dashes (-) represent wild-type alleles. Letters (P1, P2, SZ) represent alleles with TE insertions, discovered by genomic analyses. *a.* Wild-type alleles according to PCR, TE inserted alleles according to genomic sequences, may be unique to the sequenced individual of Genoscope. *b.* Pinot Noir allele according to both PCR and genomic sequences (P1: 6421 bp). *c.* Pinot noir allele according to both PCR and genomic sequences (P2: 8175 bp). *d.* Sangiovese and Zinfandel alleles according to PCR, not found in genomic sequences (SZ: 3246 bp)

Intron size expansion in MADS-box genes

In *V. vinifera* genome, 66 MADS-box domains were identified, out of which 61 genes including at least one MADS-box domain were annotated in Genoscope. There are 12 genes encode proteins with MADS-box domains only (no other peptides) and 4 genes encode proteins with two MADS-box domains. None of them have corresponding ESTs or cDNAs so they are very likely incorrect annotations. A large portion of genes containing MADS-box domains (24) occur in clusters, *i. e.* within 200 kb of physical distance. Among the 61 genes, 17 genes have *Arabidopsis* orthologs; 22 genes have *Populus* orthologs; 16 genes have both orthologs; 7 genes have conserved exon/intron structures with *Arabidopsis* and *Populus*. In all MADS-box genes, 30 large introns (> 3 kb) were identified. MADS-box large introns are predominately first intron (18 first introns (60%), one 5'-UTR and seven second introns). Detailed annotation of these introns also revealed significant number of repetitive element insertions. Six complete LTR retrotransposons were identified in MADS-box large introns. The average similarity of LTR regions is 0.9345, suggesting older ages of LTR retrotransposons in MADS-box introns than overall large introns. Over two thirds (40) of all MADS-box genes do not have corresponding EST data, and 65% (26) of them have at least one TE insertion in their introns. This suggested that significant number of null or weak alleles of MADS-box genes were produced by TE insertions in introns. However, 20 out of the 24 MADS-box genes with EST data contain at least one TE insertion in their introns, suggesting that TE invasions did not abolish the expression of these genes. Mapping large introns to a phylogenetic tree of MADS-box domains revealed that intron size expanded in specific lineages (Fig. 28).

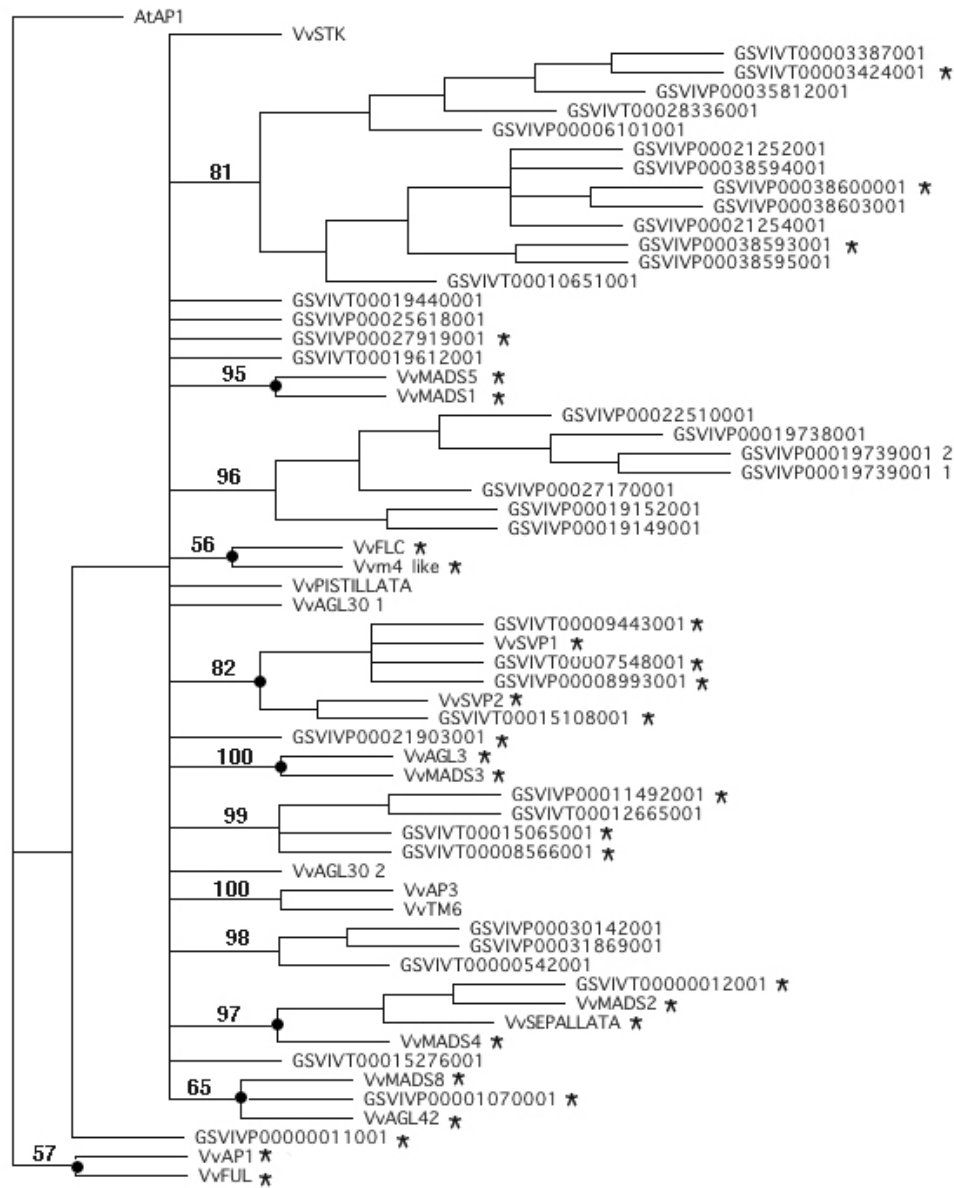


Fig. 28 Lineage-specific intron size expansion in MADS-box genes. Asterisks denote genes with intron larger than 3 kb, closed circles denote gene lineages with all descendants having intron larger than 3 kb. Genes with identified orthologous genes in *Arabidopsis* and *Populus* were named after the orthologs. Other gene names follow Genoscope terminology. Bootstrap values shown on nodes

Discussion

Transcriptome and intron identification

In most genomic studies, introns are identified by gene annotations. The investigation of intron size and structure is limited by the slow process of generating high quality gene predictions and annotations. As a result, the investigation of intron evolution has been limited to single genes or gene families, and genome-wide intron analysis is rarely conducted, especially in plants (Wendel *et al.* 2002a). In this study, introns at a whole genome scale are identified by combining a large EST data set and genomic sequences in an automated workflow. This approach was rarely used before because sufficient EST and genomic data are needed and they are only available in a few plants (Lanier *et al.* 2008). Domesticated grapevine ranks in the top 25 taxa with the largest collections of ESTs and is the fourth whole plant genome sequence available, making it ideal for the EST mapping approach described here. Using ESTs was not only facilitated by a streamlined process, but also required by the nature of this study: intron size can affect gene expression (Morello and Breviario 2008), therefore the candidate genes with large introns have to be expressed, which can be verified by the presence of EST sequences. EST cluster number in this study is higher than the predicted gene number of *V. vinifera* (Jaillon *et al.* 2007), suggesting that gene prediction based on EST consensus sequences covered a gene range comparable to genomic annotations. By manually checking selected matches between EST consensus sequences and genomic regions, the rate of correct mapping is 100%. Therefore, our workflow correctly identified the corresponding genomic regions of EST consensus sequences, thus correct intron positions and sizes.

A high rate of annotation error was revealed by comparison of ESTs to current annotations. This was expected in two ways: correct prediction of genes in genomic sequences is very sensitive to gene size as large size genes are more likely to be predicted incorrectly (Wang *et al.* 2003); the large introns present in these genes tend to cause prediction algorithms to split exons at both ends of large introns into two separate genes. Incorporating ESTs into gene prediction can clearly refine genome annotation as suggested by Coyne *et al.* (2008).

Properties of large genes

The genes with large introns were selected based on their overall coverage of genomic regions. In both *Arabidopsis* and *Populus*, genes larger than 10 kbp are very rare (10 and 50, respectively). In contrast, about 1,000 genes larger than 10 kbp were identified in *V. vinifera*. In these genes, the major components of gene size are introns larger than 3 kb as exons sizes are about the same in all compared plant genomes. The large intron size was expected to require more energy to be transcribed and spliced, so large introns may decrease gene expression (energy cost hypothesis) (Castillo-Davis *et al.* 2002). This hypothesis was supported by this study: the expression level of large genes, as represented by number of ESTs, is significantly lower than a random sample of about 1,000 predicted genes. However, there was evidence against the hypothesis in genes expressed in specific mammal organs/tissues (Huang and Niu 2008). Our results supported the hypothesis at one end of the intron size spectrum: genes with large introns are expressed at lower levels than genes randomly selected from the grapevine genome. In contrast, gene size seems to be not related to gene function in this study. Among

identified large genes, there is not particularly overrepresented gene family. Intron size may be under a balanced selection scheme as small introns are selected in highly expressed genes and large introns are selected in certain genes to house more regulatory elements (Hughes, Buckley and Neafsey 2008). In our data, there is still about 20% intronic regions not accounted for by similarity searches or repeats masking, suggesting unique sequences in introns are also evolving. Most of these unique sequences may be evolving neutrally since there may not be any selective constraint on them, as they will be spliced out of mature mRNA. However, a few of them may contain very important regulatory elements controlling gene expression patterns. For regulatory elements, not only their sequences but also their locations are crucial for their functions. Intron length dynamics may greatly change regulatory elements landscape by simply alter the distance between elements and other important sites for gene transcription. This is consistent with our observation that large introns are predominant first and second introns, which contain most of the regulatory elements (Bradnam and Korf 2008).

However, if not only large genes identified by mapping ESTs but also available annotations are considered, large introns are common in the MADS-box genes of *V. vinifera*. The expanded introns in this family are predominantly the first introns, which are proposed to contain regulatory elements (Hughes, Buckley and Neafsey 2008, Bradnam and Korf 2008). The size expansion of the first introns is consistent with the complex expression patterns of MADS-box genes. Interestingly, one of the crucial MADS-box genes of flowering process control, *APETALA1* (*API*), have large first introns (~3 kbp) in all three plants. The first introns of *Arabidopsis* and *Populus API*

genes are among the largest introns in these two species. Considering the central role of *API* in flower development, the extraordinary sizes of the first introns suggested regulatory elements in first introns for fine control of temporal and spatial expression pattern.

TEs within introns and gene expression

Detailed annotation of expanded introns revealed that a large portion (80%) of their contents is composed of repetitive elements, also known as transposable elements (TEs). TEs play important roles in plant genome evolution, such as affecting genome size (Vitte and Panaud 2005), shaping sex chromosomes (Kejnovsky *et al.* 2009) and contributing to expressed genes (Bennetzen 2005). Proliferation of TEs was considered a major source of plant genome expansion (Vitte and Panaud 2005). Intron expansion in domesticated grapevine was largely contributed by TEs, even though genome sizes are about the same in *Populus* and *Vitis*. Based on available genomic sequences from various plants, TEs are rarely found in plant exons and introns, but are very abundant in intergenic sequences (SanMiguel *et al.* 1996). There seems to be strong purifying selection against genes disrupted by TE insertions so most of the TE invaded genes were eliminated as deleterious mutations. However, in a few cases, TE insertions in coding regions or nearby non-coding regions of expressed genes were not only preserved but also led to significant evolutionary changes. For example, TEs inserted in the promoter regions of grape *VvMybA1* gene completely ceased the expression of this gene and resulted in the white skin color of the berries (This *et al.* 2007). Multiple TEs in the first introns of *Arabidopsis FLC* gene created null or weak alleles and altered the flowering time

(Michaels *et al.* 2003). In all cases, gene expression is greatly attenuated or even completely abolished by TE insertions near or within genes. In the grapevine MADS-box gene family, over half of the genes with large introns do not have expression (EST) data, suggesting that large intron size also created null or weak alleles. In genome-wide studies, TE insertions in or near introns are also shown affecting gene expression (Pereira, Enard and Eyre-Walker 2009, Luehrsen and Walbot 1992). In addition, TEs in introns can also change tissue-specificity of gene expression (Greene, Walko and Hake 1994). However, in this study, the genes with introns containing TEs all have corresponding EST sequences, suggesting that they are likely normally expressed. This is not evidence against the energy cost hypothesis since expression of these genes may be only reduced by TE insertions in introns but not completely abolished, which still suggests the negative effect of large introns and TEs on gene expression. As presented in the previous section, the presence of large introns in these genes does reduce their expression level, indicating weak alleles were created by TE insertions.

Intron size expansion, TEs, grapevine evolution and domestication

Comparisons of genome sizes and intron sizes between *Populus* and *Vitis* revealed that intron sizes expanded much more than genome sizes. Wendel *et al.* (2002a, 2002b) proposed that intron size dynamics may be decoupled with genome size evolution by showing that intron size stayed static in multiple rounds of genome expansion and contraction in cotton (*Gossypium*). However, it is shown that *Vitis* genome size barely changed while intron size experienced a 5-fold expansion compared to *Populus*. Intron size dynamics in *Vitis* seem to be decoupled from genome size evolution but our data

suggested expansion instead of conservation of intron size.

Lineage-specific intron size expansion in MADS-box genes suggested that some size increases in the gene family are ancestral to gene duplications. The maintenance of large introns in paralogous genes suggested that distribution of large introns in this family was shaped by natural selection in lineage leading to *Vitis*: in certain lineages, large introns were preserved for regulatory elements within them; in other lineages, large introns were selected against since they interfere with gene expression. Currently accepted angiosperm phylogeny suggests that Vitaceae is sister to the entire rosid clade (A. P. G. 2003). This family, including the genus *Vitis*, diverged from the common ancestor with rosids very early and experienced unique morphological character evolution. Here, size expansion of introns seems to be parallel to unique evolutionary history of *Vitis*. In particular, its unique floral traits may be a result of intron expansion in genes playing important roles in flower development such as MADS-box genes. In grapevines, MADS-box gene lineages with large introns may be related to grapevine specific phenotypes while lineages without large introns may confer conservative developmental functions.

Several studies have estimated the age of LTR retrotransposon insertions in plant genomes such as rice, Medicago and maize (Vitte, Panaud and Quesneville 2007, Wang and Liu 2008 and SanMiguel *et al.* 1996). Peaks of LTR retrotransposon activities were identified including some that were very recent and possibly related to plant domestication (Moisy *et al.* 2008 and Wawrzynski *et al.* 2008). In grapevine, age estimates of LTR retrotransposons in introns identified two bursts of LTR

retrotransposon activities: <10,000 ya (LTR similarity 100%) and 1 Mya (LTR similarity 94%) (Fig. 26b). Vitte and Panaud (2005) suggested that plant LTR retrotransposons evolved following a 'burst and contraction' model, which can explain the observed 'peaks' of LTR retrotransposon activities in various plant genomes. In grapevines, the estimated ages of LTR retrotransposon bursts are much younger than the divergence of Vitaceae from the common ancestor of rosids, suggesting that high retrotransposon activities, as intron size expansion, are also associated with unique evolutionary changes along the grapevine lineage.

Vitis vinifera expanded introns contain significantly more identical LTR pairs, suggesting that more young LTR retrotransposons were preserved in introns of domesticated grapevine (Fig. 26a). Here it is suggested that the most recent burst of LTR retrotransposons in grapevine may be associated with domestication, either as a result of artificial selection on the nearby genes or genomic regions (local effect), or a non-adaptive by-product of the domestication bottleneck and specific reproduction modes in domesticated grapevines (genome-wide effect). There is no enrichment of any particular gene families in genes with TE invaded introns. So it is very unlikely that all of the TEs and introns, or nearby genes with diverse functions, randomly distributed in genome, were selected by human in the domestication process. Instead, LTR retrotransposons in the recent burst were more likely to be preserved in introns than previous bursts due to vegetative propagation of domesticated grapevine plants. First, somaclonal propagation of plants (or tissue culture) was suggested to promote or induce LTR retrotransposon activity (Tsukahara *et al.* 2009 and Wessler, Bureaua and Whitea 1995). Domesticated

plants with vegetative propagation may generate more LTR retrotransposons than wild species and non-clonal cultivated plants. Second, plant genome expansions caused by TE proliferation are counteracted by rapid DNA removal as a result of recombination (Hawkins *et al.* 2009). However, due to domestication bottleneck and vegetative propagation, the linkage disequilibrium (LD) in cultivated grapevines can be found at extremely long distance compared to wild species (Barnaud *et al.* 2009), suggesting that recombination is repressed in large chunks of chromosomes. TEs, in particular LTR retrotransposons, in domesticated grapevines, may not be removed as effectively as in wild species. Third, LTR retrotransposons, as mutagenesis factors, may cause deleterious mutations if inserted in genes. The mutants were selected against and removed from population very quickly in natural environments (Baucom 2009). In domesticated plants, intensive care provided by human produces an environment with relaxed selective constraints, in which LTR retrotransposons invaded genes were more likely preserved in populations. Particularly, vegetative propagated cultivars such as domesticated grapevine have a uniform genetic background as a result of clonal reproduction to preserve desired traits. This reproduction mode facilitated the spread and fixation of introns with LTR retrotransposons (and other TEs) to entire clonal population or variety. This may be the major reason why excessive TEs were not observed within genes of other plant genomes, including domesticated plants without clonal propagation such as rice (Naito *et al.* 2006 and Gao *et al.* 2002). Recent, variety-specific TE insertions in introns were discovered in our characterization of large introns from multiple cultivated grapevine varieties, suggesting that TE invaded introns did spread with clonal propagations (see results, Table 9, and Costa *et al.* 2009). Interestingly, the length variation in these introns suggested

relatedness, or history of interbreeding of *V. vinifera* varieties. For example, in several varieties, heterozygosity of intron length was revealed (Table 10). If confirmed in more individuals, heterozygosity suggests possible interbreeding experienced by these otherwise asexually reproducing varieties. The intron length variation actually raised the question about the purity of certain wine grape varieties and their breeding history, as well as relationships among varieties. Interestingly, although homozygous for all the introns characterized, the sequenced genome of *V. vinifera* Pinot Noir showed high genome-wide heterozygosity (about 11% of all genomic regions), suggesting a possible interbreeding origin of this ancient variety (Jaillon O, *et al.* 2007).

In this study, the intron size dynamics in domesticated grapevine suggested that intron size expansion, mainly caused by TE invasions, is associated with the plant's unique evolutionary history. Intron size expansion and TE activity bursts may all be related to major evolutionary changes in grapevine lineage. Investigation of intron size expansion not only revealed a one-of-a-kind pattern of plant genome dynamics, but also raised an interesting question: the roles of introns, a somewhat ignored portion of genome, in plant genome architecture, function and evolution.

Summary

In this study, hypotheses about the evolution of native North American grapevines, including phylogenetic relationships, ecological niches, phenological divergence (with a focus on flowering time), molecular evolution of flowering time genes, as well as genome dynamics of domesticated grapevine, were tested with several approaches including molecular phylogenetics and evolution, specimen informatics and computational analyses of genomic sequences.

A preliminary phylogeny of native North American grapevines was reconstructed using multiple nuclear genes, resolving several monophyletic groups. Evolution of *RPB2* gene, one of the nuclear genes, was investigated beyond the Vitaceae family in the context of angiosperm evolution, revealing ancient gene duplication event and subsequent lineage-specific loss of duplicates. Based on geographic, ecological and phenological information extracted from *Vitis* specimens, geographic distribution, ecological niches and flowering times were analyzed, suggesting complex biogeographic patterns, ecological niche partitioning and divergence of flowering times among *Vitis* spp, confirming that native *Vitis* are isolated by ecological and/or phenological factors. In an association study of molecular evolution and flowering time divergence, *Vitis* orthologs of genes controlling flowering time were characterized and subjected to extensive molecular evolutionary analyses. A microsatellite region was discovered in front of flowering inhibitor, *TFL1* gene, with potential to affect species-specific expression pattern. Alternative transcription of *FCA* similar to *Arabidopsis* was discovered in *Vitis*, suggesting a conserved *FCA* auto-

regulation pathway between *Arabidopsis* and *Vitis*. However, no significant molecular evolutionary patterns were found associated with flowering time changes, suggesting genetic changes beyond flowering time gene coding regions. In the context of angiosperm evolution, flowering time genes showed variation in evolutionary rates and selective constraints. As the signal integrator of vernalization, a major flowering time control pathway, *FLC* gene showed high amino acid substitution rate. In several flowering time genes, accelerated protein evolution was found coincident with eudicot diversification. In addition, the 83kb-long first intron of *V. vinifera FLC* gene, extraordinary for plant genes, led to a genome-wide characterization of large size intron in the domesticated *V. vinifera* as well as other *Vitis* spp, revealing over 2700 introns larger than 3kb in *V. vinifera* genome. Large introns are predominantly distributed in 5' end of the genes and lowered gene expression. Repetitive elements, the major components of intronic sequences, were shown to be recently inserted into *V. vinifera* genome. Individual, variety and species-level intron length variations due to TE insertions of different age were discovered in domesticated grapevine varieties, suggesting that maintenance of large introns may be facilitated by domestication and subsequent vegetative propagation.

This study provided a foundation for further investigation on evolution and speciation in native North American grapevines. With the phylogenetic hypotheses and molecular data from more individuals, preliminary species boundaries can be delineated. Comparisons of biogeography and molecular data at population level, history of *Vitis* speciation may be illustrated. The innovative specimen informatic approaches used in this

study may be extended to other organisms with large number of museum or herbarium specimens, to investigate the biogeography and ecology. Based on characterized flowering time genes, further functional studies, such as gene expression analyses, may reveal the genetic changes underlying flowering time divergence between *Vitis* species. Finally, extensive intron length variation revealed genome diversity at and below species level, which can also reveal relatedness and breeding history among varieties of domesticated species. In addition, intron size expansions near functionally important genes are candidates for the genetic basis of important traits and targets of crop domestication and improvement. This study utilized various genetic, genomic and herbarium resources of grapevine to establish a base for future research on its evolution and domestication, including phylogeny, ecology, development and genome dynamics.

References

- Alonso-Blanco C, Mendez-Vigo B, Koornneef M. 2005. From phenotypic to molecular polymorphisms involved in naturally occurring variation of plant development. *Int J Dev Biol.* 49(5-6):717-32.
- Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, Vreugdenhil D, Koornneef M. 2009. What has natural variation taught us about plant development, physiology, and adaptation? *Plant Cell.* 21(7):1877-96.
- Altschul SF, Madden TL, Schaffer AA, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- Angiosperm Phylogeny Group. [A.P.G.] 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linnean Soc.* 141: 399-436.
- Antonovics J. 2006. Evolution in closely adjacent plant populations X: long-term persistence of prereproductive isolation at a mine boundary. *Heredity.* 97(1):33-7.
- Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 408(6814):796-815.
- Avise JC. 2004. *Molecular Markers, Natural History, and Evolution* (2nd ed.), Sinauer Associates, Sunderland, Massachusetts. 283p.
- Barnaud A, Laucou V, This P, Lacombe T, Doligez A. 2009. Linkage disequilibrium in wild French grapevine, *Vitis vinifera* L. subsp. *silvestris*. *Heredity.* [Epub ahead of print]
- Bart HL, Rios NE. 2007. GEOLocate. Tulane University Museum of Natural History.
- Bateman, R. M., and W. A. DiMichele. 1994. Saltational evolution of form in plants: a neoGoldschmidtian synthesis. In D. S. Ingram and A. Hudson [eds.], *Shape and Form in Plants and Fungi*, 63-100. Academic Press, London.
- Baucom R, Estill J, Leebens-Mack J, Bennetzen J. 2009. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. *Genome Res.* 19(2):243-54.
- Bennetzen J. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology.* 42(1):251-69.
- Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev.* 15(6):621-7.

- Blanchette M, Tompa M. 2003. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.* 31(13):3840-2.
- Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, Weigel D. 2007. Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biol.* 5(9):e236.
- Borner R, Kampmann G, Chandler J, Gleissner R, Wisman E, Apel K, Melzer S. 2000. A MADS domain gene involved in the transition to flowering in *Arabidopsis*. *Plant J.* 24(5):591-9.
- Bradley NL, Leopold AC, Ross J, Huffaker W. 1999. Phenological changes reflect climate change in Wisconsin. *Proc Natl Acad Sci U S A.* 96: 9701-9704.
- Bradnam K, Korf I. 2008. Longer first introns are a general property of eukaryotic gene structure. *PLoS One.* 3(8):e3093.
- Buckler ES, et al. 2009. The genetic architecture of maize flowering time. *Science.* 325(5941):714-8.
- Bustamante E, Búrquez A. 2008. Effects of plant size and weather on the flowering phenology of the organ pipe cactus (*Stenocereus thurberi*). *Ann Bot (Lond).* 102(6):1019-30.
- Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD. 2004. Epistatic interaction between *Arabidopsis* *FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. *Proc Natl Acad Sci U S A.* 101(44):15670-5.
- Carmona MJ, Calonje M, Martínez-Zapater JM. The *FT/TFL1* gene family in grapevine. *Plant Mol Biol.* 63(5):637-50.
- Castillo-Davis C, Mekhedov S, Hartl D, Koonin E, Kondrashov F. 2002. Selection for short introns in highly expressed genes. *Nat Genet.* 31(4):415-8.
- Chase MW, et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden.* 80:526-580.
- Chow L, Gelinas R, Broker T, Roberts R. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell.* 12(1):1-8.
- Comeaux BL. 1984. Taxonomic studies on certain native grapes of the eastern United States. Ph. D Thesis, North Carolina State University, Raleigh. U. S. A.
- Conti L, Bradley D. 2007. *TERMINAL FLOWER1* is a mobile signal controlling *Arabidopsis* architecture. *Plant Cell.* 19(3):767-78.

Costa J, de Melo D, Gouveia Z, Cardoso H, Peixe A, Arnholdt-Schmitt B. 2009. The alternative oxidase family of *Vitis vinifera* reveals an attractive model to study the importance of genomic design. *Physiol Plant*. 137(4):553-65.

Coyne RS, Thiagarajan M, Jones KM, Wortman JR, Tallon LJ, Haas BJ, Cassidy-Hanley DM, Wiley EA, Smith JJ, Collins K, Lee SR, Couvillion MT, Liu Y, Garg J, Pearlman RE, Hamilton EP, Orias E, Eisen JA, Methé BA. 2008. Refined annotation and assembly of the *Tetrahymena thermophila* genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. *BMC Genomics*. 9:562.

Crimmins T, Crimmins M, Bertelsen D, Balmat J. 2008. Relationships between alpha diversity of plant species in bloom and climatic variables across an elevation gradient. *International Journal of Biometeorology*. 52: 353-366.

Delisle FC, Lavoie MJ and Lachance D. 2003. Reconstructing the spread of invasive plants: taking into account biases associated with herbarium specimens. *J Biogeogr*. 30:1033–1042.

Devaux C, Lande R. 2008. Incipient allochronic speciation due to non-selective assortative mating by flowering time, mutation and genetic drift. *Proc Biol Sci*. 275(1652):2723-32.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11-15.

Doyle JJ. 1997. Trees within trees: genes and species, molecules and morphology. *Syst Biol*. 46(3):537-53.

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution*. 63(1):1-19.

Elgar G. 2009. Pan-vertebrate conserved non-coding sequences associated with developmental regulation. *Brief Funct Genomic Proteomic*. 8(4):256-65.

ESRI, 2009. ArcGIS software.

Fitter AH, Fitter RS. 2002. Rapid changes in flowering time in British plants. *Science*. 296: 1689-1691.

Flowers JM, Hanzawa Y, Hall MC, Moore RC, Purugganan MD. 2009. Population genomics of the *Arabidopsis thaliana* flowering time gene network. *Mol Biol Evol*. 26(11):2475-86.

Gallaher RV, Hughes L, Leishman MR. 2009. Phenological trends among Australian alpine species: using herbarium records to identify climate-change indicators. *Australian Journal of Botany*. 57:1-9.

- Gao L, McCarthy E, Ganko E, McDonald J. 2004. Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequences. *BMC Genomics*. 5(1):18.
- Getz W, Wilmers C. 2004. A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. *Ecography*. 27: 489-505.
- Gimaret-Carpentier C, Dray S, Pascal JP. 2003. Broad-scale biodiversity pattern of the endemic tree flora of the Western Ghats (India) using canonical correlation analysis of herbarium records. *Ecography*. 26:429–444.
- Glover B. 2007. *Understanding Flowers & Flowering, An Integrated Approach*, Oxford University Press, New York. 59p
- Godoy O, Richardson DM, Valladares F, Castro-Díez P. 2009. Flowering phenology of invasive alien plant species compared with native species in three Mediterranean-type ecosystems. *Ann Bot (Lond)*. 103(3):485-94.
- Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Syst Biol*. 49(4):652-70.
- Grant V. 1981. *Plant Speciation* (2nd ed.), Columbia Univ. Press, Durham, N.C. 563p.
- Greene B, Walko R, Hake S. 1994. Mutator insertions in an intron of the maize *knotted1* gene result in dominant suppressible mutations. *Genetics*. 138:1275–1285.
- Hanzawa Y, Money T, Bradley D. 2005. A single amino acid converts a repressor to an activator of flowering. *Proc Natl Acad Sci U S A*. 102(21):7748-53.
- Hawkins J, Proulx S, Rapp R, Wendel J. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc Natl Acad Sci U S A*. 106(42):17811-6.
- He Y. 2009. Control of the transition to flowering by chromatin modifications. *Mol Plant*. 2(4):554-64.
- Helliwell CA, Wood CC, Robertson M, James Peacock W, Dennis ES. 2006. The *Arabidopsis* FLC protein interacts directly in vivo with *SOC1* and *FT* chromatin and is part of a high-molecular-weight protein complex. *Plant J*. 46(2):183-92.
- Hepworth SR, Valverde F, Ravenscroft D, Mouradov A, Coupland G. 2002. Antagonistic regulation of flowering-time gene *SOC1* by *CONSTANS* and *FLC* via separate promoter motifs. *EMBO J*. 21(16):4327-37.
- Hijmans RJ, Guarino L, Cruz M. Rojas, E. 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter*. 127:15-

19.

Hijmans RJ, Spooner DM. 2001. Geographic distribution of wild potato species. *American Journal of Botany*. 88:2101-2112.

Hijmans RJ, Cameron SE, Parra JL, Jones PG and Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 25: 1965-1978.

Hodges SA, Arnold ML. 1994. Columbines: a geographically widespread species flock. *Proc Natl Acad Sci U S A*. 24;91(11):5129-32.

Holmgren PK, Holmgren NH. 1998. [continuously updated]. Index Herbarium: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. <http://sweetgum.nybg.org/ih/>

Hong X, Scofield D, Lynch M. 2006. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol*. 23(12):2392-404.

Huang T, Böhlenius H, Eriksson S, Parcy F, Nilsson O. 2005. The mRNA of the *Arabidopsis* gene *FT* moves from leaf to shoot apex and induces flowering. *Science*. 309(5741):1694-6.

Huang Y, Niu D. 2008. Evidence against the energetic cost hypothesis for the short introns in highly expressed genes. *BMC Evol Biol*. 8:154.

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

Hughes S, Buckley C, Neafsey D. 2008. Complex selection on intron size in *Cryptococcus neoformans*. *Mol Biol Evol*. 25(2):247-53.

Irimia M, Roy S. 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. *Nucleic Acids Res*. 36(5):1703-12.

Iseli C, Jongeneel C, Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*. 1999:138-48.

Jack T. 2004. Molecular and genetic mechanisms of floral control. *Plant Cell*. 16 Suppl:S1-17.

Jaeger KE, Wigge PA. 2007. FT protein acts as a long-range signal in *Arabidopsis*. *Curr Bio*. 19;17(12):1050-4.

Jaillon O, *et al*. 2007. The grapevine genome sequence suggests ancestral

hexaploidization in major angiosperm phyla. *Nature*. 449(7161):463-7.

Kardailsky I, Shukla VK, Ahn JH, Dagenais N, Christensen SK, Nguyen JT, Chory J, Harrison MJ, Weigel D. 1999. Activation tagging of the floral inducer FT. *Science*. 286:1962–1965.

Kearney M, Porter W. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecol Lett*. 12(4):334-50.

Keightley P, Gaffney D. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci. USA*. 100: 13402–13406.

Kejnovsky E, Hobza R, Cermak T, Kubat Z, Vyskot B. 2009. The role of repetitive DNA in structure and evolution of sex chromosomes in plants. *Heredity*. 102(6):533-41.

Kent W. 2002. BLAT--the BLAST-like alignment tool. *Genome Res*. 12(4):656-64.

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*. 29(2):170-9.

Knowles LL, Carstens BC, Keat ML. 2007. Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Curr Biol*. 17(11):940-6.

Kobayashi Y, Kaya H, Goto K, Iwabuchi M, Araki T. 1999. A pair of related genes with antagonistic roles in mediating flowering signals. *Science*. 286(5446):1960-2.

Kojima S, Takahashi Y, Kobayashi Y, Monna L, Sasaki T, Araki T, Yano M. 2002. *Hd3a*, a rice ortholog of the *Arabidopsis FT* gene, promotes transition to flowering downstream of *Hdl* under short-day conditions. *Plant Cell Physiol*. 43(10):1096-105.

Koonin EV. 2009. Darwinian evolution in the light of genomics. *Nucleic Acids Res*. 37(4):1011-34.

Koornneef M, Hanhart CJ, van der Veen JH. 1991. A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. *Mol Gen Genet*. 229(1):57-66.

Koornneef M, Alonso-Blanco C, Peeters AJ, Soppe W. 1998. Genetic control of flowering time in *Arabidopsis*. *Annu Rev Plant Physiol Plant Mol Biol*. 49:345-370.

Kumar A, Bennetzen J. 1999. Plant retrotransposons. *Annu Rev Genet*. 33:479-532.

Lanier W, Moustafa A, Bhattacharya D, Comeron J. 2008. EST analysis of *Ostreococcus lucimarinus*, the most compact eukaryotic genome, shows an excess of introns in highly expressed genes. *PLoS One*. 3(5):e2171.

- Lavoie C, Lachance D. 2006. A new herbarium-based method for reconstructing the phenology of plant species across large areas. *American Journal of Botany*. 93:512-516.
- Le Corre V. 2005. Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits. *Mol Ecol*. 14(13):4181-92.
- Lee JH, Cho YS, Yoon HS, Suh MC, Moon J, Lee I, Weigel D, Yun CH, Kim JK. 2005. Conservation and divergence of *FCA* function between *Arabidopsis* and rice. *Plant Mol Biol*. 58(6):823-38.
- Leimbeck RM, Valencia R, and Balslev H. 2004. Landscape diversity patterns and endemism of Araceae in Ecuador. *Biodiversity and Conservation*. 13:1755-1779.
- Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, Weigel D. 2005. Diversity of flowering responses in wild *Arabidopsis thaliana* strains. *PLoS Genet*. 1(1):109-18.
- Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 25: 1451-1452.
- Lienert J, Fischer M, Diemer M. 2002. Local extinctions of the wetland specialist *Swertia perennis* L. (Gentianaceae) in Switzerland: a revisitation study based on herbarium records. *Biol Conserv*. 103:65–76.
- Liu YJ, Hodson MC, Hall BD (2006) Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. *BMC Evol Biol*. 6:74.
- Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol*. 56: 504-514.
- Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol*. 53(1):320-8.
- Livingstone K, Anderson S. 2009. Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. *J Hered*. 100(6):754-61.
- Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci U S A*. [Epub ahead of print]
- Lu Y, Rausher MD. 2003. Evolutionary rate variation in anthocyanin pathway genes. *Mol Biol Evol*. 20(11):1844-53.
- Luehrsen K, Walbot V. 1992. Insertion of non-intron sequence into maize introns

interferes with splicing. *Nucleic Acids Res.* 20(19):5181-7.

Luo J, Yoshikawa N, Hodson MC, Hall BD. 2007. Duplication and paralog sorting of *RPB2* and *RPB1* genes in core eudicots. *Mol Phylogenet Evol.* 44(2):850-62.

Macknight R, et. al. 1997. *FCA*, a gene controlling flowering time in *Arabidopsis*, encodes a protein containing RNA-binding domains. *Cell.* 89(5):737-45.

Macknight R, Duroux M, Laurie R, Dijkwel P, Simpson G, Dean C. 2002. Functional significance of the alternative transcript processing of the *Arabidopsis* floral promoter *FCA*. *Plant Cell.* 2002 Apr;14(4):877-88.

MacDougall AS, Loo JA, Clayden SR, Goltz JG, Hinds HR. 1998. Defining conservation priorities for plant taxa in southeastern New Brunswick, Canada using herbarium records. *Biol Conserv.* 86:325–338.

Majewski J, Ott J, 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12:1827–1836.

Marais G, Nouvellet P, Keightley P, Charlesworth B. 2005. Intron size and exon evolution in *Drosophila*. *Genetics.* 170(1):481-5.

Mathieu J, Warthmann N, Küttner F, Schmid M. 2007. Export of FT protein from phloem companion cells is sufficient for floral induction in *Arabidopsis*. *Curr Biol.* 17(12):1055-60.

Messing J, Bennetzen J. 2008. Grass genome structure and evolution. *Genome Dyn.* 4:41-56.

Michaels SD, He Y, Scortecci KC, Amasino RM. 2003. Attenuation of *FLOWERING LOCUS C* activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. *Proc Natl Acad Sci U S A.* 100(17):10102-7.

Miller-Rushing AJ, Primack RB, Primack D and Mukunda S. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. *American Journal of Botany.* 93:1667-1674.

Moisy C, Garrison K, Meredith C, Pelsy F. 2008. Characterization of ten novel Ty1/copia-like retrotransposon families of the grapevine genome. *BMC Genomics.* 9:469.

Morello L, Breviario D. 2008. Plant spliceosomal introns: not only cut and paste. *Curr Genomics.* 9(4): 227–238.

Moore M. 1991. Classification and systematics of eastern North American *Vitis* L. north of Mexico. *Sida* 14: 810–815.

- Morjan CL, Rieseberg LH. 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol*. 13(6):1341-56.
- Munson TV. 1909. *Foundations of American grape culture*. T. V. Munson & Son, Denison, Texas.
- Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11(5):715-24.
- Naito K, Cho E, Yang G, Campbell M, Yano K, Okumoto Y, Tanisaka T, Wessler S. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci U S A*. 103(47):17620-5.
- National Association of American Wineries. 2007. *Grape Facts*.
- Nichols R. 2001. Gene trees and species trees are not the same. *Trends Ecol Evol*. 16(7):358-364.
- Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD. 2002. Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. *Genetics*. 160(4):1641-50.
- Oxelman B, Yoshikawa N, McConaughy BL, Luo J, Denton AL, Hall BD. 2004. *RPB2* gene phylogeny in flowering plants, with particular emphasis on asterids. *Mol Phylogenet Evol*. 32(2):462-79.
- Parkinson J, Guiliano D, Blaxter M. 2002. Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*. 3:31.
- Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M. 2004. PartiGene--constructing partial genomes. *Bioinformatics*. 20(9):1398-404.
- Pereira V, Enard D, Eyre-Walker A. 2009. The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One*. 4(2):e4321.
- Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*. 19(5):651-2.
- Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 21(5):676-9.
- Posada D and Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics*. 14 (9): 817-818.

- Presgraves DC. The molecular evolutionary basis of species formation. 2010. *Nat Rev Genet.* [Epub ahead of print]
- Primack D, Imbres C, Primack RB, Miller-Rushing AJ, and Del Tredici P. 2004. Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. *Amer. J. Bot.* 91:1260-1264.
- Putterill J, Robson F, Lee K, Simon R, Coupland G. 1995. The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell.* 80(6):847-57.
- Putterill J, Laurie R, Macknight R. 2004. It's time to flower: the genetic control of flowering time. *Bioessays.* 26(4):363-73.
- Quesada V, Macknight R, Dean C, Simpson GG. 2003. Autoregulation of *FCA* pre-mRNA processing controls *Arabidopsis* flowering time. *EMBO J.* 22(12):3142-52.
- R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rambaut A, Grassly, NC (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13: 235-238.
- Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. *Mol Biol Evol.* 26(5):1045-53.
- Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol.* 16(2):266-74.
- Rausher MD, Lu Y, Meyer K. 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. *J Mol Evol.* 67(2):137-44.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics.* 16(6):276-277.
- Ronquist F, Huelsenbeck, JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19:1572-1574.
- Sang, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology.* 37: 121–147.
- SanMiguel P, Tikhonov A, Jin Y, et al. (8 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science.* 274(5288):765-8.

SanMiguel P, Gaut B, Tikhonov A, Nakajima Y, Bennetzen J. 1998. The paleontology of intergene retrotransposons of maize. *Nature*. 20:43-45.

Scheldeman X, Willems L, Coppens d'Eeckenbrugge G, Romeijn-Peeters E, Restrepo MT, Romero Motoche J, Jiménez D, Lobo M, Medina CI, Reyes C, Rodríguez D, Ocampo JA, Van Damme P, Goetgebeur P. 2007. Distribution, diversity and environmental adaptation of highland pappas (Vasconcellea spp.) in tropical and subtropical America. *Biodiversity and Conservation*. 16:1867-1884.

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.

Shimodaira H, Hasegawa M (1999) Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution*. 16(8):1114-6.

Silvertown J, Servaes C, Biss P, Macleod D. 2006. Reinforcement of reproductive isolation between adjacent populations in the Park Grass Experiment. *Heredity*. 95(3):198-205.

Simpson GG, Dean C. 2002. *Arabidopsis*, the Rosetta stone of flowering time? *Science*. 296(5566):285-9.

Singh KP, Kushwaha CP. 2006. Diversity of flowering and fruiting phenology of trees in a tropical deciduous forest in India. *Ann Bot (Lond)*. 97(2):265-76.

Smit A, Hubley R, Green P. 2009. RepeatMasker at <http://repeatmasker.org>

Smith J, Putta S, Zhu W, Pao GM, Verma I, Hunter T, Bryant S, Gardiner D, Harkins T,

Voss S. 2009. Genic regions of a large salamander genome contain long introns and novel genes. *BMC Genomics*. 10:19.

Soltis PS, Soltis DE, Chase MW. 1999. angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*. 402(6760):402-4.

Soltis DE, Soltis PS, Endress PK, Chase MW. 2005. *Phylogeny and Evolution of Angiosperms*, Sinauer Associates, Sunderland, Massachusetts, 34p.

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

Swofford, D. L. 2002. PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer Associates, Sunderland, Massachusetts.

- Tam S, Lefebvre V, Palloix A, Sage-Palloix A, Mhiri C, Grandbastien M. 2009. LTR-retrotransposons Tnt1 and T135 markers reveal genetic diversity and evolutionary relationships of domesticated peppers. *Theor Appl Genet.* 119(6):973-89.
- Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18(12):1944-54.
- Templeton AR. 1994. The role of molecular genetics in speciation studies. *EXS.* 69:455-77.
- Templeton AR. 1998. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol Ecol.* 7(4):381-97.
- This P, Lacombe T, Cadle-Davidson M, Owens C. 2007. Wine grape (*Vitis vinifera* L.) color associates with allelic variation in the domestication gene *VvmybA1*. *Theor Appl Genet.* 114(4):723-30.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22:4673-4680.
- Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. 2009. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature.* 461(7262):423-6.
- Tuskan GA, et. al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science.* 313(5793):1596-604.
- Via S. 2009. Natural selection in action during speciation. *Proc Natl Acad Sci U S A.* 106 Suppl 1:9939-46.
- Vinogradov AE. 2010. Systemic factors dominate mammal protein evolution. *Proc Biol Sci.*[Epub ahead of print]
- Vitte C, Panaud O. 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Genome Res.* 110(1-4):91-107.
- Vitte C, Panaud O, Quesneville H. 2007. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics.* 8:218.
- Wang H, Liu J. 2008. LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. *BMC Genomics.* 9:382.
- Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. (2009) Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc Natl Acad Sci U S A.* 106(10):3853-8.

- Wang J, Li S, Zhang Y, Zheng H, Xu Z, Ye J, Yu J, Wong GK. 2003. Vertebrate gene predictions and the problem of large genes. *Nat Rev Genet.* 4(9):741-9.
- Walther GR, Post E, Convey P, Menzel A, Parmesan C, Beebee TJ, Fromentin JM, Hoegh-Guldberg O, Bairlein F. 2002. Ecological responses to recent climate change. *Nature.* 416: 389-395.
- Wawrzynski A, Ashfield T, Chen N, et al. (35 co-authors). 2008. Replication of nonautonomous retroelements in soybean appears to be both recent and common. *Plant Physiol.* 148(4):1760-71.
- Wendel J, Cronn R, Alvarez I, Liu B, Small R, Senchina D. 2002. Intron size and genome size in plants. *Mol Biol Evol.* (12):2346-52.
- Wendel J, Cronn R, Johnston J, Price H. 2002. Feast and famine in plant genomes. *Genetica.* 115(1):37-47.
- Wernersson R. 2005. FeatureExtract—extraction of sequence annotation made easy. *Nucleic Acids Res.* 33(Web Server issue): W567–W569.
- Wessler S, Bureaua, T and Whitea, S. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr Opin Genet Dev.* 5(6):814-21.
- Whittall JB, Hodges SA. 2007. Pollinator shifts drive increasingly long nectar spurs in columbine flowers. *Nature.* 447(7145):706-9.
- Wolfe DW, Schwartz MD, Lasko AN, Otsuki Y, Pool RM, Shaulis NJ. 2005. Climate Change and Shifts in Spring Phenology of Three Horticultural Woody Perennials in Northeastern USA. *International Journal of Biometeorology.* 49: 303-309.
- Wong G, Passey D, Yu J. 2001. Most of the human genome is transcribed. *Genome Res.* 11(12):1975-7.
- Woolfe A, Goodson M, Goode DK, et al. (13 co-authors). 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3(1):e7.
- Wu CI, Ting CT. 2004. Genes and speciation. *Nat Rev Genet.* 5, 144-122.
- Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35(Web Server issue):W265-8.
- Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J. 2006. The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc Natl Acad Sci U S A.* 103(51):19581-6.
- Yang YH, Zhang FM, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway

genes. *BMC Evol Biol.* 9:206.

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

Zohary D, Spiegel-Roy P. 1975. Beginnings of fruit growing in the old world. *Science.* 187:319-327.

Appendices

Appendices Table 1 *RPB2* sequence data of D- and I-copies

Taxa	GenBank #	Taxa	GenBank #	Taxa	GenBank #	Taxa	GenBank #
<i>Acer</i>	AY566621	<i>Corydalis_2</i>	DQ017098	<i>Hibbertia</i>	DQ017113	<i>Parthenocissus_D*</i>	GU947852
<i>Aextoxicon</i>	DQ017087	<i>Cyphostemma_D*</i>	GU947853	<i>Homalocladium</i>	DQ017114	<i>Parthenocissus_I*</i>	GU947851
<i>Akebia</i>	AY566614	<i>Cyphostemma_I*</i>	GU947849	<i>Hydrastis</i>	DQ017115	<i>Pereskia</i>	DQ017124
<i>Alangium</i>	AJ555916	<i>Dianthus</i>	DQ017099	<i>Ilex_D</i>	AJ557240/2	<i>Petunia_D</i>	DQ020641
<i>Antirrhinum_D</i>	DQ020642	<i>Diapensia_D</i>	AY579382	<i>Ilex_I</i>	AJ557241/3	<i>Petunia_I</i>	DQ020638
<i>Antirrhinum_I</i>	DQ020637	<i>Diapensia_I</i>	AY579381	<i>Leea_I*</i>	GU947850	<i>Platanus</i>	DQ058632
<i>Arabidopsis</i>	Z19120	<i>Dicentra</i>	DQ017100	<i>Linanthus</i>	DQ058637	<i>Psilotum</i>	DQ029108
<i>Argemone_I</i>	DQ017088	<i>Dillenia</i>	DQ017101	<i>Lindera</i>	DQ017116	<i>Pteridophyllum</i>	DQ017125
<i>Argemone_2</i>	DQ017089	<i>Dioscorea</i>	AY563268	<i>Liquidambar</i>	DQ058633	<i>Rhododendron_D</i>	DQ058627
<i>Armeria_I</i>	DQ017085	<i>Drosera</i>	DQ017102	<i>Liriodendron</i>	DQ058631	<i>Rhododendron_I</i>	DQ058628
<i>Armeria_2</i>	DQ017090	<i>Embothrium</i>	DQ017103	<i>Lonicera</i>	AJ565933	<i>Rivina</i>	DQ017127
<i>Astilbe</i>	DQ058629	<i>Escallonia_D</i>	AJ565858	<i>Lycopersicon_D</i>	U28403	<i>Sarcococca</i>	DQ017130
<i>Berberidopsis</i>	DQ020634	<i>Escallonia_I</i>	AJ557244	<i>Lycopersicon_I</i>	DQ020639	<i>Sarracenia</i>	DQ017131
<i>Buxus</i>	DQ017091	<i>Eschscholzia</i>	DQ058630	<i>Magnolia</i>	AF020841	<i>Selaginella</i>	DQ029107
<i>Callistemon</i>	DQ017092	<i>Euptelea</i>	DQ017105	<i>Mahonia</i>	DQ017117	<i>Sphagnum</i>	DQ029106
<i>Camellia_D</i>	DQ058625	<i>Gardenia_D</i>	AJ566359	<i>Meconopsis</i>	DQ017118	<i>Spinacia</i>	AF020840
<i>Camellia_I</i>	DQ058626	<i>Gardenia_I</i>	AJ558243	<i>Mimulus_D</i>	AJ565937	<i>Spirogyra</i>	DQ029103
<i>Chelidonium_1</i>	DQ017094	<i>Garrya</i>	AJ563602	<i>Mimulus_I</i>	AJ558241/2	<i>Tephrosia_I</i>	AJ556782
<i>Chelidonium_2</i>	DQ017095	<i>Ginkgo</i>	AF020843	<i>Myrtus</i>	AJ556164	<i>Tephrosia_2</i>	AJ557000/1
<i>Chimonanthus</i>	DQ017093	<i>Glaucium_1</i>	DQ017106	<i>Nelumbo</i>	DQ017120	<i>Tetracentron</i>	DQ017132
<i>Chloranthus</i>	AF041852	<i>Glaucium_2</i>	DQ017107	<i>Nepenthes</i>	DQ017121	<i>Trochodendron</i>	AY563269
<i>Cissus_I</i>	DQ017096	<i>Gunnera</i>	DQ017108	<i>Nicotiana_D</i>	DQ020640	<i>Vitis_D*</i>	GU947854
<i>Coleochaete</i>	DQ029105	<i>Hedera</i>	AJ563601	<i>Nicotiana_I</i>	DQ020636	<i>Vitis</i>	AJ556992
<i>Coriaria</i>	DQ017097	<i>Helleborus_1</i>	DQ017109	<i>Nymphaea</i>	AF043427		
<i>Corydalis_1</i>	DQ017086	<i>Helleborus_2</i>	DQ017110	<i>Papaver</i>	DQ017122		

* characterized in this study, all others from Luo *et al.* 2007 and Oxelman *et al.* 2004

Appendices Table 2 Herbarium data sources

Institution	Herbarium Code†	Data Source	No. of records
Academy of Natural Sciences	PH	images	63
Arizona State University	ASU	digital, (Southwest Environmental Information Network)	107
Auburn University	AUA	digital	154
Brooklyn Botanical Garden	BKL	download	332
California State University-Chico	CHSC	download, (Consortium of California Herbaria)	31
Carnegie Museum of Natural History	CM	e-mail request	796
Cincinnati Zoo	CZ*	download	72
Colorado State University	CS	download	16
Eastern New Mexico University	ENMU*	download, (New Mexico Biodiversity Collections Consortium)	8
Fairchild Tropical Garden	FTG	download, online herbarium	120
Field Museum of Natural History	F	download (Illinois Natural History Survey)	97
Florida State University	FSU	download, online herbarium	86
Harvard University	GH	download, (GBIF)	24
Illinois Natural History Survey	ILLS	download	1216
Indiana University Southeast	JEF	download	5
Kansas State University	KSC	download	280
Louisiana State University	LSU	loaned	166
Michigan State University	MSC	e-mail request	166
Mississippi State University	MISS	e-mail request	297
Missouri Botanical Garden	MO	download	767
Morton Arboretum	MOR	download, (Wisconsin Botanical Information System)	5
Murray State University	MUR	download	105
Nationaal Herbarium Nederland-Utrecht	U	download, (GBIF)	2
New Mexico State University (Biology)	NMC	download, (New Mexico Biodiversity Collections Consortium)	16
New Mexico State University (Range Sci.)	NMCR	download, (New Mexico Biodiversity Collections Consortium)	8
New York Botanical Garden	NY	download	21
North Kentucky University	KNK	e-mail request	13
Northern Arizona University	ASC	download, (Southwest Environmental Information Network)	54
Oregon State University	OSC+WIL LU	download, (Oregon Vascular Plant Database)	25
Pomona College	POM	download, (Consortium of California Herbaria)	1
Rancho Santa Ana Botanical Garden	RSA	images and download, (Consortium of California Herbaria)	132
Smithsonian Institution	US	download (D.C. flora project)	283
Appendices Table 2 (continued)	AMU	download	95
(to be continued)	NA	vouchers collected & deposited	257
University of Alabama	UNA	download, (GBIF)	75

University of Arizona	ARIZ	download, (Southwest Environmental Information Network)	108
University of British Columbia	UBC	download	4
University of Colorado-Boulder	COLO	download	31
University of Kansas	KANU	e-mail request	357
University of Michigan	MICH	download	55
University of Minnesota	MIN	download	145
University of Missouri	UMO	download	81
University of New Mexico	UNM	download, (New Mexico Biodiversity Collections Consortium)	94
University of North Carolina	NCU	e-mail request	479
University of Oregon	ORE	download, (Oregon Vascular Plant Database)	19
University of Tennessee-Knoxville	TENN	download, (GBIF)	35
University of Texas-Austin	TEX+LL	download	425
University of Washington	WTU	download, (GBIF)	33
University of Wisconsin	WIS	download, (Wisconsin Botanical Information System)	411
University of Wisconsin-Plattville	UWP*	download, (Wisconsin Botanical Information System)	16
University of Wisconsin-Green Bay	UWGB	download, (Wisconsin Botanical Information System)	35
University of Wisconsin-Stevens Point	UWSP	download, (Wisconsin Botanical Information System)	58
University of Wisconsin-Superior	SUWS	download, (Wisconsin Botanical Information System)	3
Utah Valley State College	UVSC	download	4
Virginia Polytechnic Institute	VPI	e-mail request	77
Western New Mexico University	SNM	download, (New Mexico Biodiversity Collections Consortium)	13

† Holmgren, P. K., and N. H. Holmgren. 1998 [continuously updated]. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. <http://sweetgum.nybg.org/ih/>

* Not an official Index Herbariorum code

Appendices Table 3 Data fields of the specimen data set.

Data field name	Data field description
ID	Unique identification number
herbarium	Index Herbarium code (see supplementary information)
herb_no	Herbarium accession number, if provided
genus	genus to which the specimen belongs (all have value 'Vitis')
species	Species to which the specimen belongs
var/ssp	Varieties or subspecies to which the specimen belongs
current	Current scientific name to account for nomenclature and synonyms
country	Country of origin of the specimen
state/dist	State or district of origin of the specimen
county	County of origin of the specimen
coll_location	Description of the collecting locations
comments	Comments on the specimen (primarily habitat information)
latitude	Latitude of the location where the specimen was collected
longitude	Longitude of the location where the specimen was collected
geosource	Origin of coordinates (from label or inferred, see 'georeferencing' in methods)
year	Year of the collection date
month	Month of the collection date
day	Day of the collection date
reprostate	Reproductive state of the specimen (flower or not)
ordinal_day	Ordinal day in the year of the collection date

Appendices Table 4 Bioclimatic variables derived from temperature and rainfall

Name of variable	Meaning of variable
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (P2/P7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (P5-P6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

Appendices Table 5 All primers used in *Vitis* flowering time gene characterization

Name	Locus	Region	Sequences
AP1_X1F1	<i>AP1</i>	exon 1	5'-GGTCTTGTGTGATGCAGAGGT-3'
AP1_X2R1	<i>AP1</i>	exon 2	5'-GATTTCAGGATCAGTTGCAGTCA-3'
COL-3UR1	<i>CO</i>	3'-UTR	5'-GGAATGAGTAATGAGAAGCTGAGTTTCAA-3'
COL-5UF1	<i>CO</i>	5'-UTR	5'-ATGACATGCACATTAAATGATTCCA-3'
FCA-3UR1	<i>FCA</i>	3'-UTR	5'-CCAAGCATCTTTGACCATCC-3'
FCA-5UF1	<i>FCA</i>	5'-UTR	5'-AGTCGCAGTTCCAAAACGAT-3'
FCA-E1F1	<i>FCA</i>	exon 1	5'-CTACGGCAACAACCCTGACT-3'
FCA-E3R1	<i>FCA</i>	exon 3	5'-CATCAATACCCATTCAAAAGCA-3'
FCA-X13R1	<i>FCA</i>	exon 13	5'-GAACAGTTAATGGCTGATCACAAC-3'
FCA-X13R2	<i>FCA</i>	exon 13	5'-GGATCAGCAAACCGAACAGT-3'
FCA-X3F1	<i>FCA</i>	exon 3	5'-CAAACCTGAAGAAGGCAAGACCT-3'
FCA-X3F2	<i>FCA</i>	exon 3	5'-GAGAAGCTGGACAACAGCAAGG-3'
FCA-X4F1	<i>FCA</i>	exon 4	5'-CATTGGGCTCTATTTACGCATTAG-3'
FCA-X4R1	<i>FCA</i>	exon 4	5'-CTAATGCGTAAATAGAGCCCAATG-3'
FCA-X4R2	<i>FCA</i>	exon 4	5'-GAGAAGGGATGGCTCACATT-3'
FCA-X8F1	<i>FCA</i>	exon 8	5'-GAAGCTGAAAGGGCAATTAGAG-3'
FLC_X1F2	<i>FLC</i>	exon 1	5'-ACTGAAGCGAATCGAAGACAAG-3'
FLC-I4R1	<i>FLC</i>	intron 4	5'-GCCTCATTTGTTCTTCCTGTGTT-3'
FLC-I4R2	<i>FLC</i>	intron 4	5'-GGGCTAGTGGGAAAATTGAT-3'
FLC-X1F1	<i>FLC</i>	exon 1	5'-CATTTTGAAGCAGAAGGCAAT-3'
FLC-X1F2	<i>FLC</i>	exon 1	5'-GCTCTCTGTTCTCTGCGATGT-3'
FLC-X2R1	<i>FLC</i>	exon 2	5'-GCCTTCTGCTTCAAAATGACTC-3'
FLC-X3R1	<i>FLC</i>	exon 3	5'-CAAGGTCAGTCACGCTCATC-3'
FLC-X4F1	<i>FLC</i>	exon 4	5'-GATGAGCGTGACTGACCTTGT-3'
FLC-X5R1	<i>FLC</i>	exon 5	5'-CTTTTCATGGAGATGCCTTATG-3'
FLC-X6R1	<i>FLC</i>	exon 6	5'-CCGTTGAGACACCAAACACTT-3'
FRI-5UF1	<i>FRI</i>	5'-UTR	5'-GGATTTTAGGGCTTAGGAGGAA-3'
FRI-5UF2	<i>FRI</i>	5'-UTR	5'-GAAGGGTTTTGGGGGTTTTA-3'
FRI-E1F1	<i>FRI</i>	exon 1	5'-CAAAACCTGCCAAACTTGTACTT-3'
FRI-E1F2	<i>FRI</i>	exon 2	5'-CTGCCAAACTTGTACTTGAATGC-3'
FRI-E1F2	<i>FRI</i>	exon 1	5'-CCTGCCAAACTTGTACTTGAATG-3'
FRI-E2R2	<i>FRI</i>	exon 2	5'-CGCAAGAACGATGTCAAAAT-3'
FRI-E3R1	<i>FRI</i>	exon 3	5'-GAGAGGAGGAAAACCTCAAAAGGA-3'
FRI-I1R1	<i>FRI</i>	intron 1	5'-GTTTAGCATCCCGGAAGGA-3'
FT-3UR1	<i>FT</i>	3'-UTR	5'-CACCAAGTTGCCTATTCAAGC-3'
FT-3UR2	<i>FT</i>	3'-UTR	5'-CCAAGTTGCCTATTCAAGCATAA-3'
FT-5UF1	<i>FT</i>	5'-UTR	5'-CAGTTTGGATAGCTCCCTCTTG-3'
FT-5UF2	<i>FT</i>	5'-UTR	5'-CTCTTGTATTGTATCGGTGAGGTG-3'
FT-5UF3	<i>FT</i>	5'-UTR	5'-GCAGATAGCACCGGACTAGTTAT-3'
FT-E1F1	<i>FT</i>	exon 1	5'-GAGAAGTAGCAAATGGCTGTGA-3'
FT-E4R1	<i>FT</i>	exon 4	5'-GGATTATGCTTCACCCCATAGA-3'
TFL1-3UR1	<i>TFL1</i>	3'-UTR	5'-TGATCTTCCCGTTGGTTATTG-3'
TFL1-5UF1	<i>TFL1</i>	5'-UTR	5'-GCCTTCCAAGAGACCAAGAGT-3'
TFL1-E1F1	<i>TFL1</i>	exon 1	5'-GTGATTGGGGATGTTGTTGA-3'
TFL1-E4R1	<i>TFL1</i>	exon 4	5'-GCAGCTACAGGAGGACCAAG-3'

Appendices Table 6 Taxa sampling and GenBank accession numbers for analysis of evolution rates of flowering time genes

		<i>API</i>		<i>CO</i>
Amborellaceae	<i>Amborella</i>	FD435144	<i>Amborella</i>	FD429972
Nymphaeaceae	<i>Nuphar</i>	ES732871		
Magnoliales	<i>Magnolia</i>	AY821777	<i>Magnolia</i>	GQ489239
Laurales	<i>Persea</i>	DQ398019		
Piperales	<i>Aristolochia</i>	FD759109	<i>Saruma</i>	DT592235
Alismatales				
Dioscoreales				
Liliales	<i>Lilium</i>	GT229480		
Asparagales	<i>Asparagus</i>	DY032502	<i>Allium</i>	GQ232751
Arecales	<i>Elaeis</i>	EL690590	<i>Elaeis</i>	ES324026
Poales	<i>Zea</i>	NM_001111863*	<i>Zea</i>	EU098140
Zingiberales	<i>Zingiber</i>	DY356542	<i>Musa</i>	DQ153049
Ranunculales	<i>Aquilegia</i>	DT738214	<i>Eschscholzia</i>	CD479828
Proteales				
Trochodendrales				
Gunnerales				
Berberidopsidales				
Caryophyllales	<i>Mesembryanthemum</i>	BE034098	<i>Chenopodium</i>	EU395770
Saxifragales	<i>Corylopsis</i>	AY306146		
				XM_0022824
Vitales	<i>Vitis</i>	GU133634	<i>Vitis</i>	73*
Myrtales			<i>Eucalyptus</i>	ES591039
				XM_0025328
Malpighiales	<i>Populus</i>	XM_002311317*	<i>Ricinus</i>	40*
Fabales	<i>Lotus</i>	AY770395	<i>Glycine</i>	DQ371243
Rosales	<i>Malus</i>	AY071921	<i>Fragaria</i>	FJ377616
Cucurbitales				
Fagales	<i>Betula</i>	X99653		
Brassicales	<i>Arabidopsis</i>	Z16421	<i>Arabidopsis</i>	AY086574
Malvales	<i>Gossypium</i>	ES850914		
Sapindales	<i>Citrus</i>	AY338974	<i>Mangifera</i>	FJ719767
Cornales				
Ericales	<i>Vaccinium</i>	CV090688		
Garryales				
Gentianales	<i>Coffea</i>	GT006211	<i>Coffea</i>	DV692807
Lamiales	<i>Syringa</i>	AY306185	<i>Olea</i>	EU860367
Solanales	<i>Ipomoea</i>	AB302848	<i>Ipomoea</i>	AF300700
Aquifoliales				
Asterales	<i>Helianthus</i>	DY942802		
Dipsacales				
Apiales	<i>Daucus</i>	AJ271147	<i>Panax</i>	DV554944

(to be continued)

Appendices Table 6 (continued)

		FLC		FT
Amborellaceae	<i>Amborella</i>	AY936234		
Nymphaeaceae	<i>Nuphar</i>	ES732871		
Magnoliales				
Laurales	<i>Chimonanthus</i>	DW223080		
Piperales	<i>Aristolochia</i>	FD761148		
Alismatales			<i>Lemna</i>	AY803292
Dioscoreales				
Liliales				
Asparagales	<i>Asparagus</i>	CV291850	<i>Oncidium</i>	EU583502
Arecales	<i>Elaeis</i>	EL930344		
Poales	<i>Zea</i>	DV535940	<i>Oryza</i>	NM_001063395*
Zingiberales				
Ranunculales	<i>Aquilegia</i>	DR933608		
Proteales			<i>Platanus</i>	GQ847823
Trochodendrales				
Gunnerales				
Berberidopsidales				
Caryophyllales	<i>Beta</i>	DQ189210	<i>Chenopodium</i>	EU128013
Saxifragales				
Vitales	<i>Vitis</i>	XM_002276105*	<i>Vitis</i>	DQ504308
Myrtales				
Malpighiales	<i>Euphorbia</i>	DV136637	<i>Populus</i>	AY515152
Fabales	<i>Glycine</i>	DB964933	<i>Medicago</i>	GU065342
Rosales	<i>Taihangia</i>	EF469601	<i>Malus</i>	AB161112
Cucurbitales			<i>Cucurbita</i>	DQ865290
Fagales				
Brassicales	<i>Arabidopsis</i>	BT030637	<i>Arabidopsis</i>	NM_105222*
Malvales	<i>Theobroma</i>	CU608380	<i>Gossypium</i>	ES826802
Sapindales	<i>Citrus</i>	EU497676	<i>Citrus</i>	AB027456
Cornales				
Ericales	<i>Actinidia</i>	FG485739		
Garryales				
Gentianales				
Lamiales	<i>Torenia</i>	AB359957	<i>Olea</i>	EU860369
Solanales	<i>Petunia</i>	AY370529	<i>Ipomoea</i>	EU178859
Aquifoliales				
Asterales	<i>Cichorium</i>	EH695107	<i>Chrysanthemum</i>	GQ925916
Dipsacales				
Apiales				

(to be continued)

Appendices Table 6 (continued)

<i>RPB2</i>			<i>TFL1</i>	
Amborellaceae				
Nymphaeaceae	<i>Nymphaea</i>	AF043427		
Magnoliales	<i>Magnolia</i>	AF020841		
Laurales				
Piperales				
Alismatales				
Dioscoreales	<i>Dioscorea</i>	AY563268		
Liliales				
Asparagales				
Arecales				
Poales			<i>Lolium</i>	AF316419
Zingiberales				
Ranunculales	<i>Akebia</i>	AY566614	<i>Aquilegia</i>	DQ286962
Proteales	<i>Nelumbo</i>	DQ017120		
Trochodendrales	<i>Tetracentron</i>	DQ017132		
Gunnerales	<i>Gunnera</i>	DQ094142		
Berberidopsidales	<i>Berberidopsis</i>	DQ020634		
Caryophyllales	<i>Spinacia</i>	AF020840	<i>Beta</i>	DQ849290
Saxifragales	<i>Astilbe</i>	DQ058629		
Vitales	<i>Vitis</i>	AJ556992	<i>Vitis</i>	XM_002276784*
Myrtales	<i>Myrtus</i>	AJ556164	<i>Eucalyptus</i>	EU573981
Malpighiales			<i>Ricinus</i>	XM_002528685*
Fabales			<i>Glycine</i>	EU912425
Rosales			<i>Cydonia</i>	AB162043
Cucurbitales			<i>Cucumis</i>	AB383154
Fagales				
Brassicales	<i>Arabidopsis</i>	Z19120	<i>Arabidopsis</i>	BT024828
Malvales			<i>Gossypium</i>	EU026443
Sapindales	<i>Acer</i>	AY566621	<i>Citrus</i>	EU400600
Cornales	<i>Alangium</i>	AJ555916		
Ericales	<i>Camellia</i>	DQ058626	<i>Impatiens</i>	AJ888756
Garryales	<i>Garrya</i>	AJ563602		
Gentianales	<i>Gardenia</i>	AJ558243	<i>Coffea</i>	DV679580
Lamiales	<i>Antirrhinum</i>	DQ020637		
Solanales	<i>Solanum</i>	DQ020639	<i>Solanum</i>	U84140
Aquifoliales	<i>Ilex</i>	AJ557241		
Asterales			<i>Helianthus</i>	DY918510
Dipsacales	<i>Lonicera</i>	AJ565933		
Apiales	<i>Hedera</i>	AJ563601		

* annotated genomic sequences

Appendices Table 7 Primers for characterizing selected introns in *Vitis* spp and varieties

Name	Locus	Region	Sequence (5'-end to 3'-end)
45_X2F	GSVIVT00027725001	exon 2	GATAAAGAGCTTAACAAAGATGTTGC
45_X3R	GSVIVT00027725001	exon 3	CTTCTGAATGTCCTCAATTTTCATCT
45_I2F	GSVIVT00027725001	intron 2	CTTCAGCCACATATTCTCAAACC
45_I2R	GSVIVT00027725001	intorn 2	CAATTGATCACATTCTTCTTCATTG
59_X4F	GSVIVT00033984001	exon 4	GAGGATGCCAAATATGTGATAAATG
59_X5R	GSVIVT00033984001	exon 5	CATGCCTTCTTCTATATCAGTTGGA
59_I4F	GSVIVT00033984001	intron 4	CTCTATCTCTACAGCAGGACTCCAC
59_I4R	GSVIVT00033984001	intron 4	AGTTATGGAGTAGGATTTTGGAAGG
140_X1F	GSVIVT0003715001	exon 1	ATCCCAGAGGAGCTACAAATG
140_X2R	GSVIVT0003715001	exon 2	CACATATAAGTTGAGCCATTTCCA
140_I1F	GSVIVT0003715001	intron 1	CTTTGCTTTTAGTCTTTTGGAGTG
140_I2R	GSVIVT0003715001	intron 1	ATATACTCCTGTGAAAATCCAGACG
148_X1F	GSVIVT00022278001	exon 1	GCAACCTTGATGCCCAGATTTCTCAG
149_X2R	GSVIVT00022278001	exon 2	CCTCCATTAGTATCTCCTTCGC
148_I1F	GSVIVT00022278001	intron 1	GCTAACAGATGTTTTATGTGCTGTG
148_I1R	GSVIVT00022278001	intron 1	CTCTATCTCTACAGCAGGACTCCAC