The Pennsylvania State University

The Graduate School

Eberly College of Science

EVOLUTION OF TWO GENE FAMILIES CONTROLLING THE FUNDAMENTAL PROCESSES OF EUKARYOTIC DEVELOPMENT: MADS-BOX GENE FAMILY AND HOMEOBOX GENE FAMILY

A Thesis in

Biology

by

Jongmin Nam

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2005

The thesis of Jongmin Nam was reviewed and approved* by the following:

Masatoshi Nei
Evan Pugh Professor of biology
Thesis Advisor
Chair of Committee

Hong Ma
Professor of Biology

Claude W. dePamphilis
Associate Professor of Biology

David M. Geiser
Assistant Professor of Plant Pathology

Douglas R. Cavener
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School.

ABSTRACT

MADS-box and homeobox genes form two large gene families of transcription factors that control fundamental processes of eukaryotic development. Therefore, studies of evolution of these gene families are expected to give some insights into the evolution of morphological characters. With this in mind, the following four related studies were conducted. (1) The origin and diversification of MADS-box genes controlling flowering in plants were studied. The results suggested that the ancestor of floral MADS-box genes diverged from the group of MADS-box genes controlling development of vegetative tissues about 650 MYA. The results also suggested that several classes of floral MADS-box genes already existed around the time of the Cambrian explosion, which is much earlier than the time of the origin of flowering plants. (2) The patterns of duplication and loss of MADS-box genes in the genomes of *Arabidopsis* and rice were studied. It was shown that type I MADS-box genes which were poorly characterized biochemically and functionally have experienced a higher rate of birth-and-death evolution than the type II MADS-box genes. Further analyses suggested that segmental duplication occurred more frequently in type I genes than in type II genes, indicating a higher rate of birth-and-death evolution in type I genes than in type II genes. (3) A simple statistical method for predicting the gene regions that are functionally differentiated between duplicate genes was developed. This method is to compare the evolutionary rates of two genes using sliding window analysis and identify the regions (or windows) of significant rate differences as possible candidates for functional differentiation. This method was applied to 23 pairs of closely related MIKC-type MADS-box genes from petunia. Of the 23 pairs, 14 pairs showed a significant rate difference. Interestingly, most of the predicted regions

were within the K-domain, which is important for the dimerization of MADS-box proteins. These predicted regions may be chosen for further experimental studies. (4) The pattern of duplication and loss of homeobox genes in the 11 species of bilateral animals was studied. There are more than 200 homeobox genes in the genome of vertebrates and about 100 homeobox genes in the genome of invertebrate studied. On the basis of a phylogenetic analysis, it was estimated that there were at least 88 homeobox genes in the MRCA of the 11 species. Of the 88 genes, similar numbers (≈ 50 ~ 60) of homeobox genes left at least one descendents (or survived) in the genome of each species, and many genes have been lost in a lineage-specific manner. The genes that survived in each evolutionary lineage experienced frequent gene duplication events, and some of duplicate genes were lost much later. Because the gene losses studied here are losses of ancient duplicate genes, these gene losses are very likely to have caused the evolutionary changes of morphological and/or physiological characters. These results suggested that gene loss might also have contributed to phenotypic evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply indebted to my thesis advisor Dr. Masatoshi Nei for his encouragements, supports, and criticisms during my graduate study. In addition, he saved me from many troubles and gave me insightful advices whenever I was struggling with deadlock questions (e.g., the biological significance of statistical significance). I would also like to thank my committee members, Drs. Hong Ma, Claude dePamphilis, and David Geiser, for their encouragements and advices, both humanly and scientifically. Drs. Ma and dePamphilis also guided me to finish the MADS-box gene works. I have been very lucky to learn from them ever since I came to Penn State.

Some of former and current members of the Nei lab should not be missed in this acknowledgment. They are Takeshi Itoh, Yoshiyuki Suzuki, Galina Glazko, Kazuharu Misawa, Helen Piontkivska, Yoshihito Niimura, Nikolas Nikolaidis, Li Hao, Joyce White, and Kathleen Seasholtz. I think that we were not just lab members but family members. I also thank a former visiting student in the Nei lab, Kerstin Kaufmann, and her advisor, Dr. Günter Theiβen, for enjoyable collaborations. And also many thanks go to several members in the Biology department for their advices and supports. They are Dr. Jan Klein, Dr. Wojciech Makalowski, Dr. Hongzhi Kong, Dr. James Leebens-Mack, Dr. Laura Zahn, Ms. Kathleen McClintock, Ms. Paula Farwell, Ms. Dianne Burpee, Dr. Esther Siegfried, and Dr. Carla Hass. I also thank my fellow graduate students, especially Wen-Ya Ko and Jamie Blair, for their friendship and discussions. I am also indebted to many Korean students in Penn State, especially Yong-Shik Kim, Young-Hoon Kim, Byong-Hoon Jeon, Dong-Joon Cho, Siho Nam, and Seok-Ki Um. Without their help, my life in State College could have been very tough.

I also thank my Masters advisor, Dr. Gynheung An at POSTECH, for his continuous encouragements and supports. Some of former and current members of the An lab also should be appreciated for continuing and inspiring communications. They are Dr. Jong-Seong Jeon, Dr. Soon-Kee Sung, Dr. Hong-Gyu Kang, Joonyul Kim, Sichul Lee, Dong-Hoon Jeong, Jinwon Lee, and Shinyoung Lee. The Rotary Club of the District 3630 also should be appreciated for the Rotary Scholarship and encouragements. A special thank goes to the members of the Euisong Rotary Club for their strong support and encouragements.

Finally, I wish to thank my wife, Heeyoung Park (a.k.a. Baqui), for her love, support, patience, and listening to many of my junk ideas. 멀리 고국에서도 항상 자식들이 탈없이 잘 지내길 빌어주시는 부모님들께도 여러 모로 감사드립니다.

CHAPTER 1

INTRODUCTION

Every organism has its unique body structure and, at the same time, shares similar structures with other organisms. The search for the factors that are responsible for such morphological uniqueness or similarity has been one of fundamental questions in biology. Recent progress in developmental biology and evolutionary biology contributed significantly to our understanding on the molecular mechanisms that are responsible for the uniqueness or similarity (Carroll, Grenier, and Weatherbee 2001; Cronk, Bateman, and Hawkins 2002). General insights based on the progress may be summarized as follows. (1) Regulatory genes encoding transcription factors and signaling proteins play major roles in development. (2) Regulatory genes that are involved in developmental processes are often shared by distantly related organisms, and these conserved regulatory genes are largely responsible for establishing and maintaining similar morphological characters among different organisms. (3) Duplication and subsequent functional differentiation of regulatory genes are often correlated with the evolution of new morphological characters. Therefore, it is of great interest to understand the evolution of regulatory genes that control developmental processes.

For the past 15 ~ 20 years, two large gene families of transcription factors, MADS-box genes and homeobox genes have been intensively characterized because of their important and various roles during development. The former in plants was originally identified as genes controlling the development of floral organs in Arabidopsis and Antirrhinum (Sommer et al. 1990; Yanofsky et al. 1990). It has later been shown that other members of MADS-box gene family are involved in various aspects of developmental processes such as activation of flowering, repression of flowering, root development, fruit development, etc (reviewed in Theissen et al. 2000). The latter in animals was identified as genes responsible for controlling the

development of body structure in fruit flies (Garber, Kuroiwa, and Gehring 1983; Scott et al. 1983). Subsequent studies showed that different members of the homeobox gene family play important roles during the development of eyes, neurons, limbs, axis formation, and other processes (reviewed in Duboule 1994). Further comparative functional analyses of MADS-box genes and homeobox genes showed that closely related genes often have similar or at least related function. Therefore, the two gene families are good model systems for studying the evolution of regulatory genes and the related developmental mechanisms.

In recent years, owing to the advances in sequencing technologies and genomics, the numbers of MADS-box genes and homeobox genes identified from various taxa have increased substantially. There are also complete or almost complete genome sequences of two plants species and more than ten animal species, and therefore almost complete inventories of genes in these genomes are known. These advances provided us a good opportunity for studying the evolution of the two gene families in various aspects.

I have therefore conducted the following four related studies to gain insights into the evolution of the MADS-box genes and homeobox genes and also to gain general insights concerned with the evolution of developmental processes. The four subjects are as follows: (1) Origin and divergence of the MADS-box genes controlling the development of flowers in plants; (2) The patterns of duplication and loss of different types of MADS-box genes in the genomes of Arabidopsis and rice; (3) Development of a method for predicting functionally differentiated regions between duplicate genes and its application to the MADS-box genes; (4) Repertoire of the homeobox genes and the pattern of duplication and loss of homeobox genes in bilateral animals.

Chapter 2 is concerned with the evolution of floral MADS-box genes that are important for flower development in plants. Although it has been suggested that the origin of floral MADS-box genes predates the origin of flowers (Purugganan 1997), the order and time of the divergence of different classes of floral MADS-box genes that are concerned with different aspects of flower development have not been studied intensively. The previous work was also needed an update, because we now have much more data than before. In this study I defined the order of the divergence of different classes floral MADS-box genes using a rooted tree. The results showed that the ancestor of floral MADS-box genes diverged from a group of genes controlling the development of vegetative tissues about 650 MYA. I also estimated the time of the divergence of different classes of floral MADS-box genes using molecular clock estimation and found that several classes of floral MADS-box genes was already existed at the time of the Cambrian explosion which is much earlier than the origin of flowers.

Chapter 3 is about a comparison of the patterns of duplication and loss of different types of MADS-box genes in angiosperms. It has been proposed that there were at least two types of MADS-box genes (type I and type II genes) before the animal-plant split (Alvarez-Buylla et al. 2000b). It has also been known that there are about 60 type I genes and about 40 type II genes in the genome Arabidopsis. However, all functionally characterized MADS-box genes are type II genes, and type I genes have been poorly characterized. I therefore compared the patterns of duplication and loss of type I genes and that of type II genes to gain insights into their evolutionary processes. The results suggested that type I genes have experienced much faster birth-and-death evolution than type II genes in the lineages leading to Arabidopsis and rice. Our further study suggested that more frequent segmental duplication and weaker

purifying selection in type I genes than in type II genes may be related to the high rate of birth-and-death evolution in type I genes. We also found that a number of type I genes had evolved under strong purifying selection.

Chapter 4 presents a method for predicting functionally differentiated regions between duplicate genes. There are many duplicated genes in the genome, and they have been subject to functional differentiation after the duplication event. If one can predict the gene region responsible for the differentiation, it will be helpful for the functional characterization of genes in multigene families. I therefore developed a simple statistical method for this purpose. This method is based on the assumption that significant rate differences between duplicate genes are correlated with the functional differentiation between them. This method was applied to 23 pairs of closely related MIKC-type MADS-box genes from petunia, and 14 pairs of sequences showed significant rate differences. A majority of the predicted regions were within the K-domain which is important for the dimerization of MADS-box proteins. These predicted regions may be particularly attractive for further experimentation.

Chapter 5 is concerned with the patterns of duplication and loss of homeobox genes in bilateral animals. I compiled the entire or almost entire homeobox gene sets from 11 species of bilateral animals ranging from *Caenorhabditis elegans* to humans. Phylogenetic analyses of the 2031 sequences suggested that there were at least 88 homeobox genes in the MRCA of nematodes and vertebrates. This result suggests that the repertoire of homeobox genes was already diverse in the ancestor of bilateral animals. Of the 88 ancestral genes only about 50 ~ 60 genes have left at least one descendent in each genome, and substantial numbers of genes have been lost in each lineage. While both the vertebrate and the invertebrate lineages lost similar numbers of ancestral genes, the vertebrate lineages gained many more homeobox genes than

the invertebrate lineages during evolution. In addition, substantial numbers of the products of fairly old duplication events also have been lost. Considering the functional importance of homeobox genes, the results suggest that both gene duplication and gene loss might have been important sources of the morphological diversity of animals.

CHAPTER 2

ANTIQUITY AND EVOLUTION OF THE MADS-BOX GENE FAMILY CONTROLLING FLOWER DEVELOPMENT IN PLANTS

SUMMARY

MADS-box genes in plants control various aspects of development and reproductive processes including flower formation. To obtain some insight into the roles of these genes in morphological evolution, we investigated the origin and diversification of floral MADS-box genes by conducting molecular evolutionary genetics analyses. Our results suggest that the most recent common ancestor of today's floral MADS-box genes evolved roughly 650 million years ago (MYA), much earlier than the Cambrian explosion. They also suggest that the functional classes T (SVP), B (and Bs), C, F (AGL20 or TM3), A, and G (AGL6) of floral MADS-box genes diverged sequentially in this order from the class E gene lineage. The divergence between the class G and E genes apparently occurred around the time of the angiosperm/gymnosperm split. Furthermore, the ancestors of three classes of genes (class T genes, class B/Bs genes, and the common ancestor of the other classes of genes) might have existed at the time of the Cambrian explosion. We also conducted a phylogenetic analysis of MADS-domain sequences from various species of plants and animals and presented a hypothetical scenario for the evolution of MADS-box genes in plants and animals, taking into account paleontological information. Our study supports the idea that there are two main evolutionary lineages (type I and type II) of MADS-box genes in plants and animals.

INTRODUCTION

MADS-box genes encode transcription factors and have been found in three eukaryotic kingdoms, plants, animals, and fungi. In plants, MADS-box genes include developmental regulatory genes comparable to homeobox genes in animals. The protein region encoded by the highly conserved MADS-box is called the MADS-domain and is part of the DNA-binding domain. It is composed of approximately 55 amino acids (aa). It has been proposed that there are at least 2 lineages (type I and type II) of MADS-box genes in plants, animals, and fungi (figure 2.1; Alvarez-Buylla et al. 2000b). Most of the well-studied plant genes are type II genes and have three additional domains compared with type I genes: intervening (I) domain (~30 codons), keratin-like coiled-coil (K) domain (~70 codons), and C-terminal (C) domain (variable length). These genes are called the MIKC-type and are specific to plants.

The plant-specific MIKC-type MADS-box genes were first discovered in flowering plants (angiosperms). They can be divided into at least nine classes on the basis of their functions and expression patterns (table 2.1). In angiosperms, several classes of MADS-box genes control flower formation and are often referred to as floral MADS-box genes. In particular, the 'ABC' model of flower formation proposes that the four floral components (organs) are controlled by the interactions of three classes of floral MADS-box genes, A, B, and C (Weigel and Meyerowitz 1994; Ma and dePamphilis 2000). More recently, this 'ABC' model was amended to include an interaction with an additional class of genes, called class E genes (Theissen and Saedler 2001). According to this amended model, which is called the quartet model, the combinatorial tetramers of four classes of floral MADS-domain proteins regulate the development of the four floral components (Honma and Goto 2001; Theissen 2001): sepals by class A genes, petals by class A, B, and E genes, stamens

by class B, C, and E genes, and carpels by class C and E genes (table 2.1). Class A, C, and E genes are also involved in floral meristem development.

Other classes include the class D genes, which are the close relatives of class C genes and control ovule development (Theissen 2001). The recently proposed class B-sister (Bs) genes also appear to control the development of ovule and seed coat, though their protein sequences are quite different from those of D genes (Becker et al. 2002; Nesi et al. 2002). In addition, another group of MADS-box genes that includes *AGL20* (*AGAMOUS-LIKE 20*) in *Arabidopsis thaliana* (thale cress; hereafter called Arabidopsis) plays a pivotal role in flower activation as an integrator of genetic and environmental flowering pathways (Lee et al. 2000). This group of genes will be called "class F genes" instead of the TM3 or orphan group as previously named (Purugganan 1997; Becker et al. 2000). Several genes such as *AGL6* in Arabidopsis seem to be involved in the development of both flowers and vegetative organs (Alvarez-Buylla et al. 2000a). We call these genes "class G genes". Furthermore, there is a group of genes that trigger flowering as an initiator or a repressor. Loss of function of some of these genes resulted in late flowering or early flowering (Hartmann et al. 2000; Michaels et al. 2003). We call these genes "class T genes". All the above genes are directly involved in flower formation of angiosperms. We therefore call them "floral MADS-box genes" in this paper, though this terminology is usually used for the class A, B, C, and E genes. Note that our classification of MADS-box genes is for simplifying the explanation of our study rather than for proposing new terminologies. There are a large number of other MADS-box genes in angiosperms. Some of them appear to control flowering time or formation of leaves, fruits, roots, etc. (Zhang and Forde 1998; Michaels and Amasino 1999; Sheldon et al.

1999; Alvarez-Buylla et al. 2000a; Hartmann et al. 2000), but the functions of other genes are unknown.

The primary purpose of this paper is to investigate the evolutionary relationships and divergence times of floral MADS-box genes. However, since most floral MADS-box genes are known to exist in gymnosperms as well (e.g. Winter et al. 1999; Becker et al. 2000), we consider the genes from both angiosperms and gymnosperms. Previously, Purugganan (1997) studied a similar problem, but this problem should be reexamined since extensive data on MADS-box genes have become available in recent years. Furthermore, to understand the long-term evolution of MADS-box genes, we will also investigate the evolutionary relationships of MADS-domain sequences from plants and animals.

MATERIALS AND METHODS

Floral MADS-box genes used

At present, MIKC-type MADS-box gene sequences are available from various species of angiosperms, gymnosperms, ferns, clubmosses, and mosses (GenBank, TIGR). There are more than 70 MADS-box genes annotated in Arabidopsis (the Arabidopsis Genome Initiative 2000 and our unpublished study). Similarly, we have identified about 70 genes from rice by conducting a TBLASTN search in the Rice Genome Database of China (Yu et al. 2002) and the TIGR Rice Genome Database. From these databases, we compiled 293 full-length MIKC-type MADS-box genes. In the phylogenetic study of floral MADS-box genes, we used 23 reproductive genes, covering all classes of genes shared by angiosperms and gymnosperm species (class B, Bs, C, F, G, and T genes). These genes were chosen from the well-studied eudicot species Arabidopsis, monocot species *Oryza sativa* (rice) and *Zea mays* (maize), and gymnosperm species *Pinus radiata* (Monterey pine), *Picea abies* (Norway spruce) and *Gnetum gnemon* (table 2.1). We did not include the gymnosperm class E gene (*PrMADS1*) reported from the pine *Pinus radiata*, because this appears to be a contaminated gene from *Eucalyptus grandis* at the time of experimentation (G. Theiβen, personal communication). Class A and E genes from angiosperms were also included from our analysis because of their importance during flower development, though these genes have not been found in gymnosperms. Class D genes were excluded from the analysis, because their protein sequences were close to C gene sequences and the distinction between C and D genes was not always clear-cut.

Protein sequences of these genes were obtained from GenBank or TIGR. The names of the proteins and their GenBank accession numbers or TIGR locus numbers

are as follows: AGL9 (At1g24260), AGL6 (At2g45650), AGL20 (At2g45660), APETALA1 (AP1) (At1g69120), APETALA3 (AP3) (At3g54340), PISTILLATA (PI) (At5g20240), AGAMOUS (AG) (At4g18960), SVP (At2g22540), OsMADS3 (S59480), OsMADS4 (T03902), OsMADS8 (AAC49817), OsMADS14 (AAF19047), OsMADS16 (AAD19872), OsMADS17 (AAF21900), OsMADS50 (BAA81886), OsMADS54 (BAA81880), DAL1 (T14846), DAL2 (S51934), DAL3 (T14848), DAL13 (AAF18377), GGM13 (CAB44459), ZMM17 (CAC81053), ABS (At5g23260), and LAMB1 (AAG08991). As is shown in table 2.1, the protein sequence of a class T gene from *G. gnemon*, GGM12, is available, but this was not used in our analysis, because it was a fragmentary sequence. In this paper, we have used simplified gene notations to make this study understandable for a wide audience.

Phylogenetic analysis of MIKC-type genes

We used protein sequences for our phylogenetic analysis, because the evolutionary pattern of protein sequences appears to be simpler than that of DNA sequences (Nei and Kumar 2000, chapter 2) and protein sequences often give more satisfactory results than DNA sequences in the study of long-term evolution (Hashimoto et al. 1994; Russo, Takezaki, and Nei 1996; Glazko and Nei 2003). In the present case, we could minimize the effect of variation in the GC content at third codon position by using protein sequences.

We aligned 293 protein sequences using the computer program ClustalX (Thompson et al. 1997) with default parameters except the gap opening parameter of 2.0. We then constructed a preliminary neighbor-joining (NJ) tree with Poisson-correction (PC) distance using the computer program MEGA2 (version 2.1) (Kumar et al. 2001). (In MEGA2, taxon input orders are randomized for all bootstrap

replications.) According to this tree, we divided 293 protein sequences into 18 groups and aligned them separately with the same parameters using ClustalX. These aligned groups were again aligned to each other using the profile alignment option in this program. After elimination of gaps in this alignment, we constructed an initial NJ tree using PC distance. As mentioned above, we selected 24 representative sequences of 142 amino acid sites without gaps, including the MADS-domain, the K-domain, and the conserved region of the I-domain. We then constructed NJ trees with *p*-distance (proportion of different amino acids), PC distance, and PC gamma distance (Nei and Kumar 2000;chapter 2) using MEGA2. In addition, we constructed maximum-likelihood (ML) trees using the PROTML program with the Poisson and JTT models (Adacchi and Hasegawa 1996) and maximum-parsimony (MP) trees using the PAUP* program with the stepwise addition and tree-bisection-reconnection (TBR) algorithm with 500 bootstrap resamplings (Swofford 1998). A distantly related MADS-box gene, *LAMB1*, from the clubmoss *Lycopodium annotinum*, was used as the outgroup in this study. According to our phylogenetic analysis, this gene was closely related to type I genes (see Supplemental material at MBE web site: http://www.molbiolevol.org). Alvarez-Buylla et al. (Alvarez-Buylla et al. 2000b) have suggested that type I proteins do not have the K-domain (putative coiled-coil structure). However, the LAMB1 protein has a domain similar to the K-domain, including regularly spaced hydrophobic amino acids (e.g., leucine, isoleucine, and valine), which are known to be important for protein-protein interaction (Moon et al. 1999). Therefore, we could align the LAMB1 protein sequence with other MADS-domain protein sequences. Moreover, *LAMB1* has been suggested to be a new MIKC-type MADS-box gene designated as MIKC*-type, whereas the other 23 genes were classical MIKC genes (MIKC^c-type; Henschel et al. 2002). There are two more

MIKC*-type genes (*PPM3* and *PPM4*) reported from the moss *Physcomitrella patens* (Henschel et al. 2002). Use of these genes as the outgroups produced essentially the same topology for the floral MADS-box genes.

Once the topology of the phylogenetic tree was determined, we estimated the times of divergence between various types of genes using the linearized tree method (Takezaki, Rzhetsky, and Nei 1995). In the LINTREE method, the timescale constructed does not apply to the outgroup. We also used Yoder and Yang's (2000) likelihood method implemented in the computer program PAML (Yang 2002) with a different evolutionary rate for class B genes of angiosperms compared with that of the remaining genes. Sanderson's (2003) penalized likelihood method was also used.

Phylogenetic analysis of MADS-domains from plants and animals

The animal species studied so far seem to have at least one type I gene and one type II MADS-box gene, but the number of the genes is generally very small (Alvarez-Buylla et al. 2000b). All of the well-studied plant MADS-box genes are type II genes, and there are many other type II genes in angiosperms and gymnosperms. The existence of plant type I genes has not been well established except in Arabidopsis, rice, and clubmoss (Alvarez-Buylla et al. 2000b and our unpublished data). To study the evolutionary relationships of type I and type II MADS-box genes, we used the MADS-domain sequences (~55 aa) of 87 representative genes from plants (Arabidopsis, rice, spruce, pine, gnetum, fern, clubmoss, and moss) and animals (human, mouse, zebrafish, fruitfly, mosquito, and nematode) (see Supplemental material at MBE web site: http://www.molbiolevol.org). In this study we used only MADS-domain sequences, because animal genes do not have the IKC domain. The 87 MADS-domain sequences were aligned by using

ClustalX, and the evolutionary relationships of the genes were examined by constructing a NJ tree with *p*-distance for 55 shared amino acids.

RESULTS

Phylogenetic tree of MIKC-type genes

The phylogenetic tree of 24 representative MADS-box genes from eudicots, monocots, and gymnosperms is presented in figure 2.2. This tree was obtained by the NJ method with PC distance, but very similar trees were obtained by NJ with *p*-distance and PC gamma distance and ML and MP methods (see Supplemental material at MBE web site: http://www.molbiolevol.org). Although the bootstrap values for interior branch a-b as well as for the B or Bs gene clades of this tree are very low, the other clades involving class E, G, A, F, and C genes are supported with reasonably high bootstrap values (> 70%). Similar patterns were observed in trees obtained by other tree-building methods. Therefore, the portion of the tree containing the class E, G, A, F, and C genes appears to be quite reliable.

This tree suggests that after separation of the class T genes from the non-T floral MADS-box genes, class B/Bs genes were the first to diverge from the rest of non-T floral MADS-box genes, though this is still provisional. Class C genes then separated from the genes belonging to class F, A, G, and E genes. The next group of genes to diverge was class F genes. Moreover, the taxonomic distribution of functional classes of floral MADS-box genes (table 2.1) suggests that class E and G genes, which diverged most recently, diverged around the time of angiosperm/gymnosperm split. Several class specific or taxon specific amino acids have been reported (e.g., Huang et al. 1995; Kramer, Dorit, and Irish 1998), but we did not find any key features of conserved amino acids supporting any clade of the tree in figure 2.2. We also compared the positions of introns among all classes of genes, but the positions were too conserved to be informative for inferring the phylogenetic relationships of MADS-box genes (data not shown).

Estimates of divergence times

Although molecular estimates of divergence times between genes or species depend on a number of assumptions and are generally very crude (Nei, Xu, and Glazko 2001; Glazko and Nei 2003), they are still useful for obtaining a rough idea of the evolutionary history of genes or species. With this caveat in mind, we estimated the times of divergence between different classes of genes. In the estimation of divergence times, the hypothesis of constant evolutionary rate should first be tested, and then the sequences whose evolutionary rate significantly deviates from constancy should be eliminated (Takezaki, Rzhetsky, and Nei 1995). In this case a number of authors have used Yang's (2002) or Gu and Zhang's (1997) likelihood method for estimating gamma parameter a. However, for the purpose of time estimation, these methods, particularly the former method, tend to give underestimates of a, and this often leads to overestimation of divergence times when ancient divergence times are estimated (Nei, Xu, and Glazko 2001; Glazko and Nei 2003). This seems to be particularly so for slowly evolving genes such as cytochrome c. Dickerson (1971) showed that in cytochrome c and hemoglobin the number of amino acid substitutions estimated by PC distance ($a = \infty$) is nearly proportional to the time since species divergence up to about 500 MYA. Nei (1987) also showed that variation in evolutionary rate among amino acid sites has a relatively small effect on time estimates unless the sequence divergence is very high. We have therefore decided to use primarily PC distance for estimating divergence times. However, we also used Dayhoff's distance to take into account backward and parallel mutations. According to Nei and Kumar (2000), Dayhoff's distance can be computed by a PC gamma distance with $a = 2.25$. We therefore used this method. Note that the use of these

distances give conservative estimates of divergence times compared with those obtained by the PC gamma distance with a likelihood estimate of *a* (see below).

We used the two-cluster test of Takezaki, Rzhetsky, and Nei (1995) to examine the applicability of the molecular clock for the tree in figure 2.2 and found that the four B genes (2 AP3 and 2 PI genes) evolved significantly faster than other genes at the 3 percent level. We therefore eliminated these four genes and constructed a linearized tree with PC distance for the remaining genes (figure 2.3*A*). The two-cluster test also showed that the spruce C gene evolved significantly slower than the Arabidopsis and rice C genes at the 5 percent level, but we retained this gene because it was important for calibration of the timescale and a relatively small deviation of a sequence from rate constancy does not affect time estimates seriously (Nei and Kumar 2000). In addition to the four B genes, we also eliminated all Bs genes because of the uncertain phylogenetic position of the genes (figure 2.2). To compare our results with previous estimates of divergence times for floral MADS-box genes by Purugganan (1997), we constructed a linearized tree for a simplified Purugganan tree topology. Purugganan studied the phylogenetic tree of many floral MADS-box genes, but the bootstrap values of the interior branches were so low that he merged several interior nodes. If we use only 24 genes as used in our study, the linearized Purugganan tree becomes as given in B of figure 2.3. We therefore estimated the divergence time for the merged node (a-b-c-d).

To calibrate the timescale of the linearized tree, a calibration point is necessary. For our dataset, the divergence times between "eudicots" and "monocots" and between "gymnosperms" and "angiosperms" may be used as the calibration point. However, there is no good fossil record for the divergence of "eudicots" and "monocots", and previous authors used various values (131 – 200 MYA) for this

divergence (Wolfe et al. 1989; Laroche, Li, and Bousquet 1995; Soltis et al. 2002). This calibration point also gives some unreasonable time estimates for our dataset (see below). By contrast, there seems to be a general consensus about the divergence time between angiosperms and gymnosperms, which is about 300 MYA. This estimate is supported by both paleontological data and molecular time estimates (Stewart and Rothwell 1993; Savard et al. 1994; Goremykin, Hansmann, and Martin 1997; Soltis et al. 2002). In addition, the angiosperm/gymnosperm split calibration will produce smaller standard errors of time estimates than the monocot/eudicot split calibration, because the former is a more ancient evolutionary event than the latter (Glazko and Nei 2003). We have therefore decided to use this time as the calibration point.

Figure 2.3*A* shows that each of class G, F, and C genes included one gymnosperm gene and two angiosperm genes. We therefore computed the average PC distance (d) between the gymnosperm and angiosperm genes and obtained $d = 0.372$. This gives an estimate of the rate of amino acid substitution (r) to be $r = d/(2\times300)$ per million years or $r = 6.2\times10^{-10}$ per year. The timescales for trees A and B in figure 2.3 were obtained by using this rate of amino acid substitution. The times of divergence between different classes of genes can then be estimated from these linearized trees. The results obtained are presented in table 2.2. This table also includes time estimates obtained by using Dayhoff and PC gamma distances. When PC distance is used, the time of divergence between the T and the non-T floral MADS-box genes is estimated to be about 652 MYA. This is well before the time of the Cambrian explosion (about 545 MYA; see figure 2.4). Table 2.2 also suggests that the divergence between class B genes and other non-T floral MADS-box genes (612 MYA) occurred before the Cambrian explosion. The divergence between class C genes and the remaining non-T floral genes (537 MYA) again appears to have

occurred around the Cambrian explosion. This might sound strange, because most animal and plant phyla are believed to have evolved no earlier than the time of the Cambrian explosion. However, recent paleontological data (Xiao, Zhang, and Knoll 1998) suggest that by this time green algae had already evolved. The fossil record suggests that the first land plants such as bryophytes appeared around 450 MYA. Our estimates in table 2.2 suggest that class A, G, and E gene lineages originated after the occurrence of land plants. Table 2.2 also includes an estimate (556 MYA) of the divergence time between B and Bs genes. In the estimation of this divergence time, the class B genes from angiosperms were excluded because of their faster rate of evolution compared to other genes, and the divergence time was estimated by dividing the distance between the B and Bs genes by $2r$, where $r = 6.2\times10^{-10}$ per year. This estimate suggests that the gymnosperm B and Bs genes diverged a long time ago, if they are clearly definable separate gene groups.

Because many of the above estimates of divergence times far exceed the times of first appearance of land plants in the fossil record (450 MYA), they might be overestimates. However, if we use Dayhoff distance or PC gamma distance with an ML estimate (1.06) of a obtained by Gu and Zhang's method, the divergence time estimates become even greater (table 2.2). This was especially so when PC gamma distance was used. In this case branch points a and b were estimated to be 816 and 743 MYA, respectively. We also used Yoder and Yang's method without eliminating B genes but with the assumption that these genes evolved faster than the other genes (two rates model). This method also gave greater estimates than those obtained by PC distance even when the Poisson model ($a = \infty$), Dayhoff model, or Poisson gamma model ($a = 1.06$) was used (table 2.2). Sanderson's penalized likelihood method gave even greater estimates than other methods (see Supplemental material at MBE web

site: http://www.molbiolevol.org). Therefore, our estimates obtained from the linearized tree method with PC distance are most conservative.

One might wonder whether we used most closely related copies (orthologous genes) of the class G, F, and C genes between angiosperms and gymnosperms for computing the timescale. Actually we tried to do so, but there is no guarantee for the use of real orthologous genes, partly because no complete genome sequence is currently available from any gymnosperm species and partly because it is not easy to determine orthologous genes even in the presence of complete genome sequences (Theissen 2002). However, if we have used nonorthologous genes for any of these gene classes, our estimates would have been lower than unbiased estimates, because the rate of amino acid substitution should have been overestimated. This factor also tends to make our estimates conservative.

As mentioned earlier, some authors used the monocot/eudicot divergence (200 MYA) as the calibration point. In our dataset, however, the use of this calibration point gave a divergence time estimate of 251 MYA between the angiosperms and gymnosperms. (The average distance of the angiosperm and gymnosperm genes from class C, F, and G genes was used.) When we used a calibration point of 150 MYA for the monocot/eudicot divergence, we obtained an estimate of divergence of 188 MYA for the angiosperm and gymnosperm split. These estimates are clearly unreasonable, because angiosperms and gymnosperms are believed to have diverged about 300 MYA. We therefore decided not to use the monocot/eudicot calibration point. Incidentally, if we use the angiosperm/gymnosperm divergence (300 MYA) as the calibration point, we obtain an expected divergence time of 239 MYA between monocots and eudicots.

In figure 2.3B, we have Purugganan's topology. If we estimate the branch point (a-b-c-d) of this topology, we obtain 575 MYA. This is considerably greater than Purugganan's estimate (476 MYA). This difference has occurred partly because Purugganan used the monocot/eudicot divergence (200 MYA) as the calibration point and partly because he used paralogous genes of E genes between monocots and eudicots.

Phylogenetic tree of 87 MADS-domains from plants and animals

Figure 2.5 shows a NJ tree of type I and type II MADS-domain sequences from plant and animal species. Type I and type II genes form their own clades, and these clades are quite well supported by the bootstrap test. Type II genes are further divided into plant and animal genes. The monophyletic cluster of animal type II genes is well supported. Plant type II genes also form a monophyletic cluster, though the bootstrap support is rather weak (51%). Animal type II genes form a well-supported monophyletic group. Animal type I gene also form a monophyletic group. By contrast, plant type I genes do not form a monophyletic cluster, though genes from Arabidopsis and rice form a well-supported cluster. This could be due to the small number of amino acids used.

Although our results are somewhat ambiguous, they generally support Alvarez-Buylla et al.'s (2000b) view that the type I and type II genes were generated by a gene duplication that occurred before the plant/animal divergence. Animal type I genes control very basic transcription processes concerned with various aspects of cell growth and differentiation and neuronal transmission, etc., whereas type II genes are responsible for muscle development (Shore and Sharrocks 1995). The function of plant type I genes is not well understood, and these genes have only been identified by

genomic sequencing of Arabidopsis and rice, though the *LAMB1* gene in the clubmoss has been suspected to be a type I gene. Many plant type II genes in figure 2.5 belong to one of the nine classes of MIKC-type MADS-box genes considered in figure 2.2. However, there are additional MADS-box genes that control various developmental processes such as root formation.

Plant type II genes form many clades of a few genes, and many of these clades are statistically supported relatively well. However, their inter-clade relationships are poorly supported. In particular, B/Bs genes are no longer monophyletic. Nevertheless, the relationships of the genes belonging to floral MADS-box gene classes A, C, E, F, G, and T are virtually the same as those in figure 2.2. Therefore, the tree in figure 2.5 may reflect the evolutionary history of MADS-box domains to some extent. The low bootstrap values for these relationships are primarily due to the fact we used many sequences with only 55 aa and that there are many other MADS-box genes which are closely related to but are distinct from floral MADS-box genes in plant genomes. It is quite possible that the nine classes of floral MADS-box genes were derived from some of these distinct MADS-box genes nearly independently. In the present case it is not meaningful to make any attempt to estimate the divergence times of these genes, because the number of amino acids per sequence is very small.

DISCUSSION

Reliability of estimates of divergence times

The fact that nonflowering gymnosperms have most classes (B, Bs, C, G, and T) of floral MADS-box genes indicate that the gene duplications that generated these genes occurred long before their angiosperm-specific functions were established. It is not clear what kinds of function these floral MADS-box genes had before their functional diversification, but they were probably involved in the regulation of broad developmental and reproductive processes, as was suggested by Becker et al. (2000). This evolutionary pattern is similar to that of homeobox genes that control segmentation of animal body structure (Zhang and Nei 1996; Purugganan 1998). Cnidarian species such as jellyfish do not have a segmented body structure, yet they have hox genes (Ferrier and Holland 2001). Actually, similar evolutionary patterns are observed with several other gene families controlling development (e.g. Burglin 1997; Meyerowitz 2002), and it appears that the occurrence of gene duplication before functional diversification is a general phenomenon with gene families controlling development.

Our conservative estimates suggest that class A and B floral genes diverged about 612 MYA, which is two times earlier than the paleontological estimates of divergence time between gymnosperms and angiosperms and far exceeds the paleontological estimate of the time of first land plants (mosses) (ca. 450 MYA). However, mosses are known to have at least two genes that are homologous to classical MIKC-type genes (Henschel et al. 2002). It should also be noted that classical MIKC-type genes have been identified even in green algae such as *Chara*, *Coleochaete*, and *Closterium* (M. Hasebe, personal communication), which evolved

earlier than land plants. Note that the oldest fossil record of green algae is 700 – 750 million years old (Chen and Xiao 1991; Butterfield 2000), though green algae do not appear to be monophyletic. These observations suggest that our estimate of the time of origin of floral MADS-box genes may not be too early.

In this discussion we used the most conservative estimates of divergence times obtained by PC distance. If we use PC gamma distance or Yoder and Yang's method, estimates of the time of origin of floral MADS-box genes become greater than 800 MYA. These estimates appear to be too early if we consider the fossil record of land plants and green algae, but we cannot rule out this possibility completely at present, because the fossil record is notoriously incomplete. It is worth noting that until recently all or most orders of placental mammals were believed to have diverged only about 65 MYA. At present, however, we know the fossil remain of a placental mammal that is about 125 million years old (Ji et al. 2002). The notion of the Cambrian explosion, in which most visible eukaryotic organisms are believed to have been absent before 545 MYA, is also slowly changing. We now know 570 million years old fossils of animal eggs (Xiao, Zhang, and Knoll 1998), 900 - 1,200 million years old fossils of red algae (Butterfield 2000), and 1,100 - 1,200 million years old trace fossils of worm (Seilacher, Bose, and Pfluger 1998; Rasmussen et al. 2002), though the authenticity of these trace fossils have been questioned (Conway Morris 2002).

Nevertheless, it is not clear what kind of function the MIKC-type genes had in ancestral non-seed plants. In recent years an intensive study has been made to identify genes orthologous to floral MADS-box genes in non-seed plants, but the study has not been very successful (e.g. Munster et al. 1997; Hasebe et al. 1998; Hohe 2002; Svensson and Engstrom 2002). What are the possible reasons for these

negative results? There seem to be at least five reasons. First, the orthologs of floral MADS-box genes in non-seed plants so far studied might have been lost in the course of evolution. Second, the orthologs of floral MADS-box genes in non-seed plants are so different from the floral MADS-box gene that it is difficult to identify orthologs now. Actually, figure 2.5 shows that several genes from nonseed and seed plants form several clades, although the bootstrap supports are very low. Third, our molecular time estimates are too old even though we used the most conservative method. This may happen if the rate of amino acid substitution was faster in the early stage of evolution of floral MADS-box genes than in the later stage. Fourth, the current fossil record is incomplete and land plants might have evolved earlier than currently believed. Fifth, the genes so far studied may be incomplete, and a complete genome search may find the genes. At the present time, however, it is difficult to resolve the current discrepancy between the theoretical and experimental studies.

Long-term evolution of MADS-box genes

As mentioned earlier, MADS-box genes are highly conserved, and the MADS-domain sequences are shared by plants, animals, and fungi, indicating that MADS-box genes have an ancient history. Therefore, studying the history of MADS-box genes, we should be able to obtain some insight into the evolution of morphological characters in eukaryotes. Unfortunately, our knowledge about the MADS-box genes and their function in early eukaryotes is quite limited at present. Nevertheless, it would be interesting to speculate about the evolution of MADS-box genes in eukaryotes taking into account both paleontological information and molecular dating. Such a plausible scenario may give some useful information for future experimental

studies. Here we consider only the evolution of plant and animal genes, because MADS-box genes in fungi other than the budding yeast are not well studied.

In figure 2.5 we have seen that both plants and animals have two different types of MADS-box genes, type I and type II genes. As indicated by Alvarez-Buylla et al. (2000b), this suggests that these two types of genes diverged by a gene duplication that occurred before the plant/animal divergence (figure 2.6). The oldest geological evidence of eukaryotes is given by a lipid biomarker, which has been dated 2,700 MYA (Brocks et al. 1999). There are also eukaryotic fossils that have been dated 2,100 MYA (Han and Runnegar 1992). There is no fossil record that indicates the time of divergence between plants and animals, but molecular data suggest that the divergence time is about 1,400 MYA (Feng, Cho, and Doolittle 1997; Wang, Kumar, and Hedges 1999; Nei, Xu, and Glazko 2001). Therefore, if these estimates are reliable, the gene duplication must have occurred some time between 1,400 MYA and 2,700 MYA (figure 2.6). Because yeast, *Caenorhabditis elegans*, and *Drosophila melanogaster* have a small number of type I and type II genes (two type I genes and two type II genes in yeast, one type I gene and one type II gene in *C. elegans* and *D. melanogaster*), it is probable that the early plants (possibly red and brown algae, Cavalier-Smith 2002; note that the monophyly of plants and these algae is still controversial) also had a small number of type I and type II genes. This hypothesis may be tested by examining the genomes of extant red and brown algae. Because these early plants have quite complex morphological characters and life cycles, this would help us to understand the ancient function of MADS-box genes during plant evolution. According to our conservative estimates of divergence times of MADS-box genes in table 2.2, a group of green algae which are believed to have evolved 700 – 750 MYA (figure 2.6) is expected to have at most one gene that is ancestral to all

the floral MADS-box genes currently present in angiosperms and gymnosperms. However, if our estimates from gamma distance are correct, green algae may have three genes that are ancestral to the current T, B (and Bs), and E (or A, C, F, G) classes of genes.

Figure 2.6 shows several evolutionary events in both animal and plant lineages. Molecular estimates of divergence times of early metazoan animals are almost always considerably earlier than paleontological estimates. For example, molecular data have suggested that the nematode lineage diverged from the vertebrate lineage 800 – 1,100 MYA (e.g. Feng, Cho, and Doolittle 1997; Wang, Kumar, and Hedges 1999; Nei, Xu, and Glazko 2001), which is about two times earlier than the times of the Cambrian explosion. The nematode *C. elegans* is known to have one type I gene and one type II MADS-box gene (Alvarez-Buylla et al. 2000b; our unpublished data). The type I and type II MADS-box genes in animals have not been studied very well, but zebrafish has several type I and type II genes (our unpublished results). These findings suggest that MADS-box genes are very ancient and evolved gradually in the long history of plants and animals.

Previously we indicated that the MADS-box gene family is an important gene family comparable to the animal homeobox gene family. In this regard, it is interesting to note that the homeobox gene family also exists in plants, animals, and fungi (Burglin 1997; Kappen 2000), and that there are at least two lineages of homeobox genes that diverged before the plant/animal/fungal split. It would be interesting to investigate how these two different multigene families controlling development coevolved.

Gene family expansion or birth-and-death evolution?

Figure 2.2 shows a pattern of functional diversification of major groups of MADS-box genes. This figure suggests that the number of genes of this multigene family has steadily increased as the reproductive system becomes more complex. However, although the gene number must have increased from the time of early plants, this tree does not give the entire picture of evolution of MADS-box genes, because we did not include many genes that are not directly related to flower formation. Our tree in figure 2.5 is not very reliable, but if it represents a general pattern of evolution of MADS-box genes, it is possible that different floral MADS-box genes were derived from other floral MADS-box genes, which have already been lost, or even from other reproductive MADS-box genes. Furthermore, the Arabidopsis genome is known to contain several MADS-box pseudogenes or truncated genes (our unpublished data), indicating that some MADS-box genes died out in the evolutionary process. These observations suggest that the MADS-box gene family might have been subjected to the birth-and-death model of evolution, in which some genes generate duplicate genes with new functions but others become nonfunctional or are deleted from the genome (Nei, Gu, and Sitnikova 1997). If this is the case, it is possible that the genome of gymnosperms or ferns contains nearly as many MADS-box genes as those of angiosperm genomes and the genes in these plants merely exert different functions that are required for the different forms of reproduction. Of course, it is also possible that the phylogenetic tree of current angiosperm genes in figure 2.2 largely reflects the history of the increase of member genes of the MADS-box gene family in gymnosperms and angiosperms. At the present time, we cannot distinguish between the two alternative hypotheses, but this can be done rather easily if the genomic sequences of gymnosperms and ferns are determined in the future. However, note

that the two hypotheses are not mutually exclusive and we are interested only in the relative importance of the two possibilities.

Table 2.1 Representatives of different classes of MADS-box genes considered in this study

Class	Gene and source			Function in Arabidopsis
	Arabidopsis (Eudicots)	Rice or Maize (Monocots)	Norway spruce, Monterey pine, or Gnetum (Gymnosperms)	
Class A (AP1 or SQUA)	Arabi A (APETALA1 or AP1)	Rice A (OsMADS14)	unknown	Sepal and petal development, floral meristem development (Weigel and Meyerowitz 1994)
Class B AP3/PI or DEF/GLO)	Arabi B-AP3 (APETALA3 or AP3) Arabi B-PI (PISTILLATA or PI)	Rice B-AP3 (OsMADS16) Rice B-PI (OsMADS4)	Spruce B (DEFICIENS-AGAMOUS-LIKE 13 or DAL13)	Petal and stamen development (Weigel and Meyerowitz 1994)
Class Bs	Arabi Bs (ABS)	Maize Bs (ZMM17)	Gnetum Bs (GGM13)	Ovule and seed coat development (Nesi et al. 2002)
Class C (AG or PLENA)	Arabi C (AGAMOUS or AG)	Rice C (OsMADS3)	Spruce C (DEFICIENS-AGAMOUS-LIKE 2 or DAL2)	Stamen and carpel development, floral meristem development (Weigel and Meyerowitz 1994)

Class D (not used in this study)	Arabi D (AGL11)	Rice D (OsMADS13)	unknown	Ovule development (Theiβen 2001)
Class E (AGL2/4/9)	Arabi E (AGL9)	Rice E (OsMADS8)	unknown	Petal, stamen, carpel and floral meristem development (Theiβen 2001)
Class F (AGL20 or TM3)	Arabi F (AGL20)	Rice F (OsMADS50)	Spruce F (DAL3)	Flowering activation (integrator of genetic and environmental flowering pathways) (Lee et al. 2000)
Class G (AGL6)	Arabi G (AGL6)	Rice G (OsMADS17)	Spruce G (DAL1)	Expressed in both vegetative and reproductive tissues (Alvarez-Buylla et al 2000a)
Class T (SVP or STMADS11)	Arabi T (SVP)	Rice T (OsMADS54)	Gnetum T (GGM12, partial sequence - not used in this study)	Flowering repression (Hartmann et al. 2000)

NOTE.—We used simplified gene names. Commonly used names and their abbreviations are given in parentheses. The function of each class of genes is based on the studies in Arabidopsis. 'Arabi' indicates Arabidopsis. All classes of genes are members of floral MADS-box genes.

Table 2.2. Estimates of divergence times (± standard errors) of floral MADS-box genes

Node	Linearized tree method			Maximum-likelihood method		
	PC distance	Dayhoff distance (a = 2.25)	PC gamma distance (a = 1.06)	Poisson model ($a = \infty$)	Dayhoff model	Poisson + gamma model (a = 1.06)
(a) T/(others)	**652 ± 72**	721 ± 91	816 ± 120	816	813	836
(b) B/(C/D-F-A-G-E)	**612 ± 62**	668 ± 77	743 ± 99	749	775	772
(c) (C/D)/(F-A-G-E)	**537 ± 44**	573 ± 54	612 ± 68	630	647	631
(d) F/(A-G-E)	**502 ± 42**	531 ± 50	566 ± 62	564	569	586
(e) A/(G-E)	**374 ± 39**	380 ± 45	388 ± 51	428	406	422
(f) G/E	**289 ± 29**	286 ± 31	282 ± 35	341	327	340
(g) B/Bs	**556 ± 65**	598 ± 78	646 ± 94	656	662	714
Node a-b-c-d (Purugganan tree)	**575 ± 49**	623 ± 59	684 ± 75	689	701	706

NOTE.—Unit of time estimates is MYA. The gymnosperm/angiosperm split (ca. 300 MYA) in classes C, F, and G was used for calibrating the timescale. Dayhoff distance was computed by using PC gamma distance with a = 2.25 (Nei and Kumar 2000, chapter 2).

In the linearized tree method, time estimates for nodes (a) ~ (f) were computed by using 16 genes. (Three Bs genes and 4 angiosperm B genes were excluded.) The time of divergence between the B and Bs genes was estimated separately (see text). Since the ML method does not give proper standard errors (Yoder and Yang 2000), these values are not presented.

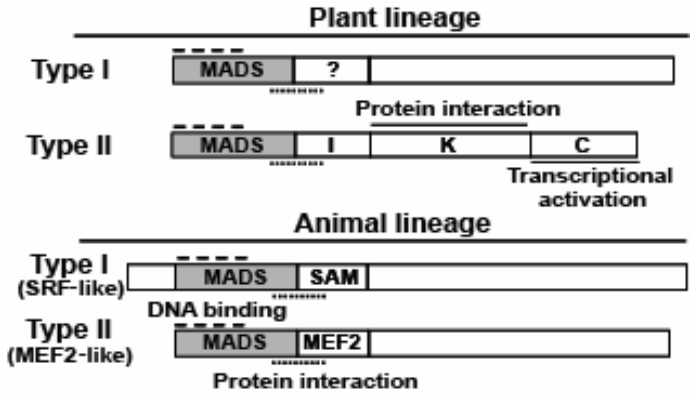


Figure 2.1. Schematic diagram of two types (types I and II) of MADS-box genes in plants and animals. The plant-specific MIKC-type MADS-domain proteins are presented with the name and function of each conserved domain. A broken line indicates the DNA-binding region, and a dotted line the protein-protein interaction region. This figure has been modified from Alvarez-Buylla et al. (2000).

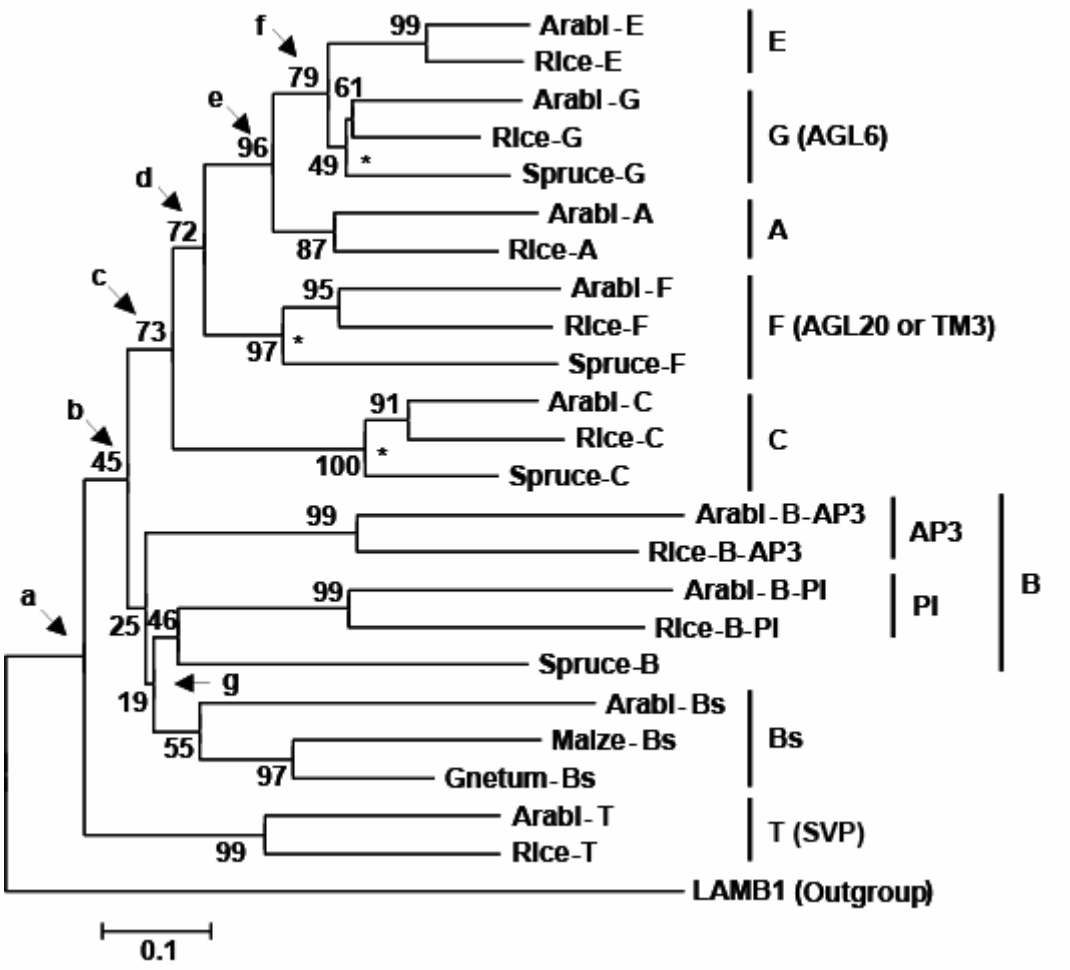


Figure 2.2. Phylogenetic tree of eight classes of MADS-box genes (A, B, Bs, C, D, E, F, G, and R) from monocots, dicots, and gymnosperms with a gene from the clubmoss *Lycopodium annotinum*, *LAMB1*, used as the outgroup. The number for each interior branch is percent bootstrap value (500 resamplings). The scale bar indicates the estimated number of amino acid substitutions per site. The number of amino acids used was 142 without gaps per sequence. AP3 and PI are abbreviations of APETALA3 and PISTILATA, respectively. Gene names were simplified to make the paper understandable to a wide audience (see table 2.1). Calibration points used for estimating divergence times are marked with "*".

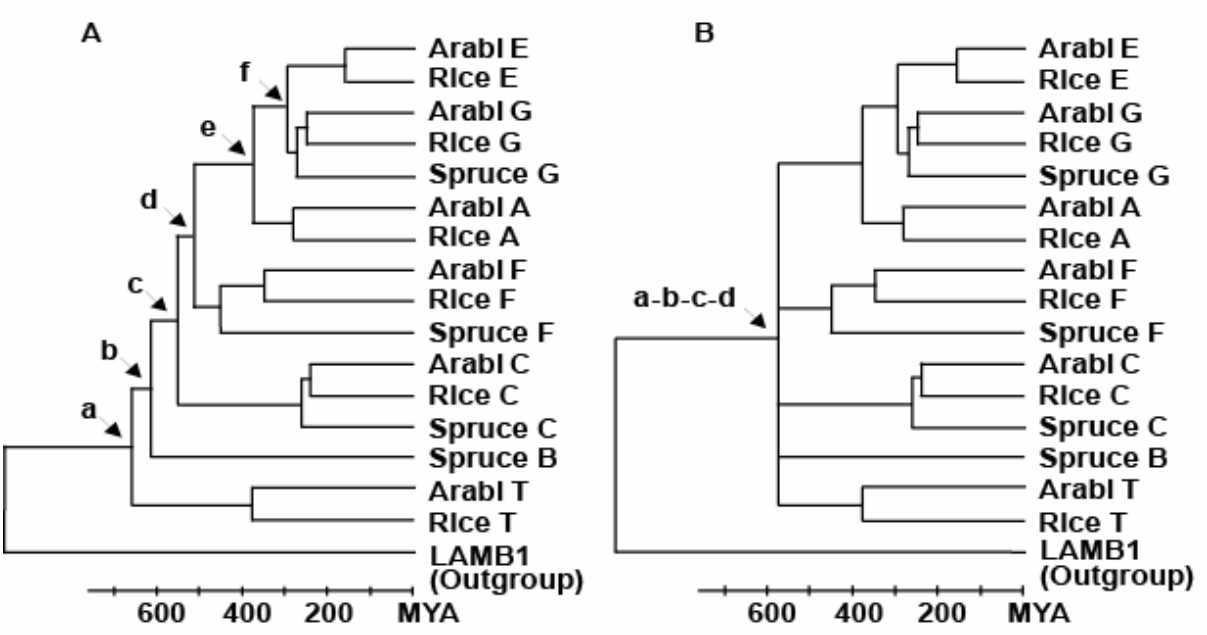


Figure 2.3. Linearized trees used for estimating divergence times. The timescale is based on the results with PC distance. (A) Topology from figure 2.2. (B) Topology when the interior branches between nodes a, b, c, and d are collapsed.

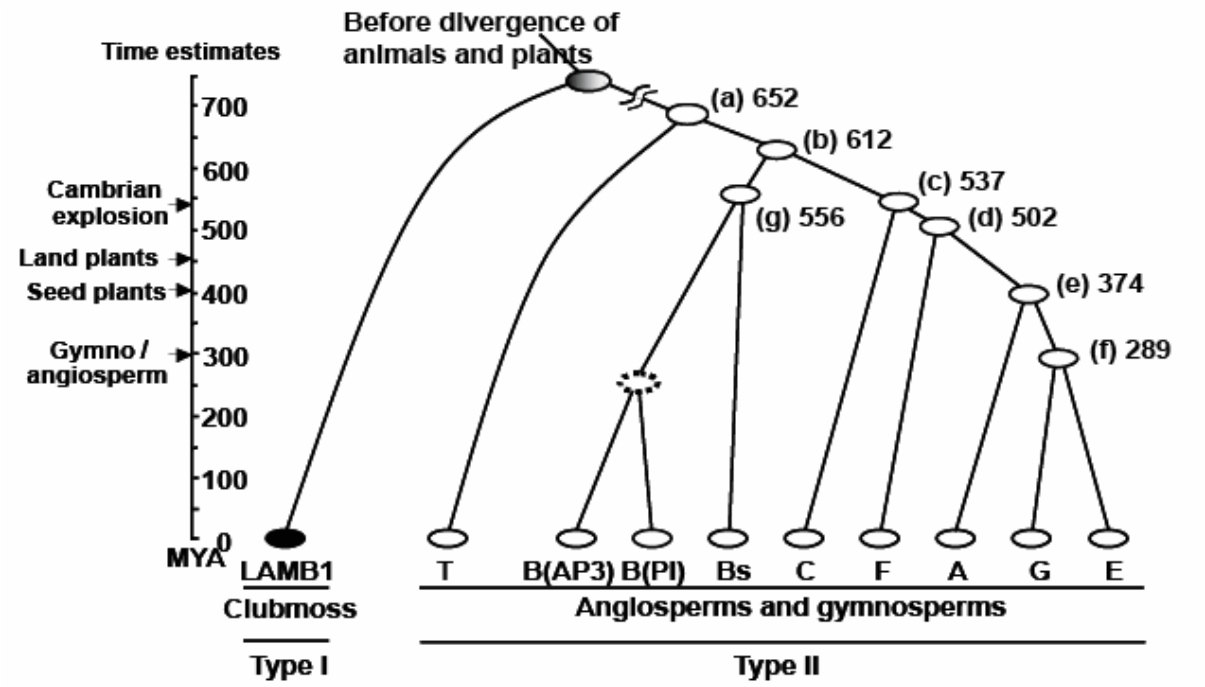


Figure 2.4. Schematic representation of the evolution of floral MADS-box genes. Divergence time estimates (Ma) are indicated for each node of the tree in figure 2.3*A*. The divergence time for node g was estimated separately (see text). Several important events in plant evolution are indicated to the left of the timescale. The time estimates of these major events are taken from Stewart and Rothwell (1993).

Figure 2.5. Phylogenetic tree of 88 MADS-domain sequences from *Arabidopsis*, rice, gymnosperms, ferns, clubmosses, mosses, and animals. This tree was constructed by the NJ method with *p*-distance for a 55-aa domain. The number for each interior branch is percent bootstrap value (500 resamplings), and only values greater than 50% are shown. The names of plant species used are the same as those of figure 2.2 except for ferns and mosses, and those of the remaining species are as follows: fern, *Ceratopteris richardii*; moss, *Physcomitrella patens*; human, *Homo sapiens*; mouse, *Mus musculus*; zebrafish, *Danio rerio*; worm, *Caenorhabditis elegans*; mosquito, *Anopheles gambiae*; fly, *Drosophila melanogaster*.

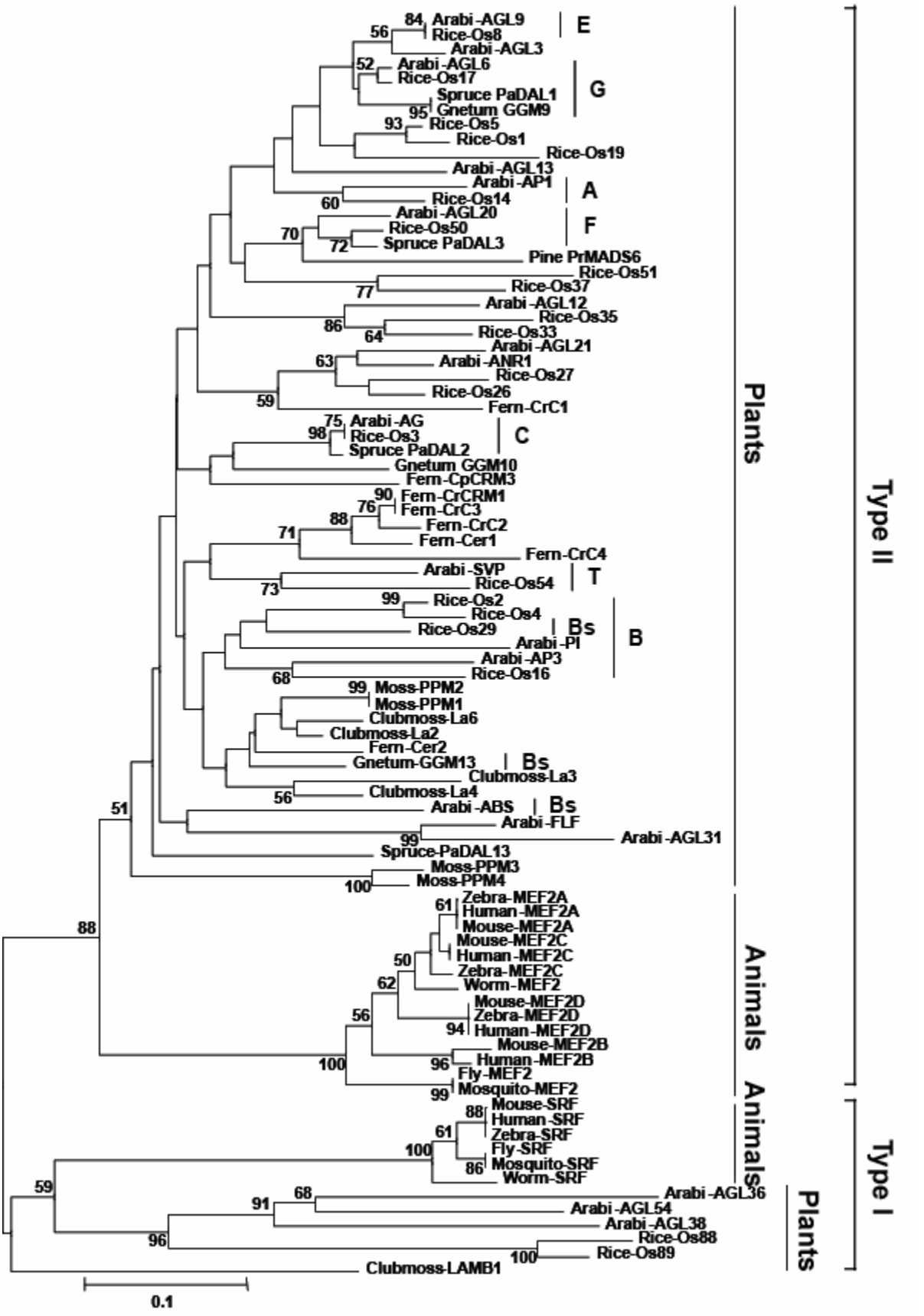
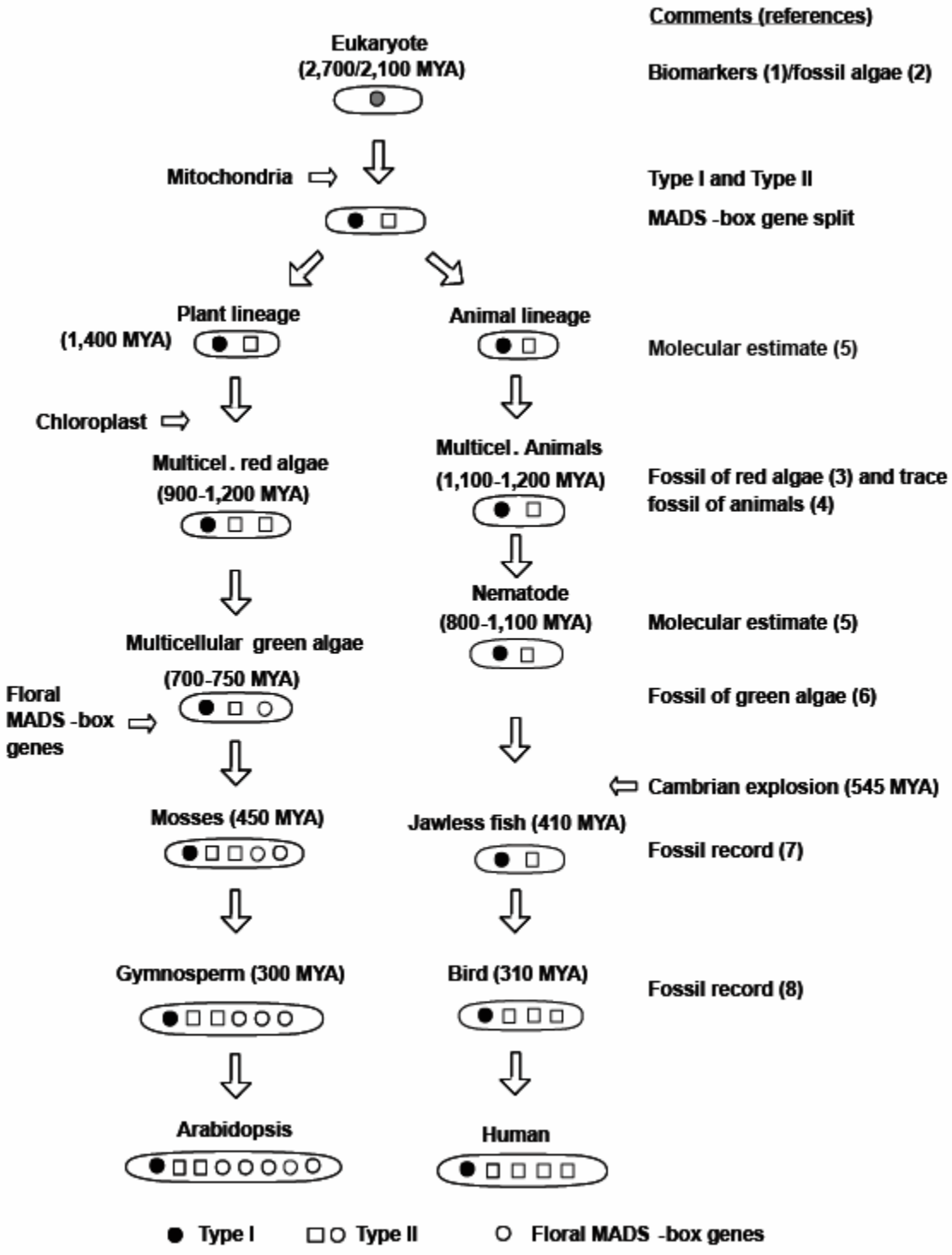

Arabi-AGL9
Rice-Os8
Arabi-AGL3
Arabi-AGL6
Rice-Os17
Spruce PaDAL1
Gnetum GGM9
Rice-Os5
Rice-Os1
Rice-Os19
Arabi-AGL13
Arabi-AP1
Rice-Os14
Arabi-AGL20
Rice-Os50
Spruce PaDAL3
Pine PrMADS6
Rice-Os51
Rice-Os37
Arabi-AGL12
Rice-Os35
Rice-Os33
Arabi-AGL21
Arabi-ANR1
Rice-Os27
Rice-Os26
Fern-CrC1
Arabi-AG
Rice-Os3
Spruce PaDAL2
Gnetum GGM10
Fern-CpCRM3
Fern-CrCRM1
Fern-CrC3
Fern-CrC2
Fern-Cer1
Fern-CrC4
Arabi-SVP
Rice-Os54
Rice-Os2
Rice-Os4
Rice-Os29
Arabi-PI
Arabi-AP3
Rice-Os16
Moss-PPM2
Moss-PPM1
Clubmoss-La6
Clubmoss-La2
Fern-Cer2
Gnetum-GGM13
Clubmoss-La3
Clubmoss-La4
Arabi-ABS
Arabi-FLF
Arabi-AGL31
Spruce-PaDAL13
Moss-PPM3
Moss-PPM4
Zebra-MEF2A
Human-MEF2A
Mouse-MEF2A
Mouse-MEF2C
Human-MEF2C
Zebra-MEF2C
Worm-MEF2
Mouse-MEF2D
Zebra-MEF2D
Human-MEF2D
Mouse-MEF2B
Human-MEF2B
Fly-MEF2
Mosquito-MEF2
Mouse-SRF
Human-SRF
Zebra-SRF
Fly-SRF
Mosquito-SRF
Worm-SRF
Arabi-AGL36
Arabi-AGL54
Arabi-AGL38
Rice-Os88
Rice-Os89
Clubmoss-LAMB1
0.1
E
G
A
F
C
T
Bs
B
Bs
Bs
Plants
Animals
Animals
Plants
Type II
Type I

Figure 2.6. A scenario of the evolution of MADS-box genes in plant and animal lineages. Important events of plant and animal evolution (divergence from the lineage leading to *Arabidopsis* or human) are presented with their estimated times. The references for these estimates are as follows: (1) time of the oldest biomarkers of eukaryotes (Brocks et al. 1999), (2) oldest fossil record of eukaryotic algae (Han and Runnegar 1992), (3) fossil record of some forms of red algae (Butterfield 2000), (4) trace fossil of animals (Seilacher, Bose, and Pfluger 1998; Rasmussen et al. 2002), (5) molecular time estimates of the animal/plant split and nematode evolution (Feng, Cho, and Doolittle 1997; Wang, Kumar, and Hedges 1999; Nei, Xu, and Glazko 2001), (6) fossil record of green algae (Chen and Xiao 1991), (7) fossil record of jawless fish (Maisey 1996, pp. 52-55), and (8) fossil record of the bird/mammal split (Benton 1993, pp. 717-771). The number of circles and squares are not representing the real gene number in each organism. The estimated numbers of MADS-box genes in the species of available genome sequences are as follows: *Arabidopsis* (>70 genes), rice (>70 genes), human (5 genes), fly (2 genes), nematode (2 genes), and fission yeast (4 genes).


Comments (references)
Eukaryote
(2,700/2,100 MYA)
Biomarkers (1)/fossil algae (2)
Mitochondria
Type I and Type II
MADS -box gene split
Plant lineage
(1,400 MYA)
Animal lineage
Molecular estimate (5)
Chloroplast
Multicel. red algae
(900-1,200 MYA)
Multicel. Animals
(1,100-1,200 MYA)
Fossil of red algae (3) and trace fossil of animals (4)
Nematode
(800-1,100 MYA)
Molecular estimate (5)
Multicellular green algae
(700-750 MYA)
Floral MADS -box genes
Fossil of green algae (6)
Cambrian explosion (545 MYA)
Mosses (450 MYA)
Jawless fish (410 MYA)
Fossil record (7)
Gymnosperm (300 MYA)
Bird (310 MYA)
Fossil record (8)
Arabidopsis
Human
Type I
Type II
Floral MADS -box genes

CHAPTER 3

TYPE I MADS-BOX GENES HAVE EXPERIENCED FASTER BIRTH-AND-DEATH EVOLUTION THAN TYPE II MADS-BOX GENES IN ANGIOSPERMS

SUMMARY

Plant MADS-box genes form a large gene family for transcription factors and are involved in various aspects of developmental processes including flower development. They are known to be subject to birth-and-death evolution, but the detailed features of this mode of evolution remain unclear. To gain a deeper insight into the evolutionary pattern of this gene family, we enumerated all available functional and nonfunctional (pseudogene) MADS-box genes from the Arabidopsis and rice genomes. Plant MADS-box genes can be classified into type I and type II genes on the basis of phylogenetic analysis. Conducting extensive homology search and phylogenetic analysis, we found 64 presumed functional and 37 nonfunctional type I genes and 43 presumed functional and 4 nonfunctional type II genes in Arabidopsis. We also found 24 presumed functional and 6 nonfunctional type I genes and 47 presumed functional and 1 nonfunctional type II genes in rice. Our phylogenetic analysis indicated that there were at least about 4 ~ 8 type I genes and about 15 ~ 20 type II genes in the most recent common ancestor (MRCA) of Arabidopsis and rice. It has also been suggested that type I genes have experienced a higher rate of birth-and-death evolution than type II genes in angiosperms. Furthermore, the higher rate of birth-and-death evolution in type I genes appeared partly due to a higher frequency of segmental gene duplication and weaker purifying selection in type I genes than in type II genes.

INTRODUCTION

Morphological/physiological evolution of organisms has been driven mainly by the evolution of genetic toolkits for developmental/physiological processes such as transcription factors and signaling pathways (Carroll, Grenier, and Weatherbee 2001). A large proportion of genetic toolkits are highly conserved even between distantly related organisms. In flowering plants (angiosperms), MADS-box genes are one of such toolkits that control various aspects of developmental processes. MADS-box genes are defined by the highly conserved 180 base pairs (bp) long motif called the MADS-box and are found in animals, fungi, and plants (Shore and Sharrocks 1995). The protein region encoded by the MADS-box is called the MADS-domain (or M-domain) and is part of the DNA-binding domain. It has been proposed that there are at least two evolutionary lineages (type I and type II) of MADS-box genes in animals, fungi, and plants (Alvarez-Buylla et al. 2000b) (figure 3.1).

There are approximately 100 MADS-box genes in *Arabidopsis thaliana* (hereafter called Arabidopsis) and more than 70 MADS-box genes in *Oryza sativa* (hereafter called rice). There are approximately 40 clearly identifiable type II MADS-box genes in each of Arabidopsis (Kofuji et al. 2003; Parenicova et al. 2003) and rice (Lee et al. 2003). Most of the plant type II genes contain three additional plant-specific domains: intervening (I) domain (~30 codons), keratin-like coiled-coil (K) domain (~70 codons), and C-terminal (C) domain (variable length) (Ma, Yanofsky, and Meyerowitz 1991) (figure 3.1). These genes are called MIKC-type genes. The MIKC-type genes can further be divided into two types based on the intron-exon structure; MIKC^{c}-type and MIKC*-type genes (Henschel

et al. 2002). The MIKC^c-type genes have been identified in most major evolutionary lineages of green plants such as angiosperms, gymnosperms, ferns, and mosses (Becker and Theissen 2003). The MIKC*-type genes were originally found in mosses and clubmosses (Svensson, Johannesson, and Engstrom 2000; Henschel et al. 2002), but these genes are also present in Arabidopsis (Kofuji et al. 2003). By contrast, the type I MADS-box genes in plants do not encode the K-domain and are sometimes called M-type genes (Kofuji et al. 2003).

It has been shown that at least 11 classes of MADS-box genes are shared between Arabidopsis and rice/maize (Munster et al. 2002; Becker and Theissen 2003). All of them are MIKC^c-type genes, and their expression patterns have been studied intensively in eudicots. Several classes of MIKC^c-type genes called floral MADS-box genes are concerned with the development of floral components (organs) such as petals, sepals, stamens, and carpels and regulation of flowering time (Weigel and Meyerowitz 1994; Theissen 2001). Other classes of MIKC^c-type genes play diverse roles during vegetative growth (Zhang and Forde 1998; Michaels and Amasino 1999; Alvarez-Buylla et al. 2000a) and fruit development (Liljegren et al. 2000). Some of the floral MADS-box genes in monocots have functions equivalent to those of their orthologs in eudicots (Kang et al. 1998; Ma and dePamphilis 2000), suggesting an ancient origin of the machinery of flower development. There are also a few other genes that are not shared (lineage-specific) between Arabidopsis and rice (Becker and Theissen 2003). The functions of MIKC*-type and M-type (or type I) genes are poorly understood.

MADS-box genes are important regulators of development of angiosperms (and probably nonflowering plants as well), and therefore the study of evolution of MADS-

box genes is expected to give important clues for understanding the morphological evolution of plants. In our previous paper (Nam et al. 2003), we indicated that the MADS-box gene family has been subject to the model of birth-and-death evolution, in which new genes are generated by gene duplication and some duplicate genes stay in the genome as differentiated genes whereas others are inactivated into pseudogenes or deleted from the genome (Nei 1969a; Nei, Gu, and Sitnikova 1997). Although it has been discussed that type I and type II genes might have experienced different modes of gene duplication (De Bodt et al. 2003), little is known about the detailed evolutionary process of MADS-box genes. Here we investigate the pattern of birth-and-death evolution in the MADS-box gene family using all available MADS-box functional genes and pseudogenes from Arabidopsis and rice.

MATERIALS AND METHODS

Identification of MADS-box functional genes and pseudogenes

In this paper, we will assume that the annotated genes that encode a complete MADS-domain are functional genes and the other genes are pseudogenes. To find functional proteins with MADS-domain from Arabidopsis, we performed the PSI-BLAST search (Altschul et al. 1997) with an E-value of $\leq 10^{-5}$ against the entire annotated proteins of Arabidopsis downloaded from the GenBank (as of Dec. 2002). We used 149 MADS-domain sequences from Arabidopsis, rice, animals, and fungi as queries. Similarly, we searched for functional MADS-box genes from annotated proteins of rice that are available from TIGR. Because annotation of rice gene was still in progress at the time of this study, we ourselves conducted gene annotation by using the computer program FGENESH (http://www.softberry.com) from the genome sequences obtained from TIGR and Rice Genome Database of China (Yu et al. 2002).

In order to screen for pseudogenes from Arabidopsis, we first masked all annotated MADS-box gene loci (105 loci) in the genome sequence of Arabidopsis. We then performed the TBLASTN search with an E-value of $\leq 10^{-5}$ against every possible reading frame of this MADS-masked genome with all MADS-domain protein sequences (105 sequences) from Arabidopsis as queries.

A similar search as that of Arabidopsis genes was performed to screen for MADS-box pseudogenes in rice. However, there could be a number of artificially fragmented genes by assembling errors in BAC/PAC clone sequences, resulting in a high false positive rate of pseudogenes. For this reason we used only MADS-domains to

screen for pseudogenes and regarded a gene as a pseudogene if there is at least one stop codon in the MADS-box. (See File 1, which is published as supporting information on the PNAS web site for all the sequences described above and their genomic locations.)

Phylogenetic analysis

Type II proteins generally contain both MADS- and K-domains that can be used for phylogenetic analysis. For this reason, type II genes were analyzed by using the "M- and K-domains" data set. When we constructed a tree for the entire type I and type II genes, we used "M-domains" data set, because type I proteins do not contain the K-domain. The classification of the MIKC-type proteins was made on the basis of the Hidden Markov Model (HMM) search using the computer program HMMer (Eddy 2001) and a K-domain matrix (see File 2, which is published as supporting information on the PNAS web site).

Protein sequences in each data set were aligned by using the computer program MAFFT (Katoh et al. 2002) with the FFT-NS-i option. We then constructed neighbor-joining (NJ) trees (Saitou and Nei 1987) using the computer program MEGA2 (version 2.1) (Kumar et al. 2001). In addition, we constructed a maximum-parsimony (MP) consensus tree of the sequences in the "M- and K-domains" data set using the PAUP* program with the TBR search with 100 bootstrap resamplings (Swofford 1998). Because our data set contained a large number of sequences, we did not use maximum-likelihood method. We also did not use Bayesian phylogenetics, because this method often gives excessively high posterior probabilities even for wrong topologies (Suzuki, Glazko, and Nei 2002; Cummings et al. 2003; Misawa and Nei 2003).

RESULTS

MADS-box functional genes and pseudogenes in Arabidopsis and rice

Our homology search in Arabidopsis initially detected 105 functional MADS-domain protein sequences and 43 MADS-box pseudogenes. Sixteen of these 43 pseudogenes contained the MADS-box. After correcting two possibly misannotated pseudogenes, we finally found 107 functional genes and 41 pseudogenes. Seven out of these 41 pseudogenes were annotated pseudogenes in the GenBank.

In rice, we identified 71 nonredundant functional MADS-box genes. Our preliminary homology search also identified 7 pseudogenes that contained stop codons in the MADS-box. Because the complete rice genome sequence is not publicly available at the present time, the numbers of MADS-box functional genes and pseudogenes may increase in the future.

Of the 178 functional MADS-domain proteins from Arabidopsis and rice, our HMM search detected K-domains in 39 and 37 sequences from Arabidopsis and rice, respectively. These sequences were included in the "M- and K-domains" data set, and all MADS-domains of 178 functional MADS-box genes and the MADS-domains of 21 pseudogenes were included in the "M-domains" data set.

Number of ancestral MADS-box genes in the MRCA of Arabidopsis and rice

Because phylogenetic trees of type II genes were more reliable than those of type I genes, we first inferred the number of ancestral type II genes. Figure 3.2 shows the evolutionary relationships of 79 type II genes from Arabidopsis, rice, mosses, and clubmosses. This tree has low bootstrap supports (< 50%) for deep interior branches that

determine interclade relationships. To infer the number of ancestral type II genes, however, this phylogenetic tree is quite informative. There are several clades including Arabidopsis and rice genes that are supported by a bootstrap value of ≥ 50%. We will call them "shared clades". The genes in a "shared clade" are likely descendents of an ancestral MADS-box gene in the MRCA of Arabidopsis and rice. Thus, the number of "shared clades" is a minimum estimate of the number of MADS-box genes in the MRCA.

Figure 3.2 shows that there are 11 such shared clades, and most type II genes are members of these shared clades. A clade for so-called Bs genes was supported by a low bootstrap value (45%). However, we will consider this clade as a shared clade, because in a previous study they appeared to be monophyletic (Becker et al. 2002). Therefore, 12 shared clades of type II genes were identified. Of these, 11 clades (classes A, B-AP3, B-PI, Bs, C/D, E or AGL2, F or AGL20, G or AGL6, T or SVP, AGL12, and ANR1) were previously reported as different shared clades (Munster et al. 2002; Becker and Theissen 2003; Nam et al. 2003) and belong to the MIKC^c-type. The simplified class names F, G, and T are used according to Nam et al. (20). Class S is a novel shared gene class, and the genes in this class appear to be orthologous to the MIKC*-type genes from the moss. There are also three groups of genes (member genes of class E, F, and ANR1) that are not shared but are sister groups for three "shared clades" belonging to class E, class F, and class ANR1, respectively. Because these sister relationships are reasonably well supported (≥ 65%), it appears that at least two ancestral genes existed in each of classes E, F, and ANR1 in the MRCA of Arabidopsis and rice. Identification of 12 shared clades and 3 sister clades suggests that there were at least 15 ancestral type II MADS-box genes in the MRCA. A similar result was obtained from the MP tree (data not shown).

There are other type II genes (class FLC genes, *AGL15*, *AGL18*, and *OsMADS32*) that are not members of the above 15 clades. Orthologs of each of these genes might have been lost in either the Arabidopsis or the rice lineage, or our phylogenetic analysis could not resolve their evolutionary relationships. If the former is the case, the number of ancestral type II genes can be approximately 20 in the MRCA of Arabidopsis and rice. Of course, we cannot exclude the possibility that this number is an underestimate because of the incomplete genome sequencing in rice.

Figure 3.3 shows the phylogenetic tree of all MADS-box genes from Arabidopsis and rice and some of type I and type II MADS-box genes from animals (223 sequences). According to this tree, all MADS-box genes that encode detectable K-domains form a clade together with animal type II genes, though they are not statistically well supported. Interestingly, there are also 13 MADS-box genes that do not encode an intact K-domain (at least in their predicted ORFs, genomic sequences, and adjacent genomic sequences) but are apparently very similar to MIKC-type (type II) genes. These genes might have lost the K-box during the evolution, or the absence or fragmentation of the K-box could be due to assembling error of genome sequences. In this paper these genes will be included in the type II genes on the basis of their close evolutionary relationships with other type II genes. The remaining MADS-box genes appear to have diverged from type II genes before the animal/plant split, though the bootstrap support is weak. These genes correspond to the type I genes proposed by Alvarez-Buylla et al. (Alvarez-Buylla et al. 2000b). Figure 3.3 also suggests that at least one type I gene existed in the MRCA of animals and plants. This observation is consistent with that of other researchers (Alvarez-Buylla et al. 2000b; Becker and Theissen 2003; De Bodt et al. 2003).

Most of the shared classes observed in figure 3.2 remain unchanged in the original tree of figure 3.3 (data not shown). We also identified another novel shared clade (P) supported by a bootstrap value of 79%. Class P genes from Arabidopsis were previously classified as type I by Alvarez-Buylla et al. (Alvarez-Buylla et al. 2000b). However, we also observed that MIKC*-type genes from mosses and class S and class P genes from Arabidopsis and rice formed a clade, when we used different sets of genes (data not shown). It is therefore possible that class P genes are also closely related to the MIKC*-type (type II) genes from mosses as proposed by other researchers (De Bodt et al. 2003; Kofuji et al. 2003; Parenicova et al. 2003), though they are not orthologous to the latter genes. On the basis of the tree shown in figure 3.3, the remaining type I genes can further be subdivided into classes Mα, Mβ, and Mγ in agreement with Parenicova et al.'s (Parenicova et al. 2003) classification, though bootstrap supports of these classes are very low and class Mγ genes are not monophyletic. Although our classification of type I genes is very crude, it suggests that at least about 4 ~ 8 ancestral type I genes existed in the MRCA of Arabidopsis and rice. The numbers of functional genes and ancestral genes estimated in this way for each type of MADS-box genes in Arabidopsis and rice are shown in figure 3.3 (see numbers in parentheses).

Our study of ancestral MADS-box genes therefore leads to the hypothesis that there were at least about 15 ~ 20 type II genes and at least about 4 ~ 8 type I genes in the MRCA of Arabidopsis and rice. Because there are 43 type II genes and 64 type I genes in Arabidopsis, the results of the present study suggest that type I genes have experienced a higher birth rate than type II genes in the Arabidopsis lineage. A similar pattern was also observed in rice, though it is preliminary. In addition, this pattern is quite general across

most gene classes except class FLC in Arabidopsis and class AGL12 in rice (see numbers in parentheses in figure 3.2 and figure 3.3). One may argue that if we use more stringent criteria for estimating the number of ancestral type I genes, the number may change, and therefore the rate of gene birth would change. However, this does not affect our conclusion that type I genes have experienced a higher birth rate than type II genes. This is because many type I genes in each of classes Mα, Mβ, and Mγ from either Arabidopsis or rice appear to be monophyletic, suggesting that they were duplicated after the Arabidopsis and rice split.

Classification of MADS-box pseudogenes

Existence of pseudogenes means that functional genes die sometimes in the evolutionary process. To examine whether there are differences in the death rate among different types of MADS-box genes, we classified pseudogenes on the basis of sequence similarity to functional MADS-box genes. In Arabidopsis 4 pseudogenes were most similar to the type II genes (see Table 3.1, which is published as supporting information on the PNAS web site), and none of these pseudogenes had the MADS-box. The remaining 37 pseudogenes were most similar to the type I genes. Fourteen of these 37 pseudogenes had the MADS-box. In the case of rice, only one of the 7 pseudogenes belonged to the type II, and the remainder were type I genes. These results show that the proportion of pseudogenes is significantly different between type II and type I genes in both Arabidopsis and rice. When we applied the same criterion of pseudogenes as that of rice pseudogenes (existence of stop codons in the MADS-box), we detected 9 type I pseudogenes and no type II pseudogenes in Arabidopsis. Our homology search and

phylogenetic analysis also showed that several pseudogenes belonging to class Mα are monophyletic (see figure 3.3), suggesting that the number of pseudogenes has increased recently in this lineage. Even if we exclude such lineage-specific pseudogenes, the difference in the proportion of pseudogenes between type I and type II genes is still substantial. Although type I genes are expected to include more pseudogenes than the type II genes because of their higher birth rate, this factor alone does not explain the difference in pseudogenes between type I and type II genes. Therefore, type I genes should have had a higher death rate than type II genes.

It is not easy to have an unambiguous definition of pseudogenes, because even a fragmentary gene can be functional (Chen et al. 2002; Shin et al. 2002) and young pseudogenes may not be distinguishable from functional genes. Therefore, different criteria for pseudogenes may change our conclusion about the death rates of type I and type II genes. As mentioned above, however, our conclusions about the death rates based on two different criteria in Arabidopsis are essentially the same. Note also that our searches for pseudogenes are apparently biased for pseudogenes similar to more conserved functional genes (type II genes in this study) than for less conserved functional genes. Therefore, our conclusion about the difference in death rate between type I and type II genes is conservative.

Genomic organization of MADS-box genes in Arabidopsis

The genomic locations of all MADS-box genes in Arabidopsis are shown in figure 3.4. In general, MADS-box genes are scattered all over the chromosomes. However, we also observed a number of clusters of closely located MADS-box genes in

Arabidopsis. Most of these genes belonged to type I genes, and in general the genes in each cluster are evolutionarily closely related. These closely related MADS-box genes were probably generated by recent segmental duplication. The genomic locations of pseudogenes are also shown in figure 3.4. Most pseudogenes are closely located to each other as well as to their closely related functional MADS-box gene, though there are several exceptions. We also found a genomic cluster of type I pseudogenes without any functional MADS-box genes (but there are other genes) on chromosome 3 (genes with vertical bars in figure 3.4). The genomic locations and the phylogenetic tree of MADS-domain sequences (figure 3.3) suggest that this gene cluster was formed by segmental duplication of an ancestral pseudogene cluster, which was in turn duplicated from another pseudogene cluster on chromosome 2 (genes with gray bars in figure 3.4).

Rice MADS-box genes are also scattered all over the chromosomes, and more clusters of type I genes were found than those of type II genes (our unpublished data).

DISCUSSION

We have seen that type I genes have experienced faster birth-and-death evolution than type II genes in the Arabidopsis and rice lineages. The higher birth rate of type I genes is apparently caused by a higher rate of gene duplication, because duplicate genes generally do not cause harmful effects. In fact, type I genes are associated with a higher frequency of segmental duplications than type II genes in Arabidopsis (see figure 3.4). (We do not think the genome duplication is responsible for the different birth rates of type I and type II genes, because in this case the birth rate should be the same for all genes. Therefore, we will not discuss this factor.) By contrast, the death of functional genes may have harmful effects, and therefore the death rate may be influenced by functional requirements of duplicate genes as well as genomic events and fixation by genetic drift. Our estimates of the numbers of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) and synonymous nucleotide substitutions per synonymous site (d_S) suggested that type I genes have been under weaker purifying selection than type II genes (see figure 3.5). This observation may explain why type I genes have experienced a higher death rate than type II genes, because the death of type I genes could be less harmful than that of type II genes. It is possible that, after duplication, type II genes became functionally differentiated in a relatively short time and therefore have been maintained as functional genes in the genome. This might be related to the extensive morphological diversification of angiosperms.

Although type I genes are apparently under weaker purifying selections than type II genes, they still might have played some important roles, because most of the recently duplicated type I gene pairs show significantly lower d_N than d_S (figure 3.5). Recently, it

has been proposed that the expression of a type I MADS-box gene, *PHERES1*, in Arabidopsis is associated with seed abortion in a certain mutant background (Kohler et al. 2003). However, the functions of type I genes are not well understood. If there are functionally redundant duplicate genes, it would be difficult to study their functions by mutagenesis experiments. Moreover, if type I genes are involved in a short period of developmental processes, it may also be difficult to study their functions. At the present time, gaining insights into the functional constraints of type I genes by evolutionary analysis may be of some help for future experimentation. Our study suggests that type I genes may be more variable among different angiosperm species than type II genes because of faster birth-and-death evolution than that of type II genes. In addition, type I genes are generally less conserved than type II genes.

There are a substantial number of type II duplicate genes, though the birth rate of type II genes is lower than that of type I genes. Therefore, some extent of functional redundancy or differentiation is expected to be observed among highly similar type II genes. For example, three class E genes (*AGL2*/*4*/*9* or *SEP1*/*2*/*3*) in Arabidopsis are known to be functionally redundant, because single gene mutations showed only subtle phenotypic changes, while triple mutants showed significant phenotypic changes in flowers of Arabidopsis (Pelaz et al. 2000). Nevertheless, our d_N and d_S analysis suggests that these genes are generally subject to strong purifying selection (File 3). Therefore, more careful study of single gene mutations may reveal some unrecognized phenotypic effects in plants. Moreover, there are substantial conservation or differentiation in gene expressions (Kofuji et al. 2003; Parenicova et al. 2003) and in protein coding region (Immink et al. 2003; Lamb and Irish 2003; Vandenbussche et al. 2003) among

paralogous MADS-box genes. By combining experimental studies with evolutionary analyses, we may be able to have a better insight into gene functions.

Table 3.1. Numbers of ancestral genes, functional genes, and pseudogenes for two types of MADS-box genes in Arabidopsis and rice

Type	Class (# of ancestral genes)	Functional genes		Pseudogenes	
		Arabidopsis	Rice	Arabidopsis	Rice
Type I	**E (2)**	4	6	-	-
	G (1)	2	2	-	-
	A (1)	4	4	-	-
	F (2)	6	7	-	-
	C/D (1)	4	3	-	1
	AGL12 (1)	1	5	-	-
	Bs (1)	1	3	-	-
	B-PI (1)	1	2	-	-
	B-AP3 (1)	1	1	-	-
	ANR1 (2)	4	6	-	-
	T (1)	2	4	1	-
	S (1)	3	2	-	-
	FLC (1?)	6	-	1	-
	U (5 ~ 7?)	5	2	2	-
Subtotal		43	47	4	1
Type II	**Mα (1 ~ 4)**	20	15	23	2
	Mβ (1)	17	-	5	1
	Mγ (1 ~ 2)	21	8	8	3
	P (1)	4	1	1	-
Subtotal		64	24	37	6
Total		107	71	41	7

NOTE.—Total number of MADS-box genes in Arabidopsis includes 2 presumably misannotated genes. The class of a pseudogene refers to that of a functional gene that is most similar to the pseudogenes. The rice genome sequence is still incomplete, so the number of genes may increase in the future. Note that the criteria of pseudogene for Arabidopsis and rice genes are different (see materials and methods).

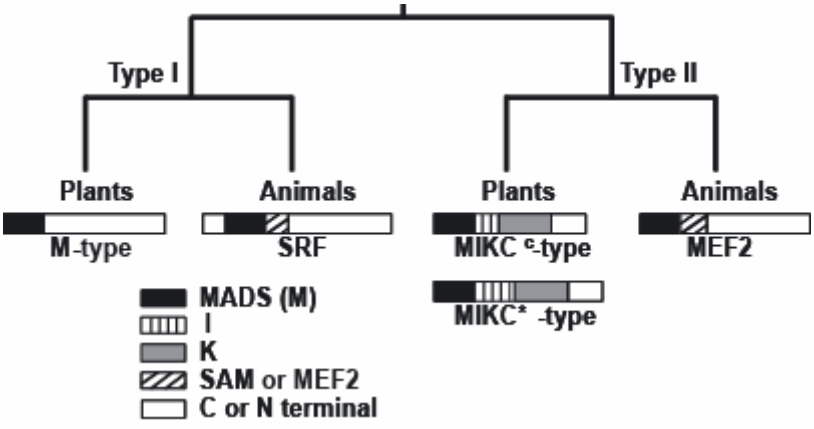


Figure 3.1. Domain structures of types I and II MADS-box genes in plants and animals. Adapted from ref. (Alvarez-Buylla et al. 2000b) about the structures of type I and type II genes and from ref. (Henschel et al. 2002) about the structures of MIKC^c-type and MIKC*-type genes.

Figure. 3.2. Phylogenetic tree of 79 MIKC-type (type II) genes from Arabidopsis, rice, mosses, and clubmosses. This tree was constructed by the NJ method with PC-distance. One hundred five amino acids were used after all alignment gaps were eliminated. The number for each interior branch is the percent bootstrap value (500 resamplings), and only the values greater than 50% are shown. The scale bar indicates the number of amino acid substitutions per site. We generally followed ref. (Parenicova et al. 2003) about the notations of Arabidopsis genes. Simplified class names following reference (Nam et al. 2003) were used except for classes A, B-AP3, B-PI, C/D, E, FLC, and S. The genes marked with (U) are unassigned genes for any classes. The three numbers in parentheses below each class name refer to the numbers of ancestral MADS-box genes, MADS-box genes in Arabidopsis, and MADS-box genes in rice, respectively. Two MIKC*-type genes (*PPM3* and *PPM4*) from the moss *Physcomitrella patens* (Svensson, Johannesson, and Engstrom 2000) and one MIKC*-type gene (*LAMB1*) from the clubmoss *Lycopodium annotinum* (Henschel et al. 2002) were used as reference sequences.

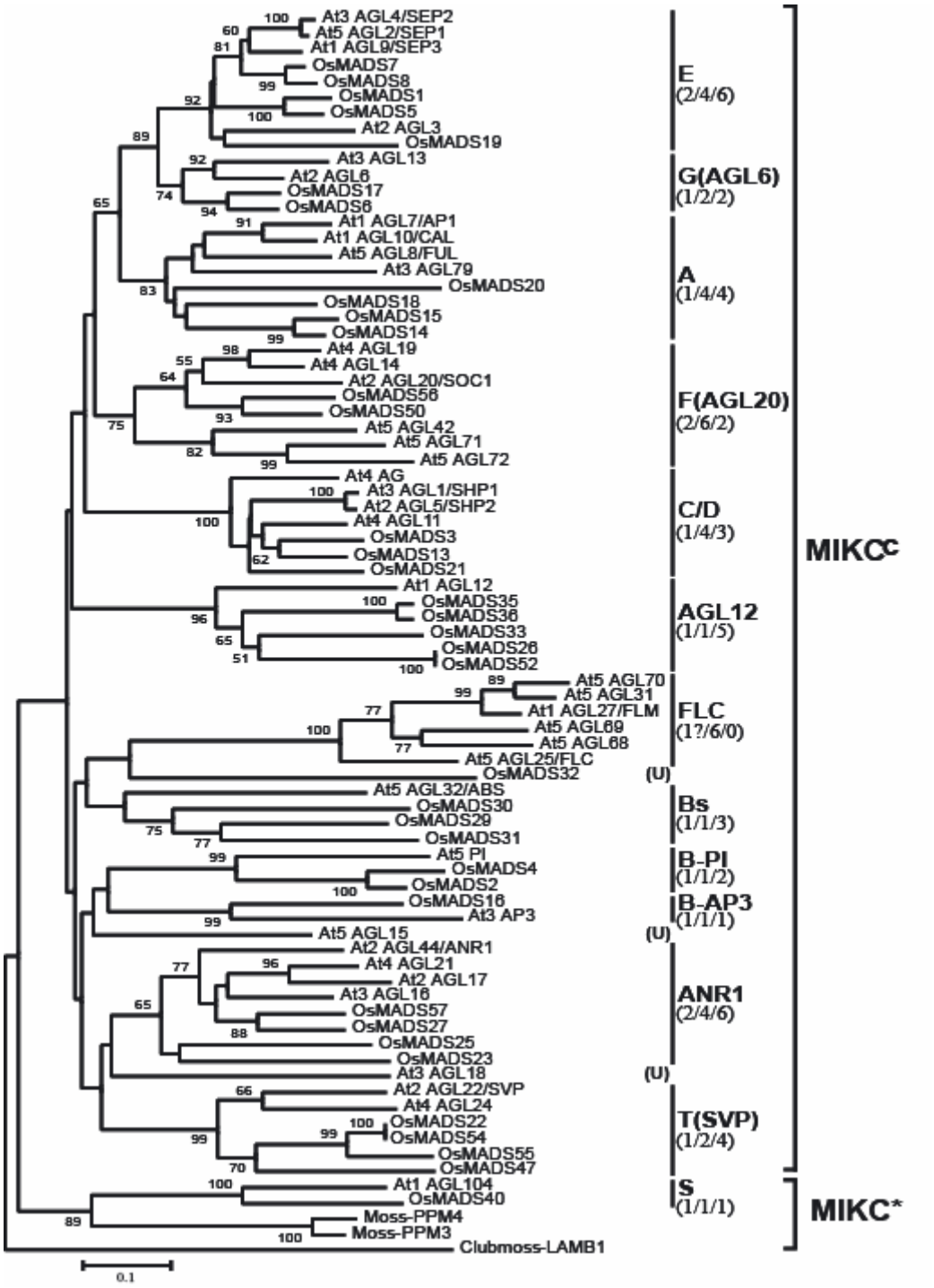
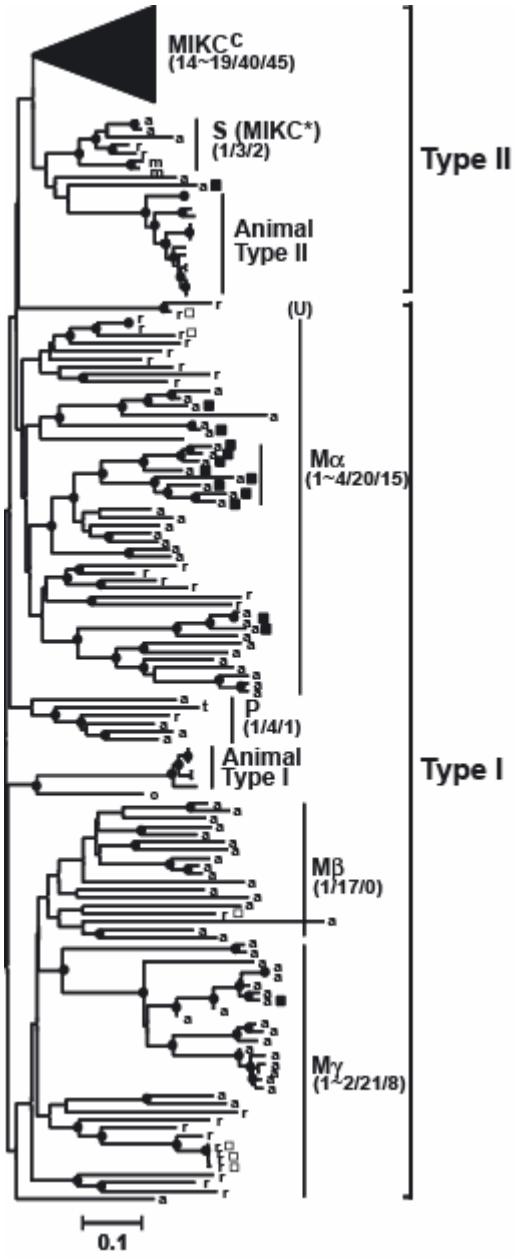

At3 AGL4/SEP2
At5 AGL2/SEP1
At1 AGL9/SEP3
OsMADS7
OsMADS8
OsMADS1
OsMADS5
At2 AGL3
OsMADS19
At3 AGL13
At2 AGL6
OsMADS17
OsMADS6
At1 AGL7/AP1
At1 AGL10/CAL
At5 AGL8/FUL
At3 AGL79
OsMADS20
OsMADS18
OsMADS15
OsMADS14
At4 AGL19
At4 AGL14
At2 AGL20/SOC1
OsMADS56
OsMADS50
At5 AGL42
At5 AGL71
At5 AGL72
At4 AG
At3 AGL1/SHP1
At2 AGL5/SHP2
At4 AGL11
OsMADS3
OsMADS13
OsMADS21
At1 AGL12
OsMADS35
OsMADS36
OsMADS33
OsMADS26
OsMADS52
At5 AGL70
At5 AGL31
At1 AGL27/FLM
At5 AGL69
At5 AGL68
At5 AGL25/FLC
OsMADS32
At5 AGL32/ABS
OsMADS30
OsMADS29
OsMADS31
At5 PI
OsMADS4
OsMADS2
OsMADS16
At3 AP3
At5 AGL15
At2 AGL44/ANR1
At4 AGL21
At2 AGL17
At3 AGL16
OsMADS57
OsMADS27
OsMADS25
OsMADS23
At3 AGL18
At2 AGL22/SVP
At4 AGL24
OsMADS22
OsMADS54
OsMADS55
OsMADS47
At1 AGL104
OsMADS40
Moss-PPM4
Moss-PPM3
Clubmoss-LAMB1
0.1
E
(2/4/6)
G(AGL6)
(1/2/2)
A
(1/4/4)
F(AGL20)
(2/6/2)
C/D
(1/4/3)
AGL12
(1/1/5)
FLC
(1?/6/0)
(U)
Bs
(1/1/3)
B-PI
(1/1/2)
B-AP3
(1/1/1)
(U)
ANR1
(2/4/6)
(U)
T(SVP)
(1/2/4)
S
(1/1/1)
MIKCc
MIKC*

Figure. 3.3. Phylogenetic tree of 223 MADS-domain sequences from Arabidopsis, rice, mosses, clubmosses, and animals. This tree was constructed by the NJ method with *p*-distance and the pairwise deletion option (Kumar et al. 2001) of about 55 amino acids. *p*-distance is known to be more efficient in obtaining the correct topology when the sequence length is short (Takahashi and Nei 2000). The genes from Arabidopsis and rice are labeled with "a" and "r", respectively. The reference sequences from the moss *Physcomitrella patens* and the clubmoss *Lycopodium annotinum* are labeled with "m" and "c", respectively. Genes labeled with black squares (■) are pseudogenes from Arabidopsis and those with open squares (□) are pseudogenes from rice. Interior branches with bootstrap values (500 bootstraps) greater than 50% are indicated by black dots (●). The portion of the tree corresponding to the MIKC^c-type genes is compressed, because it is essentially the same as that in figure 3.2. The numbers in parentheses below each class name are the numbers of ancestral MADS-box genes, MADS-box genes in Arabidopsis, and MADS-box genes in rice in this order.


MIKC^C
(14~19/40/45)
S (MIKC*)
(1/3/2)
Animal
Type II
Type II
(U)
Mα
(1~4/20/15)
P
(1/4/1)
Animal
Type I
Type I
Mβ
(1/17/0)
Mγ
(1~2/21/8)
0.1

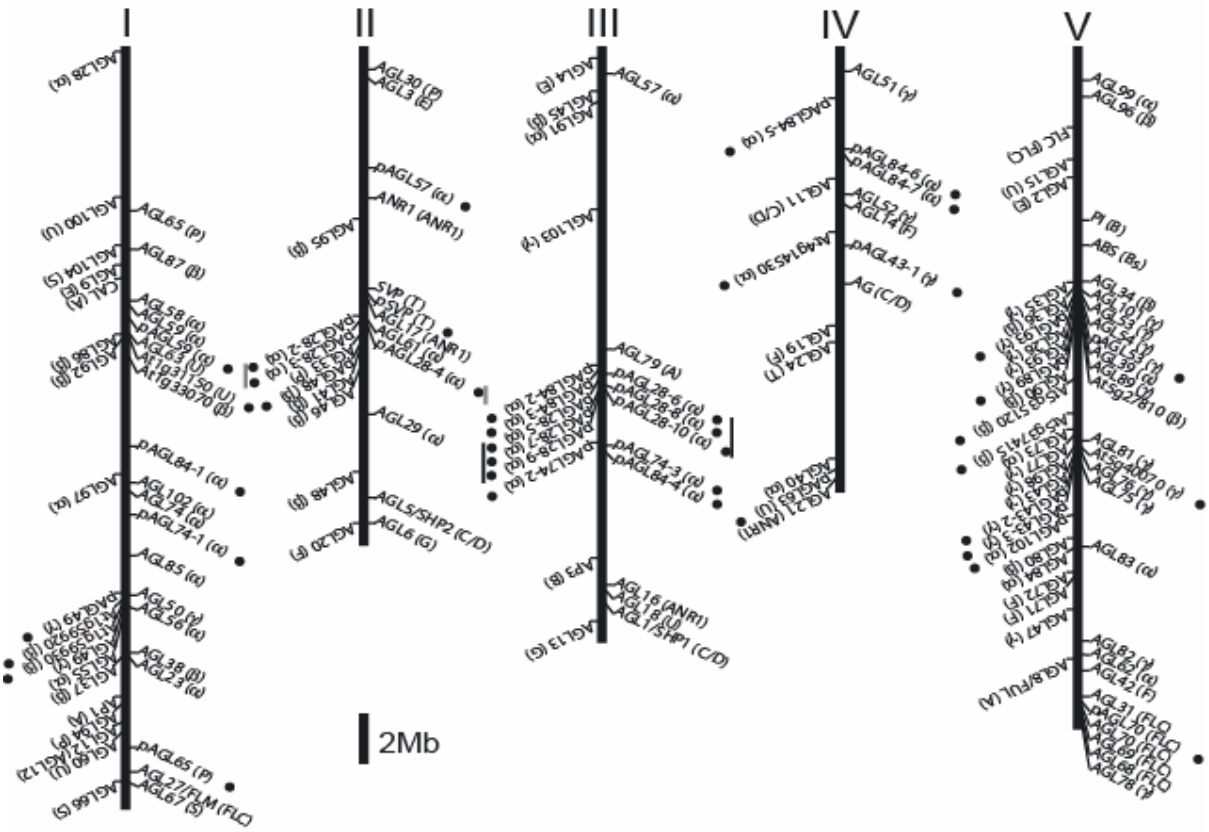


Figure. 3.4. Genomic organization of MADS-box genes in Arabidopsis. Genes with black dots (●) are pseudogenes. For 7 annotated pseudogenes, we used their gene codes from the Genbank. Unannotated pseudogenes are indicated by "p" in front of the name of the functional gene that is most similar to the pseudogene. For example, a pseudogene that is most similar to "*SVP*" is designated as "pSVP". The class name of each gene is given in parentheses at the end of the gene name. Of these class names, "(α)", "(β)", "(γ)", and "(U)" refer to "Mα", "Mβ", "Mγ", and "Unassigned", respectively. I, II, III, IV, and V represent chromosome numbers. The scale bar below chromosome II is for 2 megabase pairs (Mb).

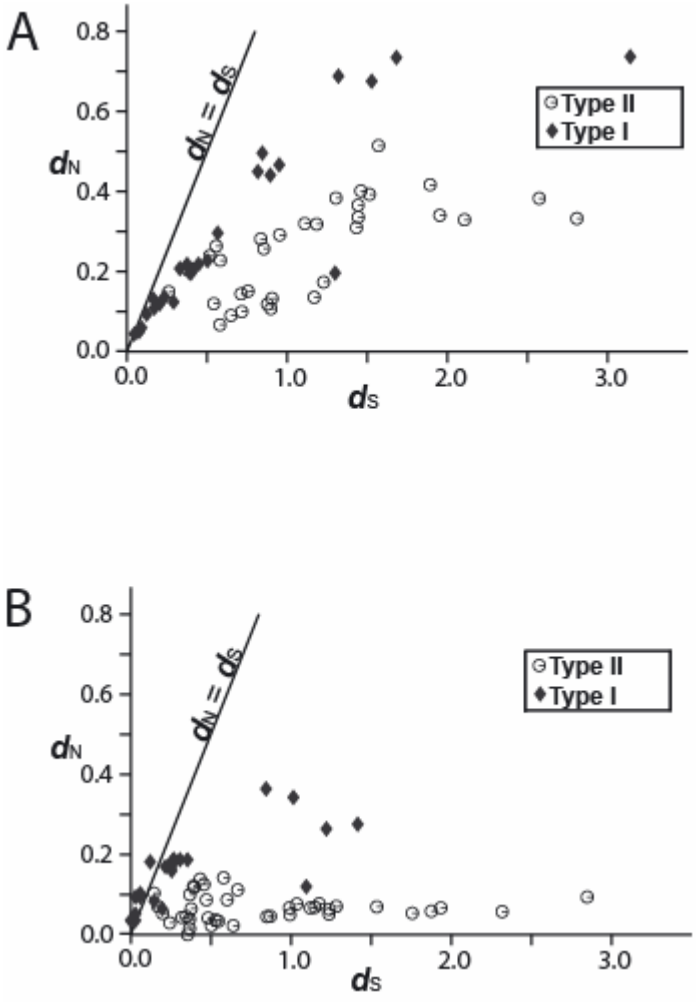


Figure 3.5. Relationships between d_N and d_S for about 60 selected pairs. The d_N and d_S values were estimated by the Nei-Gojobori method (Nei and Kumar 2000). Type I and type II genes are marked with diamonds (♦) and open circles (o), respectively. A linear line is for $d_N = d_S$. A. The entire aligned regions for 59 selected gene pairs. B. Only MADS-box regions for 63 selected gene pairs. The genes used are available from the supplementary materials.

CHAPTER 4

A SIMPLE METHOD FOR PREDICTING THE FUNCTIONAL DIFFERENTIATION OF DUPLICATE GENES AND ITS APPLICATION TO MIKC-TYPE MADS-BOX GENES

SUMMARY

A simple statistical method for predicting the functional differentiation of duplicate genes was developed. This method is based on the premise that the extent of functional differentiation between duplicate genes is reflected in the difference in evolutionary rate because the functional change of genes is often caused by relaxation or intensification of functional constraints. With this idea in mind, we developed a window analysis of protein sequences to identify the protein regions in which the significant rate difference exists. We applied this method to MIKC-type MADS-box proteins that control flower development in plants. We examined 23 pairs of sequences of floral MADS-box proteins from petunia and found that the rate differences for 14 pairs are significant. The significant rate differences were observed mostly in the K domain, which is important for dimerization between MADS-box proteins. These results indicate that our statistical method may be useful for predicting protein regions that are likely to be functionally differentiated. These regions may be chosen for further experimental studies.

INTRODUCTION

The functional differentiation of duplicate genes is thought to be an important mechanism of evolution of organisms (Lewis 1951; Nei 1969b; Ohno 1970; Zhang 2003). This differentiation is often associated with the relaxation or intensification of purifying selection in certain regions of protein sequences. Therefore, comparison of the evolutionary rates of paralogous protein sequences may give some insights into their functional differentiation. With this idea a number of authors have developed statistical methods for predicting functional differentiation by examining the evolutionary rates. Dermitzakis and Clark (Dermitzakis and Clark 2001) suggested that this functional differentiation may be revealed by examining the heterogeneity of substitution rate between two pairs of duplicated genes. Considering two groups of paralogous duplicate proteins (*A* and *B* in figure 4.1A), Gu (Gu 1999) and Knudsen and Miyamoto (Knudsen and Miyamoto 2001) respectively proposed a Bayesian and a maximum likelihood method of detecting amino acid sites that show a significant rate difference between the two groups. In these methods the number of sequences in each group (*A* or *B*) must be relatively large to have reliable results. When the groups *A* and *B* include only one sequence, their methods are not applicable. This is also true with Dermitzakis and Clark's method.

In real experimental studies, it is customary to identify the functional difference by comparing a sequence with known functional domains with a new sequence by using domain swapping or site-directed mutagenesis. However, it is time-consuming and expensive to use this method for a large number of pairs of sequences. It is therefore useful to develop a statistical method for identifying protein domains that are likely to be

functionally differentiated. For this reason, we propose a new method in which only two sequences are compared at a time after construction of a rooted tree. This method will then be illustrated by an analysis of a number of MIKC-type MADS-box genes that control the development of flowers in plants.

METHODS AND RESULTS

Statistical methods

In our method two protein sequences to be compared (*A* and *B* in figure 4.1B) and an outgroup sequence (*C*) will be used after sequence alignment (figure 4.1B). A phylogenetic tree for the sequences is constructed to determine the root of the sequences *A* and *B*. Here we suggest that the *p*-distance (proportion of different amino acids) (Nei and Kumar 2000) be used, because the sequences to be compared are usually closely related and the *p*-distance has a smaller variance than any other distance measure. However, if a pair of divergent sequences is to be tested (e.g. *p*-distance > 0.3), the Poisson-correction or some other distance may be used (Nei and Kumar 2000). To identify the protein regions that show a significant rate difference, we use a sliding window analysis. Let n be the total number of amino acid (codon) sites used and w be the window size (the number of amino acids considered for one window). This window analysis may be done by sliding the window by one amino acid position consecutively or by skipping s amino acid positions each time. The total number of windows to be considered (T) is then $(n - w)/s + 1$. Here T should be an integer. For example, if T happens to be 55.2, it should be reset to 55.

For each window, we now estimate the number of amino acid substitutions a and b for branch (sequence) *A* and *B* in figure 4.1B, respectively. The branch lengths a and b may be estimated by the least squares method and are given by

$$\hat{a} = (d_{AB} + d_{AC} - d_{BC}) / 2,$$

$$\hat{b} = (d_{AB} + d_{BC} - d_{AC}) / 2,$$

where d_{AB}, d_{AC}, etc., are the observed distances between sequences *A* and *B*, *A* and *C*, etc., respectively. We are now interested in testing the significance level of the difference $\hat{a} - \hat{b}$, that is,

$$D = \hat{a} - \hat{b}.$$

The variance of this *D* can be obtained by the formula given in (Takezaki, Rzhetsky, and Nei 1995). We can then consider

$$Z = D / \sqrt{V(D)},$$

where $V(D)$ is the variance of *D*. This *Z* is approximately normally distributed as long as *w* is about 30 or greater. Therefore, the significance level can easily be determined. When $w < 30$, the above *Z* is distributed as the *t* distribution with $w - 1$ degrees of freedom (Rao 1998). In reality, unless $w \geq 30$, the statistical power of the window test is not very high. We therefore recommend that the window size is equal to or greater than 30.

It should be noted that in this sliding window analysis the *Z* values obtained for consecutive windows are highly correlated. Therefore, the significance levels of *Z* values for consecutive window analyses may not be accurate. However, if the *Z* value for one of the windows is significant, one can take it seriously. Furthermore, our purpose is to identify protein regions that should be subjected to experimental tests. Therefore, any consecutive windows showing significant *Z* values should be considered biologically important. Actually, for this purpose, even a region showing *Z* values with a significant level of 10% may be considered for experimentation.

In the above method we considered the case where each of *A*, *B*, and *C* contains only one sequence. However, the above approach can easily be extended to the case where protein sequences are classified into two groups, *A* and *B*, and the average rate

difference between the two groups of proteins is studied. In this case, because the above test is a special case of Takezaki et al.'s (Takezaki, Rzhetsky, and Nei 1995) two-cluster method, we can directly apply the two-cluster method to test the rate difference between the two groups for each window. The outgroup may also contain many sequences. This is true even when *A* and *B* contain one sequence each.

Another extension of the above method is to consider the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) (Nei and Gojobori 1986) or the number of radical nucleotide substitutions (substitutions causing the changes of amino acid charge, hydrophobicity, and size) per site (Hughes, Ota, and Nei 1990; Zhang 2000). At present, however, it is unclear how useful these quantities are, because they generate rather large variances.

Application to MADS-box genes controlling flower development in plants

Floral MIKC-type MADS-box genes encode transcription factors that control flower development in plants. Major floral MADS-box genes can be classified into at least eight classes (Becker and Theissen 2003) in terms of their function and evolutionary relationships, that is, *A*, *B*, *C*, *D*, *E*, *F*, *G*, and *T* classes according to the simplified notation in (Nam et al. 2003). Each of these classes of genes encodes a protein consisting of the MADS (M) domain (DNA-binding site with about 60 amino acids (aa)), intervening (I) domain (~ 30 aa), keratin-like (K) domain (~ 70 aa), and C-terminal (C) domain (variable number of aa) (Ma, Yanofsky, and Meyerowitz 1991) (figure 4.1C). The M domain is composed of DNA-binding α helices, carries a nuclear localization signal and is involved in dimerization of proteins together with the I domain (Shore and

Sharrocks 1995; Riechmann, Wang, and Meyerowitz 1996). The K domain mediates protein-protein interaction, whereas the C domain possesses transcriptional activation function in some MADS-box proteins (Shore and Sharrocks 1995; Fan et al. 1997; Cho et al. 1999; Moon et al. 1999; Honma and Goto 2001) and might also be involved in protein dimerization (Tzeng, Liu, and Yang 2004) or formation of multimeric complexes (Egea-Cortines, Saedler, and Sommer 1999). Among these domains, the I and K domains are most well known for determining the pattern of homodimerization or heterodimerization of MADS-box proteins. The K domain is involved in protein-protein interaction and is characterized by three strings of heptad repeats $(\boldsymbol{abcdefg})_n$ which are potentially forming coiled coils, with hydrophobic amino acids predominantly in positions ***a*** and ***d*** (Fan et al. 1997; Yang, Fanning, and Jack 2003). The proteins encoded by different classes of floral MADS-box genes interact with one another or with some other proteins to form a particular organ. According to the floral Quartet model, the formation of petals is controlled by a combination of tetramers of class A, B, and E proteins, and that of stamens is by tetramers of class B, C, and E proteins (Weigel and Meyerowitz 1994; Ma and dePamphilis 2000; Honma and Goto 2001; Theissen 2001). However, to explain the development of various forms of flowers in different species, we have to know detailed aspects of protein-protein interaction within each class of proteins. For this reason, many experimentalists are now studying protein-protein interaction by using techniques such as yeast two-hybrid analysis, domain swapping, and site-specific mutagenesis.

Immink et al. (Immink et al. 2003) studied the gene expression and protein-protein interaction patterns of 23 floral MADS-box genes in petunia using Northern hybridization and yeast Cytotrap experiments. They identified a number of MADS-box

proteins interacting with each other (see figure 4.2). Their results showed that even closely related MADS-box proteins often have different numbers of protein interaction partners. This suggests that there was some kind of functional differentiation between these MADS-box proteins. We therefore decided to apply our new statistical method for predicting protein regions responsible for the functional differentiation using our Perl script (see http://www.bio.psu.edu/People/Faculty/Nei/Lab/software.htm). We first constructed a phylogenetic tree for the 23 petunia MADS-box protein sequences together with 22 rice and 23 *Arabidopsis* sequences. The rice and *Arabidopsis* sequences were used to classify the petunia sequences into the eight classes of genes mentioned above and to find proper outgroup genes.

Figure 4.2 shows the phylogenetic tree obtained by the neighbor-joining method (Saitou and Nei 1987) for all 68 genes (see the supplementary material for the sequence alignment). Parsimony analysis (Swofford 1998) produced essentially the same tree (see the supplementary material for the maximum-parsimony tree). The topology of this tree for major gene classes is essentially the same as that of our previous trees for floral MADS-box genes (Nam et al. 2003; Nam et al. 2004), and the eight gene classes (A, B, C, D, …, T) form separate monophyletic clades, though class C and D genes often form a mixed group and class B genes are decomposed into three classes (Bs, B-AP3, and B-PI) in figure 4.2. This indicates that petunia also has all classes of genes (figure 4.2). The number of sequences for classes A, B-AP3, B-AP1, C, D, E, F, G, and T were 3, 2, 2, 2, 2, 5, 4, 1, and 2, respectively. We applied our statistical method for all gene pairs within each gene class, testing 23 pairs of genes (see the supplementary material for the 23 data sets). In this analysis we considered consecutive windows with $s = 1$ and $w = 30$. We

used the *p*-distance for this analysis. According to this analysis, 14 out of the 23 pairs of genes studied contained protein regions that showed at least one window with a *Z* value exceeding 1.96 (5% level).

The results of our test for a pair of class T genes (FBP25 and FBP13) are given in figure 4.1C. In this case the rice gene OsMADS47 and OsMADS54 were used as outgroups. The *Z* value line in this figure shows three peaks in which *Z* exceeds the 5% and 10% ($Z = 1.65$) significance levels (one each in the I, K, and C domains). As mentioned earlier, the I and K domains are important for homodimerization and heterodimerization of MADS-box proteins, whereas the C domain is involved in transcriptional activation in some proteins. It is possible that all the three domains are involved in the functional differentiation between FBP25 and FBP13. It is also interesting to note that protein FBP13 is known to have nine protein interaction partners, whereas protein FBP25 has no known interacting partners (Immink et al. 2003).

Figure 4.3 shows five more examples of our test in which *Z* became significant at the 5% level. The results of this analysis for a pair of class A genes (FBP29 and PFG) are presented in figure 4.3A. In this case the K domain has two peaks in which *Z* exceeds the 10% level. These peaks are located in a 30 amino acids region of the K domain. Therefore, experimentalists may focus on this region if they are interested in finding functional differentiation. The C domain also has two peaks of *Z* values which are significant at the 10% level. Therefore the C domain may also be tested for the possible functional differentiation. In the other four examples given in figure 4.3, only the K domain appears to have diverged significantly.

As mentioned above, there were eight more cases in which our test gave positive results (see table 4.1). In most of these cases the K domain again showed a *Z* value significant at the 5% or 10% level, though the M or I domain occasionally showed a significant region.

DISCUSSION

In our statistical analysis we implicitly assumed that different amino acid sites have evolved independently. If there are highly conserved regions or hyper-variable regions in the proteins studied, our test would not give accurate significance levels, and the test will be too liberal or too conservative depending on the data set. For example, if conserved protein regions are studied, the test results may be too liberal because some amino acid sites may not have changed at all and therefore the actual number of degrees of freedom may be smaller than w - 1. By contrast, if a differentiated protein region is longer than the window size, the test will be conservative because a test based on the entire protein region would give a higher Z-value than the regular window test due to the smaller sampling variance of D. However, since our test is intended to identify approximate protein regions to be tested biochemically, it does not need to be very accurate in terms of the statistical significance.

It should be noted that the positive results of our test do not necessarily mean that the identified regions are functionally differentiated. Therefore, if the biochemical test to be used is available, it is always recommended that both statistical and biochemical tests should be conducted. It should also be noted that our functional differentiation test is not necessarily related to the positive Darwinian selection examined by the ratio of the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) to the number of synonymous nucleotide substitutions per synonymous site (d_S). We are only interested in the functional differentiation of duplicate genes whether the d_N/d_S is higher than 1 or not. Actually the functional change of a gene may have been caused by a few amino acid changes in the functionally important region or by many substitutions in other

regions. Here, d_S is irrelevant under the assumption that synonymous nucleotide substitutions are neutral. Strictly speaking, this assumption is incorrect (e.g. Ikemura 1985; Akashi 1995), but for our purpose the violation of this assumption is not important.

When we applied our method to floral MADS-box genes, we found that the extent of the difference of evolutionary rate is not necessarily correlated with the number of interacting protein partners. This is different from some of the previous observations that the extent of evolutionary rate differences is sometimes negatively correlated with the number of protein partners (Hahn, Conant, and Wagner 2004). This difference could be due to the fact that we studied a specific protein group or may mean that our test does not necessarily detect the region where functional differentiation has occurred. These problems should be studied experimentally in the future.

It is interesting to note that functional specificity of class A, B and C genes in *Arabidopsis* is not determined by their DNA binding domain (Riechmann and Meyerowitz 1997). Therefore, I, K, and C domains may be critical for determining functional specificity of floral MADS-box genes. In our study the difference of evolutionary rate was often observed in the K domain, while it was not observed as often in the I and C domains. It has been proposed that internal repeats of proteins give favorable conditions for evolutionary change, because their functional constraint may change with time (Andrade, Perez-Iratxeta, and Ponting 2001). Therefore, the frequent observation of significant rate differences in the K domain may be related to the presence of heptad repeats, which can be subdivided into the K1, K2, and K3 regions (Yang, Fanning, and Jack 2003). Because the M, I, and C domains also showed significant rate differences in some pairs, it is possible that these domains have also been subject to the

functional differentiation. If functional differentiation occurs in different domains of a protein, the effect of such combinatorial differentiation on the regulatory network may be more significant than the case where only one domain is functionally differentiated. Our method may be useful for studying this problem as well.

Table 4.1. Summary of the analysis of 23 pairs of sequences.

Class	# of genes	Gene A	Gene B	Significant region(|Z| > 1.96); 1 = detected, 0 = not detected			
				M	I	K	C
A	3	FBP29	FBP26	0	0	0	1
		FBP29	PFG	0	0	1	1
		FBP26	PFG	0	0	0	0
B-AP3	2	Not analyzed					
B-PI	2	FBP1	PMADS2	0	1	0	0
C	2	PMADS3	FBP6	0	0	0	0
D	2	FBP11	FBP7	0	0	0	0
E	5	FBP9	FBP23	0	0	0	0
		FBP9	FBP4	0	0	0	0
		FBP9	FBP2	0	0	0	0
		FBP9	FBP5	0	0	1	0
		FBP23	FBP4	0	0	1	0
		FBP23	FBP2	0	0	1	0
		FBP23	FBP5	0	0	1	0
		FBP4	FBP2	0	0	0	0
		FBP4	FBP5	0	0	1	0
		FBP2	FBP5	0	0	1	0
F	4	FBP20-UNS	FBP21	0	0	0	0
		FBP20-UNS	FBP28	0	0	0	0
		FBP21	FBP28	0	0	1	0
		FBP20-UNS	FBP22	1	0	0	0
		FBP21	FBP22	0	0	1	0
		FBP28	FBP22	0	0	1	0
G	1	Not analyzed					
T	2	FBP13	FBP25	0	1	1	1
Total	**23 genes**	**23 pairs**		**1**	**2**	**11**	**3**
				14 pairs*			

*Some pairs showed significant rate differences in more than one domain.

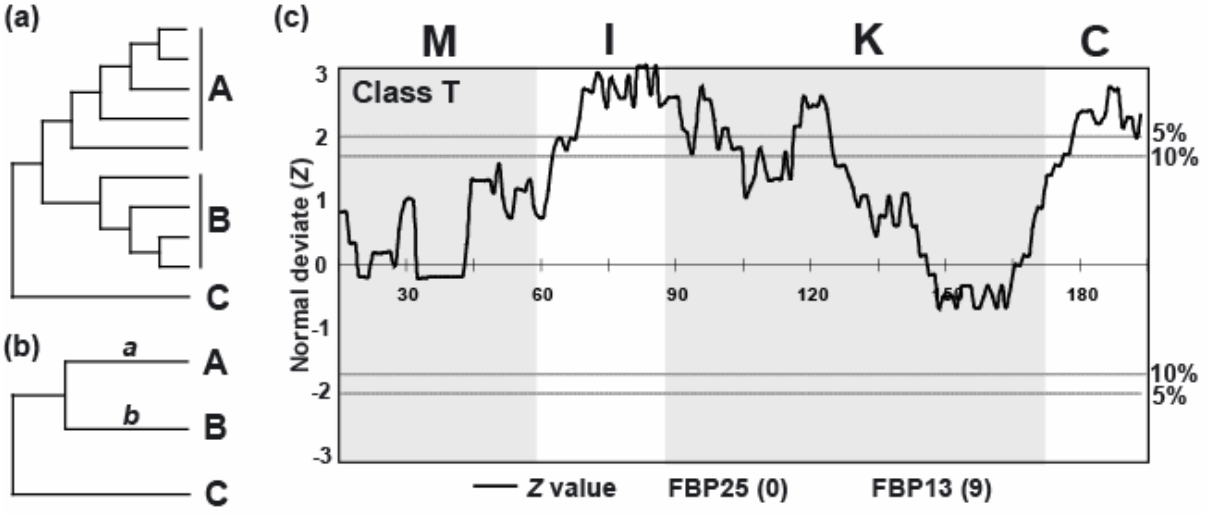


Figure 4.1. Comparison of Gu and Knudsen and Miyamoto's methods and our methods. (a) Gu (Gu 1999) and Knudsen and Miyamoto's tests (Knudsen and Miyamoto 2001). Groups *A* and *B* include many sequences to be compared, and *C* is the outgroup. (b) *A* and *B* represent two sequences to be compared, and *C* is the outgroup. (c) Comparison of the protein sequences of 2 class T MADS-box genes from petunia. For the outgroup two rice sequences were used. If FBP25 (the former) evolved faster than FBP13 (the latter) in a window, the *Z* value is positive, and if the former evolved slower than the latter, the *Z* value becomes negative. The number in parenthesis for each gene is the number of interacting protein partners given by Immink et al. (Immink et al. 2003). Horizontal lines with "5%" or "10%" correspond to the cutoff *Z* value of 1.96 or 1.65, respectively. The amino acid positions are given on the $Z = 0$ line. M, I, K, and C represent the M, I, K, and C domains. Window size (w) and skipping size (s) are 30 aa and one aa, respectively.

Figure 4.2. Phylogenetic tree of 68 MIKC-type MADS-box genes from petunia, *Arabidopsis*, and rice. This tree was constructed by the neighbor-joining method with *p*-distance. One hundred fifty one amino acids were used after removing all sites with alignment gaps. The number for each interior branch is the percent bootstrap value (500 bootstraps). The bootstrap values lower than 50% are not shown. The genes in bold characters with gray shadows are from petunia, and "At" and "Os" indicate *Arabidopsis* and rice genes, respectively. The numbers in parentheses refer to the numbers of interacting protein partners in the yeast Cytotrap system described in (Immink et al. 2003).

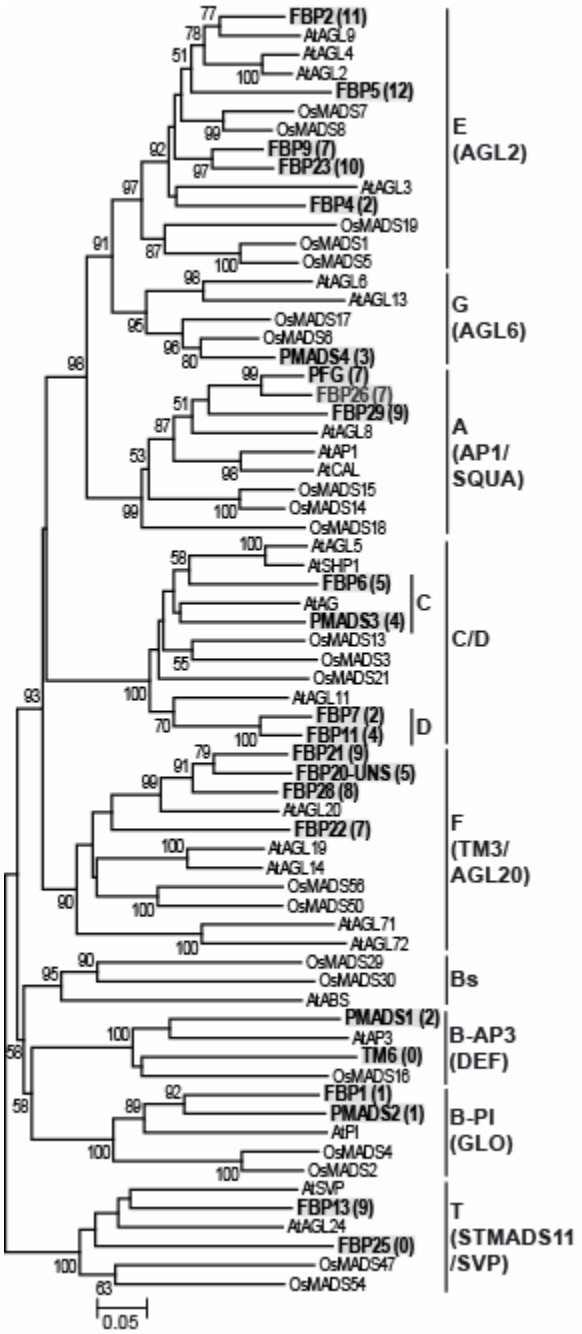
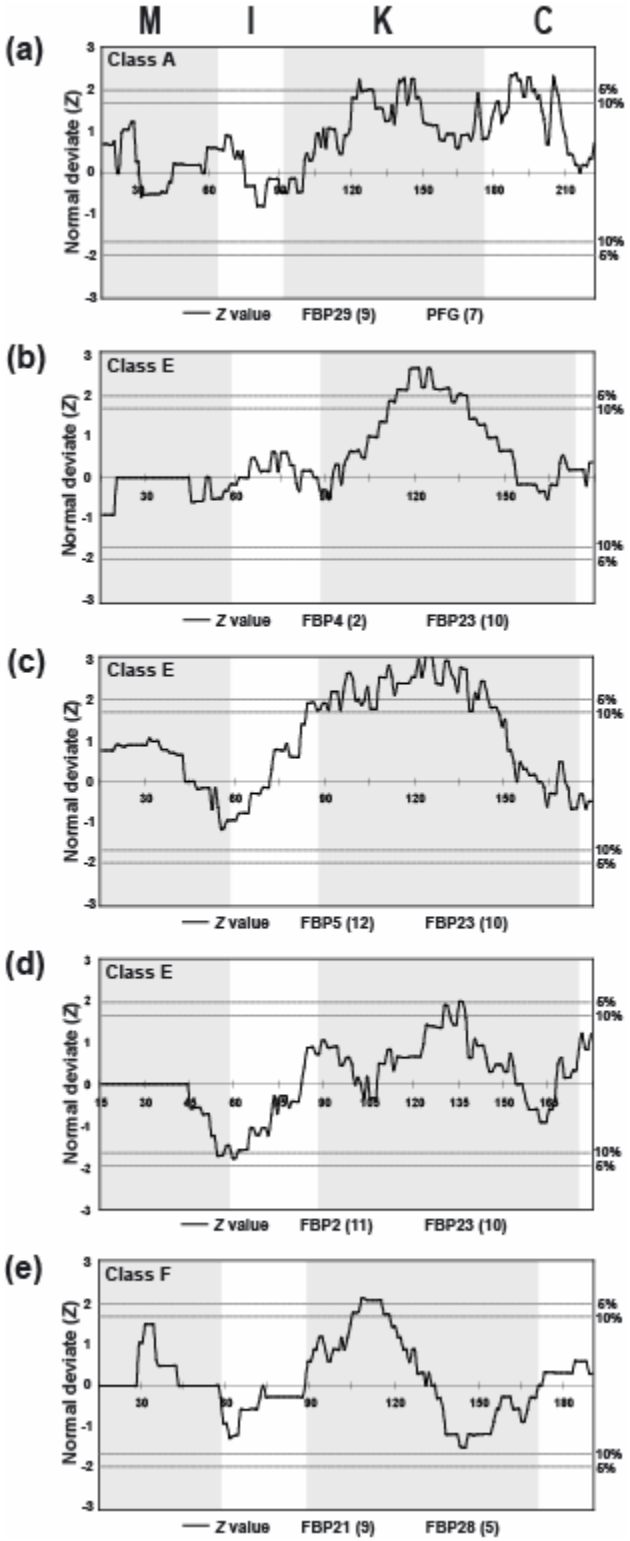

FBP2 (11)
AtAGL9
AtAGL4
AtAGL2
FBP5 (12)
OsMADS7
OsMADS8
FBP9 (7)
FBP23 (10)
AtAGL3
FBP4 (2)
OsMADS19
OsMADS1
OsMADS5
AtAGL6
AtAGL13
OsMADS17
OsMADS6
PMADS4 (3)
PFG (7)
FBP26 (7)
FBP29 (9)
AtAGL8
AtAP1
AtCAL
OsMADS15
OsMADS14
OsMADS18
AtAGL5
AtSHP1
FBP6 (5)
AtAG
PMADS3 (4)
OsMADS13
OsMADS3
OsMADS21
AtAGL11
FBP7 (2)
FBP11 (4)
FBP21 (9)
FBP20-UNS (5)
FBP28 (8)
AtAGL20
FBP22 (7)
AtAGL19
AtAGL14
OsMADS56
OsMADS50
AtAGL71
AtAGL72
OsMADS29
OsMADS30
AtABS
PMADS1 (2)
AtAP3
TM6 (0)
OsMADS16
FBP1 (1)
PMADS2 (1)
AtPI
OsMADS4
OsMADS2
AtSVP
FBP13 (9)
AtAGL24
FBP25 (0)
OsMADS47
OsMADS54
0.05
E (AGL2)
G (AGL6)
A (AP1/ SQUA)
C
C/D
D
F (TM3/ AGL20)
Bs
B-AP3 (DEF)
B-PI (GLO)
T (STMADS11 /SVP)

Figure 4.3. Five more cases in which significant rate differences were observed at the 5% level. The outgroup sequences used for each analysis are as follows: (a) OsMADS14/15/18, (b, c, and d) OsMADS1/5/19, and (e) AGL20.


M I K C
(a)
Class A
Normal deviate (Z)
5%
10%
10%
5%
Z value FBP29 (9) PFG (7)
(b)
Class E
Normal deviate (Z)
5%
10%
10%
5%
Z value FBP4 (2) FBP23 (10)
(c)
Class E
Normal deviate (Z)
5%
10%
10%
5%
Z value FBP5 (12) FBP23 (10)
(d)
Class E
Normal deviate (Z)
5%
10%
10%
5%
Z value FBP2 (11) FBP23 (10)
(e)
Class F
Normal deviate (Z)
5%
10%
10%
5%
Z value FBP21 (9) FBP28 (5)

CHAPTER 5

EVOLUTIONARY CHANGE OF THE NUMBERS OF HOMEOBOX GENES IN BILATERAL ANIMALS

SUMMARY

It has been proposed that the conservation or diversity of homeobox genes is responsible for the similarity and variability of some of morphological or physiological characters among different organisms. To gain some insights into the evolutionary pattern of homeobox genes in bilateral animals, we studied the change of the numbers of these genes during the evolution of bilateral animals. We analyzed 2031 homeodomain sequences compiled from 11 species of bilateral animals ranging from *Caenorhabditis elegans* to human. Our phylogenetic analysis using a modified reconciled-tree method suggested that there were at least 88 homeobox genes in the common ancestor of bilateral animals. About 50 – 60 genes of them have left at least one surviving descendant in each of the 11 species, suggesting that about 30 – 40 genes were lost in a lineage-specific manner. Although similar numbers of ancestral genes have survived in each genome, vertebrate lineages gained many more genes by duplication than invertebrate lineages, resulting in more than 200 homeobox genes and about 100 in vertebrates and invertebrates, respectively. After these gene duplications, a substantial number of old duplicate genes have also been lost in each lineage. Because many old duplicate genes were lost, it is likely that lost genes had already been differentiated from other groups of genes at the time of gene loss. We conclude that both gain and loss of homeobox genes were important for the evolutionary change of phenotypic characters in bilateral animals.

INTRODUCTION

Homeobox genes that regulate morphogenesis were first discovered by Garber, Kuroiwa, and Gehring (1983) and Scott et al. (1983) in *Drosophila melanogaster* (fruitfly). Subsequent studies of homeobox genes in fruitflies, frogs, and humans revealed a highly conserved motif of about 180 bp called the homeobox (McGinnis et al. 1984; Scott and Weiner 1984). The homeobox encodes a DNA-binding domain called the homeodomain. In the genomes of animals and plants homeobox genes form a large transcription factor gene family, with more than 200 genes in humans and about 80 genes in *Arabidopsis*.

Animal homeobox genes were previously classified into about 30 different groups or families based on their sequence similarity and protein domain structure (Burglin 1994). Additional groups of homeobox genes were identified later (e.g., PBX, MEIS, KNOX, TGIF, and IRX; Burglin 1997), and now the homeobox genes in animals can be classified into at least 49 different gene families. Member genes of the same family are often functionally related, and different families of homeobox genes are concerned with different aspects of development (Burglin 1994). For example, the genes of the HOX, CDX, and EVX families and their cognate genes play important roles in different steps of pattern formation during early embryogenesis of animals. The PAX6, SIX, VAX, and EMX gene families are concerned with the development of eyes, whereas the LIM and HMX gene families are important in the development of neurons (reviewed in Duboule 1994). Because of their important roles in development, homeobox genes have been studied intensively by both developmental and evolutionary biologists.

Homeobox genes are generally highly conserved and control similar phenotypic characters among distantly related organisms (reviewed in De Robertis 1994). However, they are also responsible for controlling different phenotypic characters among relatively closerly related species (e.g., Galant and Carroll 2002; Ronshaugen, McGinnis, and McGinnis 2002). The formation of similar phenotypic characters can be explained by the conservation of shared homeobox genes. By contrast, different phenotypic characters are believed to be generated by duplication of homeobox genes and their functional differentiation. It has also been hypothesized that the loss of some homeobox genes are responsible for the morphological differentiation (Ruddle et al. 1994). Therefore, it is interesting to study the pattern of duplication and loss of homeobox genes to have some insights into the evolutionary change of phenotypic characters. It is likely that the number of homeobox genes is related to the complexity of organisms.

Although the patterns of gain and loss of some families of homeobox genes have been studied (e.g., Zhang and Nei 1996; Aparicio et al. 1997; Wada et al. 2003; Amores et al. 2004; Edvardsen et al. 2005), no one appears to have studied the gain and loss of the entire set of homeobox genes covering diverse bilateral animals. We have therefore decided to study the evolutionary change of the homeobox gene superfamily examining 11 completely or nearly completely sequenced genomes from bilateral animals. The results obtained will be presented in this paper.

MATERIALS AND METHODS

Identification of homeodomain-containing proteins

To find homeodomain-containing proteins, we performed homology search using the tool PSI-BLAST (Altschul et al. 1997) for the entire set annotated proteins of *Caenorhabditis elegans*, *C. briggsae,* mosquito (*Anopheles gambiae*), fruitfly (*Drosophila melanogaster*), tunicate (*Ciona intestinalis*), zebrafish (*Danio rerio*), pufferfish (*Fugu rubripes*), frog (*Xenopus tropicalis*), rat, mouse, and human. All sequence data except for the tunicate were downloaded from the ENSEMBL (ftp://ftp.ensembl.org) as of 21 February 2005. The tunicate data set (version 1) was downloaded from the Joint Genome Institute (http://genome.jgi-psf.org/). We used 194 homeodomain sequences from animals, plants, and fungi as queries, with an E-value $\leq 10^{-5}$ (see supplementary material). We also searched for homeobox genes from the EST database of the tunicate from the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/), because Wada et al. (2003) reported several unannotated homeobox genes from the EST database of this organism.

Phylogenetic analysis

Because the homeodomain is the only alignable region between different groups of homeodomain-containing proteins, we used only this domain ($\approx$ 60 aa) for phylogenetic analysis. The homeodomain sequences were aligned against the alignment of 194 query sequences (seed alignments) using the profile alignment of the ClustalX program (Thompson et al. 1997). We then constructed a neighbor-joining tree (Saitou and

Nei 1987) using the computer program NJBOOT (Takezaki, Rzhetsky, and Nei 1995) with the pairwise deletion option, proportional amino acid difference (*p*-distance), and 1000 bootstrap resamplings (Nei and Kumar 2000). Because of the large number of sequences used, other tree construction methods such as maximum-parsimony and maximum-likelihood methods were not used.

Each homeobox gene was assigned to one of the 49 previously defined groups according to the sequence similarity and protein domain structure. Domain structure was examined by using the computer program HMMER (Eddy 2001) for each protein domain profile downloaded from the Pfam (http://pfam.wustl.edu/). A phylogenetic tree for the 49 families of genes was constructed to find their evolutionary relationships.

Estimation of the number of genes in the ancestral species

When there is a rooted tree of m species, the tree has m - 1 ancestral nodes or species (Nei and Kumar 2000). We are interested in estimating the number of homeobox genes in each of the ancestral species and how the number has changed in the evolutionary process. This can be studied by comparing the species tree with the gene tree for a given set of genes and constructing a reconciled tree (Goodman et al. 1979; Page and Charleston 1997). In this paper we use a slightly modified version of this reconciled-tree method, in which multifurcating branching patterns are taken into account.

For simplicity, let us consider the tree of three species **α**, **β**, and **γ** in figure 5.1A and assume that species **α**, **β**, and **γ** have 3, 2, and 2 genes, respectively. Suppose that the gene tree inferred for the 7 genes from the three species is given by figure 5.1B and that this tree represents the true gene tree. This gene tree can be decomposed into three groups

of genes (I, II, and III), in which genes a, b, and c come from species **α**, **β**, and **γ**, respectively. Group I genes do not include any c gene but contain two pairs of genes a and b. Therefore, to reconcile this portion of the gene tree with the species tree under the principle of parsimony, we must assume that one deletion of gene c and one duplication of the ancestral gene of genes a and b occurred (figure 5.1C). Similarily, to reconcile the group II genes with the species tree, we will have to assume that gene b was deleted. In the case of group III genes we have to consider the deletion of genes a and b. In this case we assume that one event of deletion occurred in the ancestral lineage of genes a and b. The reconciled tree in figure 5.1C therefore suggests that the ancestral species **δ** had three genes (ancestral genes of the three groups of genes).

A similar inference indicates that the ancestral species **ε** also contained three genes, i. e., 2 genes in group I, 1 gene in group II, and 0 gene in group III (figure 5.1C). Figure 5.1A now shows the change of the number of genes in the evolutionary lineages for **α**, **β**, and **γ**.

In the above estimation of the number of genes we assumed that the gene tree is correct. In practice, however, some interior branches are often weakly supported in terms of bootstrap values. For example, one interior branch may have a low bootstrap value (< 50% in the present study). In this case, the existence of this branch is questionable, so that the length of this branch is reduced to 0, and a condensed tree (Nei and Kumar 2000) is constructed (figure 5.1D). The real gene tree in this case will be one of the three possible trees given in figures 5.1E, F, and G. Tree E is identical with tree B, and we already estimated the number of genes in the ancestral species **δ** and **ε**. In the case of tree F, the reconciled tree is given by figure 5.1H, and the numbers of genes in species **δ** and **ε**

are given in figure 5.1I. This tree is more parsimonious than tree A with respect to the change of gene number. Tree G is different from tree F in that group II genes are closer to gene c, but the number of genes estimated in the ancestral organisms becomes the same as those for tree B. Therefore, we assume that tree I gives the actual numbers of ancestral species.

When there are several branches with low bootstrap values, the numbers of genes in ancestral species are estimated by the same procedure as the above under the principle of parsimony. Therefore, one can estimate the number of genes for any number of ancestral species. Obviously, the number of genes estimated would be minimal, but since homeobox genes evolve very slowly, the present method appears to give reasonably good estimates (see below). When m is large, the computation can be quite complicated, and we have developed a computer program (see http://www.bio.psu.edu/People/Faculty/Nei/Lab/software.htm).

RESULTS

Number of homeobox genes in the genome

Table 5.1 shows the numbers of non-redundant homeobox genes obtained from the annotated gene sets of 11 species. The majority of the homeobox genes encode only one homeodomain (single-homeobox genes), but some encode more than one domains (multi-homeobox genes). The number of homeodomains encoded by a multi-homeobox gene was less than ten with some exceptions. All vertebrate species studied (pufferfish, zebrafish, frog, mouse, rat, and human) had about 200 or more homeobox genes, and all invertebrate species (*C. elegans*, *C. briggsae*, fruitfly, mosquito, and tunicate) had about 100 or fewer homeobox genes. All of the sequences used are presented as supporting information on the webpage (see file 1).

Evolutionary relationships of different groups of homeobox genes

The majority of the homeobox genes were assigned into the 49 previously defined groups. The remaining homeobox genes were either highly divergent or multi-homeobox genes, so that they were not used in this study. The list of the genes in each of the 49 groups is available from the supplementary material. Figure 5.2 shows the evolutionary relationships of the 49 groups of genes. The homeobox gene superfamily was initially classified into two groups, the typical and atypical homeobox gene groups (Burglin 1994). The homeobox of typical genes encodes a 60 amino acid-long homeodomain composed of three helical regions, and the homeobox of atypical genes encode additional amino acids either between helices 1 and 2 or between helices 2 and 3 (Burglin 1994). The

typical and atypical groups of genes are presented in figure 5.2. One group of atypical homeobox genes encode three additional amino acids between helices 1 and 2 and are called TALE (Three Amino acids Loop Elongation) class genes. Bharanthan et al. (1997) and Burglin (1997) proposed that the typical and TALE homeobox genes diverged before the animal-plant split. Our tree shows the separation of most typical homeobox genes (except SIX group genes) and TALE homeobox genes and supports their notion. However, because we do not know the exact position of the root in the tree, it is difficult to know whether the early evolved groups of typical homeobox genes (e.g., SIX) are evolutionarily closer to TALE genes than to other typical genes or not.

At least 13 groups of homeobox genes (gene groups with orange boxes in figure 5.2) encode other evolutionarily conserved domains in addition to the homeodomain. Interestingly, a majority of them appear to be early evolved groups of homeobox genes. Therefore, our results suggest that there were already many homeobox genes in the most recent common ancestor (MRCA) of the 11 species, and their domain structures were already quite complex. Since we did not include the multi-homeodomain proteins and the proteins that are not assigned to any of the 49 groups, the evolutionary history of the entire set of homeobox genes should be more complex than that shown in figure 5.2.

Figure 5.2 also shows the numbers of homeobox genes for each of the 49 families as well as for those of multi-homeobox genes and unassigned genes. In most groups vertebrates have about two to four times more genes than invertebrates. However, there are several gene groups that do not show this pattern. For example, the NOT gene, which is important for the development of notochords in zebrafish (Talbot et al. 1995), has been found as a single-copy gene in three invertebrate species, fruitflies, mosquitoes, and

tunicates, and three vertebrate species, zebrafish, pufferfish, and frog. Recently, it has been claimed that a mouse ortholog of the NOT gene was found (Abdelkhalek et al. 2004; Plouhinec et al. 2004), but a phylogenetic analysis presented actually suggests that it is at best a paralog of the NOT genes. There are also other families of genes where vertebrates have fewer or no more genes than those of invertebrates. Because the genome sequencing or annotation of several species has not been completed, the absence of some families of genes should be reexamined, though it may not change the general pattern. It is also possible that the genes have not been lost completely but only their homeoboxes are missing and other regions of the gene are still functional as in the case of some PAX genes (Chi and Epstein 2002).

Evolutionary change of the number of homeobox genes in bilateral animals

Knowing that there were already many homeobox genes in the MRCA of all the 11 species (archi-MRCA), we estimated the numbers of homeobox in the ancestral organisms and their increase and decrease in different stages of the evolution of bilateral animals. We conducted a phylogenetic tree of 2031 homeodomain sequences compiled from single- and multi-homeodomain-containing proteins in relation to the species tree (figure 5.3). The species tree is based on the observation that insects, tunicates, and vertebrates are coelomates, but nematodes are pseudocoelomates (Coelomata hypothesis) as well as phylogenetic analyses using more than 100 nuclear proteins from several species (e.g., Blair et al. 2002; Wolf, Rogozin, and Koonin 2004). (The Ecdysozoa hypothesis in which insects and nematodes are sister groups will be considered later.) Because the protein sequences used are generally closely related, we used a bootstrap

cutoff of 50% for generating a multifurcating node in the gene tree. As mentioned above, the number of homeoboxes in multi-homeobox genes varies with gene, and therefore it is difficult to estimate the real number of ancestral genes for these genes. We therefore decided to regard each homeobox as one gene. However, this will not affect our results significantly, because the number of multi-homeobox genes included was small.

To check the reliability of our estimates, we first analyzed HOX group genes. The numbers of ancestral HOX group genes at several evolutionary time points have already been estimated by several researchers (e.g., Holland and Garcia-Fernandez 1996; Zhang and Nei 1996; Stellwag 1999; Wada et al. 2003). In the case of HOX genes, estimation of the numbers of ancestral genes is relatively easy, because information about the conserved genomic locations of HOX genes can also be used for reconstructing the ancestral states. We compared our estimates of the numbers of ancestral genes with the previous estimates, assuming that the previous estimates are correct (figure 5.3A). The previous estimates for ancestral species **α**, **β**, **γ**, **δ**, and **ζ** in figure 5.3A were 5, 6, 9, 43, and 39, respectively, whereas our estimates for the same ancestral species were 4, 6, 9, 24, and 26 in this order. This result suggests that our estimates of the numbers of genes in the ancestral species tend to be smaller than the previous ones. This has happened partly because some of paralogous homeodomains from different species have identical or nearly identical sequences, and therefore the interior branches involved became 0 or had low bootstrap values. In this case, the numbers of genes lost are expected to be underestimated. However, we note that the number of genes gained at the exterior branch of each lineage is very likely to have been overestimated in our method, because the numbers of ancestral genes were likely underestimated. We also note that the incomplete

annotation of genes will increase the numbers of genes lost in the external branches. However, this effect should be quite minor for the interior branches, because the possibility of missing the same genes in two different genomes should be very low.

Keeping in mind this possibility of underestimation, we estimated the numbers of ancestral genes and the numbers of genes lost and gained during the evolution of the entire homeobox gene superfamily. Figure 5.3B shows that there were at least 88 ancestral homeobox genes in the archi-MRCA. This archi-MRCA already had several genes in some homeobox families (e.g., 6 genes in PAX group) (figure 5.2). However, some families of genes (LBX, VAX/NOT, MEOX, and ZF) were not found in nematodes (figure 5.2), and separation of these families of genes from other families was not always clear-cut in the tree of 2031 sequences (data not shown). Therefore, these families of genes were not considered in the estimation of genes in the archi-MRCA. However, figure 5.2 shows that these groups of genes already diverged from other groups of genes before the nematode and mammal split. Therefore, it is possible that the genes from these four families actually existed in the archi-MRCA. If this is the case, the number of homeobox genes in the archi-MRCA would increase to about 92.

After the divergence of Coelomates and Pseudocoelomates the number of homeobox genes increased almost three-fold in the vertebrate lineages. In invertebrates, however, the increase was small or moderate, and our results suggest that the number of homeobox genes did not merely increase during the evolutionary process, but the number sometimes decreased. For example, the MRCA of insects and vertebrates had at least 118 homeobox genes, but fruitflies have 102 at present. Tunicates also have fewer homeobox genes than the MRCA of tunicates and vertebrates. In the case of vertebrate lineages the

number of genes increased primarily in two time periods, that is, the early stages of Coelomate evolution (between nodes **α** and **β** in figure 5.3B) and the early stages of vertebrate evolution (between nodes **γ** and **δ** in figure 5.3B). The major increase of gene number in these two time periods are consistent with that of the total number of genes in the genome by Gu, Wang, and Gu (2002). Note that if a gene is duplicated and one of the two duplicate genes was lost during the same time interval, our approach cannot detect them, as mentioned earlier. Therefore, the gene losses estimated here are losses of the genes of fairly old duplication events.

The Ecdysozoa hypothesis (e.g., Aguinaldo et al. 1997; Dopazo and Dopazo 2005) proposes that two molting animals, nematodes and insects, are more closely related to each other than to vertebrates. We used the species tree based on this hypothesis to estimate the numbers of ancestral homeobox genes. The estimated numbers of ancestral genes and those of gene losses are much higher than those based on the Coelomata hypothesis (see supplementary material). Therefore, in terms of the numbers of ancestral genes and gene losses, the species tree based on the Coelomata hypothesis gives us more conservative estimates than those based on the Ecdysozoa hypothesis.

Retention and loss of ancestral homeobox genes in each species

It is interesting to know how many gene families of the archi-MRCA have left descendent genes in the 11 species and how many gene families have been lost during this evolutionary period. Figure 5.2 shows that in nematodes 12 of the 49 gene families in the archi-MRCA have been lost, whereas 32 of them have been retained, the remaining five gene families existing in neither archi-MRCA nor in nematodes. In vertebrates,

however, all the archi-MRCA genes were found. In addition, some new gene families originated in insects, tunicates, and vertebrates. The HOX gene family has the largest number of member genes in vertebrates, and the archi-MRCA also had a relatively high number of member genes (4 genes). The highest number of archi-MRCA genes was observed in the LIM family, but the number of genes did not increase very much compared with the HOX gene family. Some gene families such as the BSH and VSX families had small numbers of genes in all species.

We also studied the numbers of ancestral homeobox genes lost during the time period from the archi-MRCA to the present species (column 2 in Table 5.2). Let us use figure 5.1C to illustrate how we counted the numbers of genes lost. This figure shows three groups of genes (I, II, and III) that were derived from the three genes in the ancestral species **δ**. In group I genes, gene a and b are retained, but gene c was lost. In group II genes, gene a and c are retained, but gene b was lost. Similarly, in group III genes, gene c is retained, but genes a and b were lost. Therefore, in species **α** two ancestral genes are retained and one gene was lost. Similarly, the number of genes lost is 2 in species **β** and is one in species **γ**. The total number of genes lost in each extant species is given by the sum of these losses for all gene families. Note that this number of gene losses can be computed for each MRCA.

Table 5.2 shows that the invertebrate lineages lost about 30 ~ 38 genes from the 88 ancestral genes in the archi-MRCA, while the vertebrate lineages lost about 25 ~ 28 genes during this evolutionary period. This suggests that invertebrates lost somewhat more genes from the archi-MRCA than vertebrates. However, the difference is much smaller than that observed with full genome analysis, which suggests that about two-fold

or more gene losses occurred in the lineage leading to *C. elegans* than the lineage leading to human (Hughes and Friedman 2004; Koonin et al. 2004; Ogura, Ikeo, and Gojobori 2005). Similarly, the numbers of genes lost from the ancestor **β** to insects and tunicates were somewhat higher than those to the vertebrates (see column 3 in Table 5.2). Within vertebrates, the numbers of lost genes are more or less the same from most ancestral organisms to each of the descendent species. However, since the major increase of gene number occurred in the early stage of vertebrate evolution (between **γ** and **δ**), fishes lost somewhat smaller number of genes than other vertebrates.

These results suggest that the degree of gene loss varies significantly among different families of homeobox genes, but it was not so different among different species.

DISCUSSION

In this study we showed that there were at least 88 homeobox genes in the archi-MRCA of bilateral animals. Previously we mentioned that our statistical method would give minimum estimates of the numbers of ancestral genes. However, our estimate of the total number of genes in the archi-MRCA is close to the current number of genes in nematodes and insects. This is also true with the number in each gene family. These observations suggest that our estimates may not be too far off from the true numbers. Furthermore, the similarity of the estimates for the archi-MRCA and those for nematodes or insects suggest that the archi-MRCA had the same degree of phenotypic complexity as those of current nematodes or insects. Since vertebrates gained more homeobox genes than invertebrates, it appears that this increase in the number of homeobox genes is responsible for the formation of more complex characters in vertebrates than in invertebrates.

We have also seen that many homeobox genes have been lost in the process of evolution of phenotypic characters. This loss of homeobox genes might have been either inactivation of redundant genes after gene duplication or loss of functionally differentiated genes (Ruddle et al. 1994; Wagner, Amemiya, and Ruddle 2003). The genes lost in our study are losses of fairly old duplicates, and therefore it is likely that the genes lost were already functionally differentiated from one another at the time of gene loss. This raises the question of why these genes could be lost so often. There are at least three possible reasons. First, without closely related paralogs, functional redundancy can be achieved by something called distributed robustness (reviewed in Wagner 2005). In other words, loss (or mutation) of a homeobox gene can be buffered by the rewiring of

functionally different parts of the regulatory network. If so, it is possible that losses of homeobox genes might not have caused any noticeable changes of phenotypes. Second, it is also possible that gene loss occasionally has beneficial effects. For example, loss of genes may be related to the reduction of unused characters, and in this case loss of genes controlling the character may have beneficial effects for the development of the entire set of phenotypic characters. Third, the phenotypic changes caused by the loss of homeobox genes have been more or less neutral with respect to fitness. In the case of multi-functional genes, this is possible, if the critical functions are shared by duplicates, but other functions are not.

It should be noted that the gain and loss of homeobox genes are opportunistic but these events probably changed the evolutionary courses of different organisms. However, the possible causes of gene loss mentioned above are all speculative, and more detailed studies are needed to identify the real reason.

Table 5.1. Estimates of the numbers of homeobox genes in 11 species and the MRCA of all 11 species

	Archi-MRCA	Nematodes		Insects		Tunicate	Teleosts		Frog	Rodents		Human
		C. ele.	*C. bri.*	Fruit-fly	Mosquito		Puffer-fish	Zebra-fish		Mouse	Rat	
Single-homeodomain proteins	82	83	68	97	81	97	229	279	191	234	205	217
Multi-homeodomain proteins	(6)	9 (30)	2 (4)	2 (5)	2 (4)	2 (5)	10 (36)	10 (37)	7 (27)	8 (27)	11 (35)	13 (40)
Total number of proteins (homeodomains)	(88)	92 (113)	70 (72)	99 (102)	83 (85)	99 (102)	239 (265)	289 (316)	198 (218)	242 (261)	216 (240)	230 (257)

—The estimates of the numbers of homeobox genes are shown outside parenthesis, and those of the numbers of homeodomains are shown in parentheses. These two numbers are different because of multi-homeodomain proteins.

Table 5.2. Estimates of the numbers of genes lost from each MRCA to each species

Species	Ancestral nodes in figure 5.3					
	α	**β**	**γ**	**δ**	**ε**	**ζ**
	88	**118**	**116**	**248**	**209**	**214**
C. elegans	34					
C. briggsae	38					
Fruitfly	30	53				
Mosquito	35	59				
Tunicate	35	57	47			
Zebrafish	25	40	29	71		
Pufferfish	27	43	31	78		
Frog	26	43	34	90	42	
Mouse	26	43	34	91	41	26
Rat	28	46	38	97	47	36
Human	26	42	33	86	34	17

—Estimates of the numbers of genes lost in each species from the same MRCA are shown in the same column.

Figure 5.1. A simple example of the method for estimating the numbers of ancestral, gained, and lost genes. We assume that there are 3 species (species **α**, **β**, and **γ**), and species **α**, **β**, and **γ** have 3, 2, and 2 genes, respectively. **A.** The species tree and the numbers of ancestral, gained, and lost genes. The most recent common ancestor (MRCA) of species **α**, **β**, and **γ** and the MRCA of species **α** and **β** are labeled **δ** and **ε**, respectively. The numbers within square boxes are the numbers of genes in extant species (species **α**, **β**, and **γ**) or ancestral species (species **δ** and **ε**). The numbers of genes gained and genes lost in each ancestral branch are shown on the right and left sides of each branch, respectively. **B.** The gene tree of the 7 genes. **C.** The reconciled tree of figure 5.1A and figure 5.1B. Black and grey dots stand for speciation events (sp.), empty black circles for gene duplication events, and crosses for gene losses. **D.** The condensed tree of figure 5.1B. **E-G.** Three possible gene trees that can be inferred from the condensed tree in figure 5.1D. **H.** the simplest reconciled tree of figure 5.1D. **I.** The species tree and the numbers of ancestral, gained, and lost genes counted from figure 5.1H.

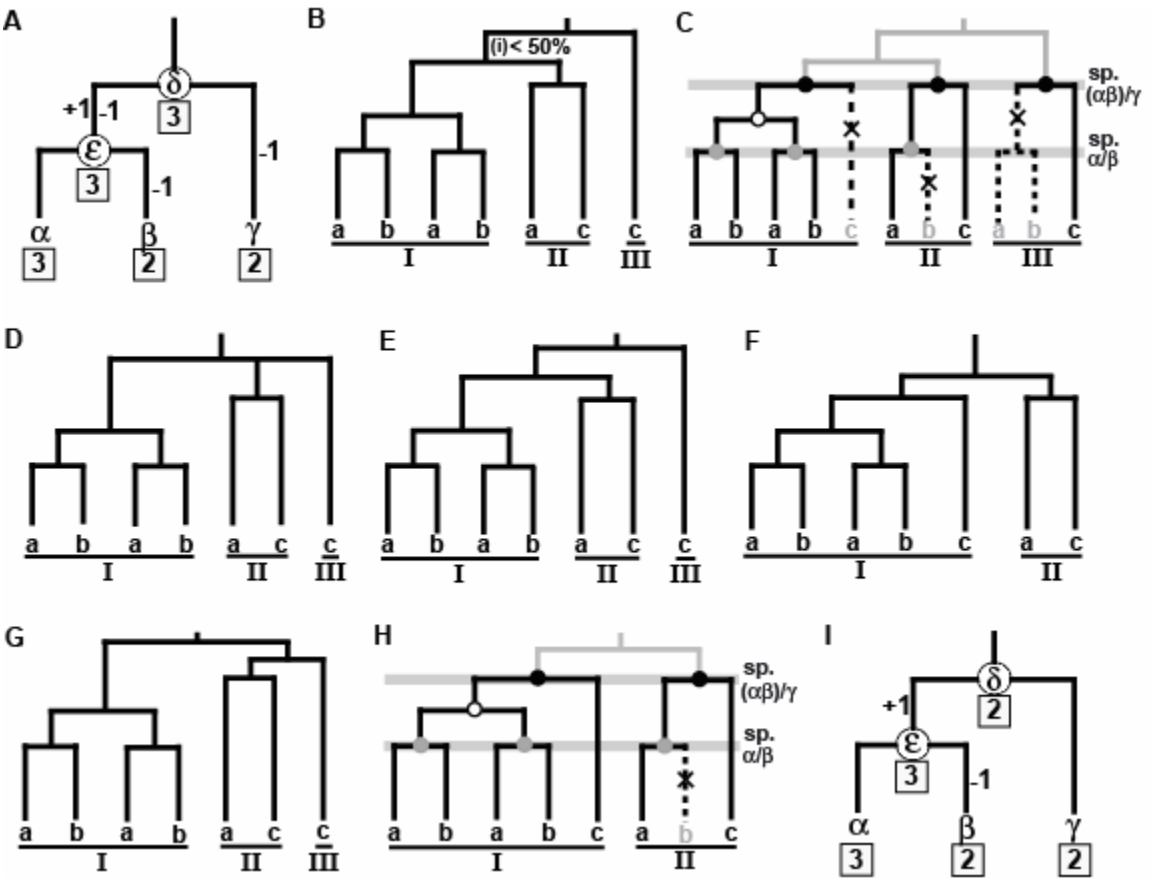
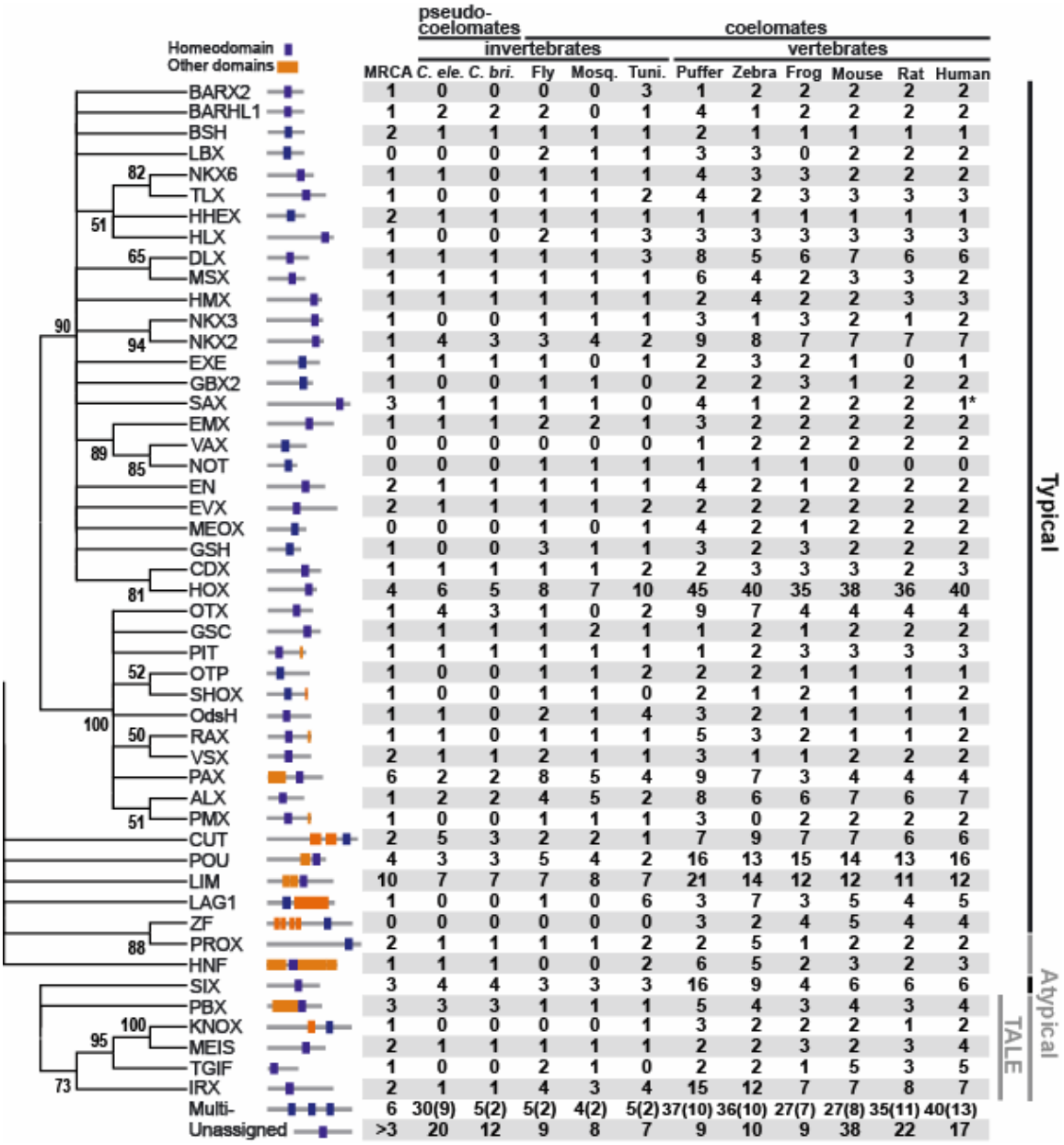

A
δ
3
+1
-1
ε
3
-1
-1
α
3
β
2
γ
2
B
(i)< 50%
a b a b a c c
I II III
C
sp.
(αβ)/γ
sp.
α/β
a b a b c a b c a b c
I II III
D
a b a b a c c
I II III
E
a b a b a c c
I II III
F
a b a b c a c
I II
G
a b a b a c c
I II III
H
sp.
(αβ)/γ
sp.
α/β
a b a b c a b c
I II
I
δ
2
+1
ε
3
-1
α
3
β
2
γ
2

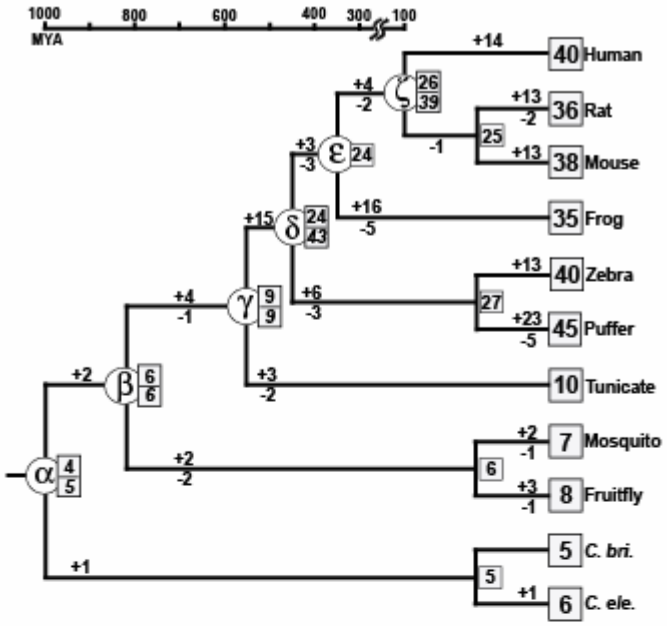
Figure 5.2. The evolutionary relationships of 49 different families of homeobox genes and their phylogenetic distribution in the 11 species of bilateral animals. The tree is constructed by the neighbor-joining method using average *p*-distances between 49 groups and is a 50% bootstrap consensus tree (100 bootstrap replications). Bootstrap values higher than 50% are shown. Representative domain organization is shown on the right-hand side of each family name. Black vertical lines indicate the typical homeobox gene family, and gray vertical lines the atypical homeobox gene family. Each blue square indicates a homeodomain, and orange squares indicate the conserved family specific domains. Gray horizontal lines indicate full-length proteins. Domains of E-value < 0.01 in the HMM search are shown. The numbers of homeobox genes for each family in each species is also shown. Numbers in parentheses are the numbers of homeoboxes from multi-homeobox genes. Numbers under "MRCA" are the estimated numbers of homeobox genes in the MRCA of the 11 species using species trees based on the Coelomata hypothesis. No SAX family gene was found in the annotated data set of human genes from the ENSEMBL. However, the annotation data set of human genes from the GenBank contains one copy of SAX group gene. We therefore included this gene in this tabulation (number with "*" mark). Note that the SIX family genes are typical homeobox genes and that the gene numbers for the HOX gene family are slightly higher than those of the genes in the HOX cluster. This is because other closely related genes (e.g., IPF) were also included in this family.



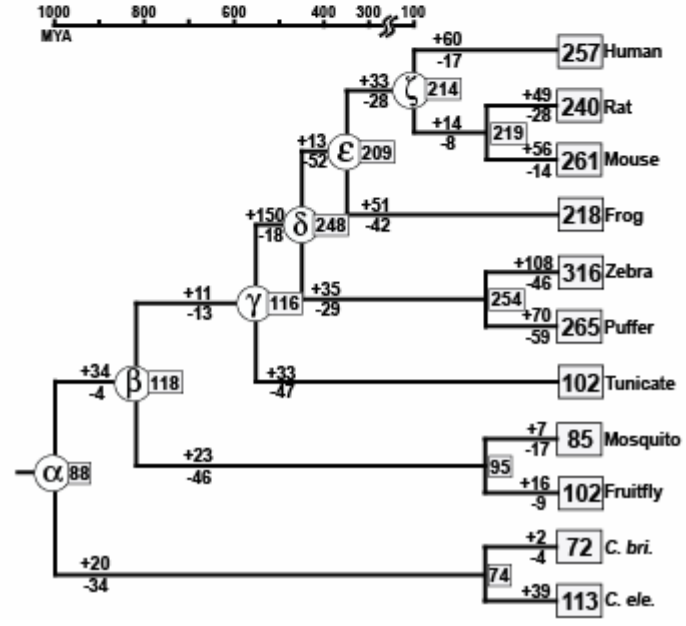
	MRCA	*C. ele.*	*C. bri.*	Fly	Mosq.	Tuni.	Puffer	Zebra	Frog	Mouse	Rat	Human
BARX2	1	0	0	0	0	3	1	2	2	2	2	2
BARHL1	1	2	2	2	0	1	4	1	2	2	2	2
BSH	2	1	1	1	1	1	2	1	1	1	1	1
LBX	0	0	0	2	1	1	3	3	0	2	2	2
NKX6	1	1	0	1	1	1	4	3	3	2	2	2
TLX	1	0	0	1	1	2	4	2	3	3	3	3
HHEX	2	1	1	1	1	1	1	1	1	1	1	1
HLX	1	0	0	2	1	3	3	3	3	3	3	3
DLX	1	1	1	1	1	3	8	5	6	7	6	6
MSX	1	1	1	1	1	1	6	4	2	3	3	2
HMX	1	1	1	1	1	1	2	4	2	2	3	3
NKX3	1	0	0	1	1	1	3	1	3	2	1	2
NKX2	1	4	3	3	4	2	9	8	7	7	7	7
EXE	1	1	1	1	0	1	2	3	2	1	0	1
GBX2	1	0	0	1	1	0	2	2	3	1	2	2
SAX	3	1	1	1	1	0	4	1	2	2	2	1*
EMX	1	1	1	2	2	1	3	2	2	2	2	2
VAX	0	0	0	0	0	0	1	2	2	2	2	2
NOT	0	0	0	1	1	1	1	1	1	0	0	0
EN	2	1	1	1	1	1	4	2	1	2	2	2
EVX	2	1	1	1	1	2	2	2	2	2	2	2
MEOX	0	0	0	1	0	1	4	2	1	2	2	2
GSH	1	0	0	3	1	1	3	2	3	2	2	2
CDX	1	1	1	1	1	2	2	3	3	3	2	3
HOX	4	6	5	8	7	10	45	40	35	38	36	40
OTX	1	4	3	1	0	2	9	7	4	4	4	4
GSC	1	1	1	1	2	1	1	2	1	2	2	2
PIT	1	1	1	1	1	1	1	2	3	3	3	3
OTP	1	0	0	1	1	2	2	2	1	1	1	1
SHOX	1	0	0	1	1	0	2	1	2	1	1	2
OdsH	1	1	0	2	1	4	3	2	1	1	1	1
RAX	1	1	0	1	1	1	5	3	2	1	1	2
VSX	2	1	1	2	1	1	3	1	1	2	2	2
PAX	6	2	2	8	5	4	9	7	3	4	4	4
ALX	1	2	2	4	5	2	8	6	6	7	6	7
PMX	1	0	0	1	1	1	3	0	2	2	2	2
CUT	2	5	3	2	2	1	7	9	7	7	6	6
POU	4	3	3	5	4	2	16	13	15	14	13	16
LIM	10	7	7	7	8	7	21	14	12	12	11	12
LAG1	1	0	0	1	0	6	3	7	3	5	4	5
ZF	0	0	0	0	0	0	3	2	4	5	4	4
PROX	2	1	1	1	1	2	2	5	1	2	2	2
HNF	1	1	1	0	0	2	6	5	2	3	2	3
SIX	3	4	4	3	3	3	16	9	4	6	6	6
PBX	3	3	3	1	1	1	5	4	3	4	3	4
KNOX	1	0	0	0	0	1	3	2	2	2	1	2
MEIS	2	1	1	1	1	1	2	2	3	2	3	4
TGIF	1	0	0	2	1	0	2	2	1	5	3	5
IRX	2	1	1	4	3	4	15	12	7	7	8	7
Multi-	6	30(9)	5(2)	5(2)	4(2)	5(2)	37(10)	36(10)	27(7)	27(8)	35(11)	40(13)
Unassigned	>3	20	12	9	8	7	9	10	9	38	22	17

Figure 5.3. The numbers of ancestral, gained, and lost genes during the evolution of bilateral animals. Species name is given on the right-hand side of each external node. Ancestral species of our interest are labeled by **α** to **ζ**. The number within a square box is the number of genes in each extant species or ancestral species. The numbers above and below each branch are the numbers of gained and lost genes, respectively. The divergence times for ancestral nodes **α**, **β**, **δ**, **ε**, and **ζ** are based on the molecular clock (Kumar and Hedges 1998; Nei, Xu, and Glazko 2001) and that for node **γ** is based on the fossil record (Shu et al. 2001). The remaining ancestral nodes are not on the time scale. **A.** Evolution of the HOX family genes. For ancestral nodes **α**, **β**, **γ**, **δ**, and **ζ**, the numbers in italic are the numbers of ancestral HOX genes estimated by other studies, and those above the italic are the estimated numbers in this study. Other studies are as follows: node **α**, Zhang and Nei (1996); node **β**, Holland and Garcia-Fernandez (1996); node **γ**, Wada et al. (2003); nodes **δ** and **ζ**, Stellwag (1999). **B.** Evolution of the entire homeobox gene superfamily. The numbers of gained and lost genes for the external branches are not so reliable (see Results). The notations used are the same as those of figure 5.3A.

A


1000
800
600
400
300
100
MYA
40 Human
36 Rat
38 Mouse
35 Frog
40 Zebra
45 Puffer
10 Tunicate
7 Mosquito
8 Fruitfly
5 C. bri.
6 C. ele.
α
β
γ
δ
ε
ζ

B


1000
800
600
400
300
100
MYA
257 Human
240 Rat
261 Mouse
218 Frog
316 Zebra
265 Puffer
102 Tunicate
85 Mosquito
102 Fruitfly
72 C. bri.
113 C. ele.
α
β
γ
δ
ε
ζ

BIBLIOGRAPHY

Abdelkhalek, H. B., A. Beckers, K. Schuster-Gossler, M. N. Pavlova, H. Burkhardt, H. Lickert, J. Rossant, R. Reinhardt, L. C. Schalkwyk, I. Muller, B. G. Herrmann, M. Ceolin, R. Rivera-Pomar, and A. Gossler. 2004. The mouse homeobox gene Not is required for caudal notochord development and affected by the truncate mutation. Genes Dev. **18**:1725-1736.

Adacchi, J., and M. Hasegawa. 1996. MOLPHY: A Computer Program Package for Molecular Phylogenetics. Version 2.3.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature **387**:489-493.

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139**:1067-1076.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389-3402.

Alvarez-Buylla, E. R., S. J. Liljegren, S. Pelaz, S. E. Gold, C. Burgeff, G. S. Ditta, F. Vergara-Silva, and M. F. Yanofsky. 2000a. MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. Plant J. **24**:457-466.

Alvarez-Buylla, E. R., S. Pelaz, S. J. Liljegren, S. E. Gold, C. Burgeff, G. S. Ditta, L. Ribas de Pouplana, L. Martinez-Castilla, and M. F. Yanofsky. 2000b. An

ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc. Natl. Acad. Sci. USA **97**:5328-5333.

Amores, A., T. Suzuki, Y. L. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J. H. Postlethwait. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res. **14**:1-10.

Andrade, M. A., C. Perez-Iratxeta, and C. P. Ponting. 2001. Protein repeats: structures, functions, and evolution. J. Struct. Biol. **134**:117-131.

Aparicio, S., K. Hawker, A. Cottage, Y. Mikawa, L. Zuo, B. Venkatesh, E. Chen, R. Krumlauf, and S. Brenner. 1997. Organization of the Fugu rubripes Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. Nat. Genet. **16**:79-83.

Becker, A., K. Kaufmann, A. Freialdenhoven, C. Vincent, M. A. Li, H. Saedler, and G. Theissen. 2002. A novel MADS-box gene subfamily with a sister-group relationship to class B floral homeotic genes. Mol. Genet. Genomics **266**:942-950.

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

Becker, A., K. U. Winter, B. Meyer, H. Saedler, and G. Theissen. 2000. MADS-Box gene diversity in seed plants 300 million years ago. Mol. Biol. Evol. **17**:1425-1434.

Bharathan, G., B. J. Janssen, E. A. Kellogg, and N. Sinha. 1997. Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Proc. Natl. Acad. Sci. USA **94**:13749-13753.

Blair, J. E., K. Ikeo, T. Gojobori, and S. B. Hedges. 2002. The evolutionary position of nematodes. BMC Evol. Biol. **2**:1-7.

Brocks, J. J., G. A. Logan, R. Buick, and R. E. Summons. 1999. Archean molecular fossils and the early rise of eukaryotes. Science **285**:1033-1036.

Burglin, T. R. 1997. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic Acids Res **25**:4173-4180.

Burglin, T. R. 1994. A comprehensive classification of homeobox genes. Pp. 25-72 *in* D. Duboule, ed. Guidebook to the Homeobox Genes. Oxford University Press, New York.

Butterfield, N. J. 2000. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology **26**:386-404.

Carroll, S. B., J. K. Grenier, and S. D. Weatherbee. 2001. From DNA to diversity. Blackwell Science, Malden, MA.

Cavalier-Smith, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int. J. Syst. Evol. Microbiol. **52**:297-354.

Chen, F., H. Kook, R. Milewski, A. D. Gitler, M. M. Lu, J. Li, R. Nazarian, R. Schnepp, K. Jen, C. Biben, G. Runke, J. P. Mackay, J. Novotny, R. J. Schwartz, R. P. Harvey, M. C. Mullins, and J. A. Epstein. 2002. Hop is an unusual homeobox gene that modulates cardiac development. Cell **110**:713-723.

Chen, M., and Z. Xiao. 1991. Discovery of the macrofossils in the Upper Sinain Doushantuo Formation at Miaohe, eastern Yangtze Gorges. Sci. Geol. Sinica **4**:317-324.

Chi, N., and J. A. Epstein. 2002. Getting your Pax straight: Pax proteins in development and disease. Trends Genet. **18**:41-47.

Cho, S., S. Jang, S. Chae, K. M. Chung, Y. H. Moon, G. An, and S. K. Jang. 1999. Analysis of the C-terminal region of Arabidopsis thaliana APETALA1 as a transcription activation domain. Plant Mol. Biol. **40**:419-429.

Conway Morris, S. 2002. Ancient animals or something else entirely? Science **298**:57-58.

Cronk, Q. C., R. M. Bateman, and J. A. Hawkins. 2002. Deveopmental Genetics and Plant Evolution. Taylor & Francis, London, UK.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. **52**:477-487.

De Bodt, S., J. Raes, K. Florquin, S. Rombauts, P. Rouze, G. Theissen, and Y. Van De Peer. 2003. Genomewide Structural Annotation and Evolutionary Analysis of the Type I MADS-Box Genes in Plants. J. Mol. Evol. **56**:573-586.

De Robertis, E. M. 1994. The homeobox in cell differentiation and evolution. Pp. 13-23 *in* D. Duboule, ed. Guidebook to the Homeobox Genes. Oxford University Press, New York.

Dermitzakis, E. T., and A. G. Clark. 2001. Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. **18**:557-562.

Dickerson, R. E. 1971. The structures of cytochrome c and the rates of molecular evolution. J. Mol. Evol. **1**:26-45.

Dopazo, H., and J. Dopazo. 2005. Genome-scale evidence of the namatode-arthropod clade. Genome Biol.

Duboule, D. 1994. Guidebook to the Homeobox Genes. Oxford University Press, New York.

Eddy, S. R. 2001. HMMER: Profile hidden Markov models for biological sequence analysis (http://hmmer.wustl.edu/).

Edvardsen, R. B., H. C. Seo, M. F. Jensen, A. Mialon, J. Mikhaleva, M. Bjordal, J. Cartry, R. Reinhardt, J. Weissenbach, P. Wincker, and D. Chourrout. 2005. Remodelling of the homeobox gene complement in the tunicate Oikopleura dioica. Curr. Biol. **15**:R12-13.

Egea-Cortines, M., H. Saedler, and H. Sommer. 1999. Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in Antirrhinum majus. EMBO J. **18**:5370-5379.

Fan, H. Y., Y. Hu, M. Tudor, and H. Ma. 1997. Specific interactions between the K domains of AG and AGLs, members of the MADS domain family of DNA binding proteins. Plant J. **12**:999-1010.

Feng, D. F., G. Cho, and R. F. Doolittle. 1997. Determining divergence times with a protein clock: Update and reevaluation. Proc. Natl. Acad. Sci. USA **94**:13028-13033.

Ferrier, D. E., and P. W. Holland. 2001. Ancient origin of the Hox gene cluster. Nat. Rev. Genet. **2**:33-38.

Galant, R., and S. B. Carroll. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. Nature **415**:910-913.

Garber, R. L., A. Kuroiwa, and W. J. Gehring. 1983. Genomic and cDNA clones of the homeotic locus Antennapedia in Drosophila. EMBO J. **2**:2027-2036.

Glazko, G. V., and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. **20**:424-434.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. **28**:132-163.

Goremykin, V. V., S. Hansmann, and W. F. Martin. 1997. Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes:revised molecular estimates of two seed plant divergence times. Pl. Syst. Evol. **206**:337-351.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16**:1664-1674.

Gu, X., Y. Wang, and J. Gu. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat. Genet. **31**:205-209.

Gu, X., and J. Zhang. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106-1113.

Hahn, M. W., G. C. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J. Mol. Evol. **58**:203-211.

Han, T. M., and B. Runnegar. 1992. Megascopic eukaryotic algae from the 2.1-billion-year-old negaunee iron-formation, Michigan. Science **257**:232-235.

Hartmann, U., S. Hohmann, K. Nettesheim, E. Wisman, H. Saedler, and P. Huijser. 2000. Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. Plant J. **21**:351-360.

Hasebe, M., C. K. Wen, M. Kato, and J. A. Banks. 1998. Characterization of MADS homeotic genes in the fern Ceratopteris richardii. Proc. Natl. Acad. Sci. USA **95**:6222-6227.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia. Mol. Biol. Evol. **11**:65-71.

Henschel, K., R. Kofuji, M. Hasebe, H. Saedler, T. Munster, and G. Theissen. 2002. Two ancient classes of MIKC-type MADS-box genes are present in the moss Physcomitrella patens. Mol. Biol. Evol. **19**:801-814.

Hohe, A., S. A. Rensing, M. Mildner, and R. Reski. 2002. Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-box gene in the moss Physcomitrella patens. Plant Biol. **4**:595-602.

Holland, P. W., and J. Garcia-Fernandez. 1996. Hox genes and chordate evolution. Dev. Biol. **173**:382-395.

Honma, T., and K. Goto. 2001. Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature **409**:525-529.

Huang, H., M. Tudor, C. A. Weiss, Y. Hu, and H. Ma. 1995. The Arabidopsis MADS-box gene AGL3 is widely expressed and encodes a sequence-specific DNA-binding protein. Plant Mol. Biol. **28**:549-567.

Hughes, A. L., and R. Friedman. 2004. Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. J. Mol. Evol. **59**:827-833.

Hughes, A. L., T. Ota, and M. Nei. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7**:515-524.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13-34.

Immink, R. G., S. Ferrario, J. Busscher-Lange, M. Kooiker, M. Busscher, and G. C. Angenent. 2003. Analysis of the petunia MADS-box transcription factor family. Mol. Genet. Genomics **268**:598-606.

Ji, Q., Z. X. Luo, C. X. Yuan, J. R. Wible, J. P. Zhang, and J. A. Georgi. 2002. The earliest known eutherian mammal. Nature **416**:816-822.

Kang, H. G., J. S. Jeon, S. Lee, and G. An. 1998. Identification of class B and class C floral organ identity genes from rice plants. Plant Mol. Biol. **38**:1021-1029.

Kappen, C. 2000. Analysis of a complete homeobox gene repertoire: implications for the evolution of diversity. Proc. Natl. Acad. Sci. USA **97**:4481-4486.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **15**:3059-3066.

Knudsen, B., and M. M. Miyamoto. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. USA **98**:14512-14517.

Kofuji, R., N. Sumikawa, M. Yamasaki, K. Kondo, K. Ueda, M. Ito, and M. Hasebe. 2003. Evolution and divergence of MADS-box gene family based on genome wide expression analyses. Mol. Biol. Evol. **20**:1963-1977.

Kohler, C., L. Hennig, C. Spillane, S. Pien, W. Gruissem, and U. Grossniklaus. 2003. The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. Genes Dev. **12**:1540-1553.

Koonin, E. V., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, I. B. Rogozin, S. Smirnov, A. V. Sorokin, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. **5**:R7.

Kramer, E. M., R. L. Dorit, and V. F. Irish. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the APETALA3 and PISTILLATA MADS-box gene lineages. Genetics **149**:765-783.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917-920.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Lamb, R. S., and V. F. Irish. 2003. Functional divergence within the APETALA3/ PISTILLATA floral homeotic gene lineages. Proc. Natl. Acad. Sci. USA **100**:6558-6563.

Laroche, J., P. Li, and J. Bousquet. 1995. Mitochondrial DNA and monocot-dicot divergence time. Mol. Biol. Evol. **12**:1151-1156.

Lee, H., S. S. Suh, E. Park, E. Cho, J. H. Ahn, S. G. Kim, J. S. Lee, Y. M. Kwon, and I. Lee. 2000. The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in Arabidopsis. Genes Dev. **14**:2366-2376.

Lee, S., J. Kim, J. S. Son, J. Nam, D. H. Jeong, S. Jang, J. Lee, D. Y. Lee, H. G. Kang, and G. An. 2003. Systemic Reverse Genetic Screening of T-DNA Tagged Lines in Rice for the Functional Genomic Analyses: MADS-box genes as a test case. Plant Cell Physiol. **44**:1403-1411.

Lewis, E. B. 1951. Pseudoallelism and gene evolution. Cold Spring Harbor Symp. Quant. Biol. **16**:159-174.

Liljegren, S. J., G. S. Ditta, Y. Eshed, B. Savidge, J. L. Bowman, and M. F. Yanofsky. 2000. SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. Nature **404**:766-770.

Ma, H., and C. dePamphilis. 2000. The ABCs of floral evolution. Cell **101**:5-8.

Ma, H., M. F. Yanofsky, and E. M. Meyerowitz. 1991. AGL1-AGL6, an Arabidopsis gene family with similarity to floral homeotic and transcription factor genes. Genes Dev. **5**:484-495.

McGinnis, W., M. S. Levine, E. Hafen, A. Kuroiwa, and W. J. Gehring. 1984. A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. Nature **308**:428-433.

Meyerowitz, E. M. 2002. Plants compared to animals: the broadest comparative study of development. Science **295**:1482-1485.

Michaels, S. D., and R. M. Amasino. 1999. FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell **11**:949-956.

Michaels, S. D., G. Ditta, C. Gustafson-Brown, S. Pelaz, M. Yanofsky, and R. M. Amasino. 2003. AGL24 acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization. Plant J. **33**:867-874.

Misawa, K., and M. Nei. 2003. Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. J. Mol. Evol. **57**:S290-S296.

Moon, Y. H., H. G. Kang, J. Y. Jung, J. S. Jeon, S. K. Sung, and G. An. 1999. Determination of the motif responsible for interaction between the rice APETALA1/AGAMOUS-LIKE9 family proteins using a yeast two-hybrid system. Plant Physiol. **120**:1193-1204.

Munster, T., W. Deleu, L. U. Wingen, M. Ouzunova, J. Cacharron, W. Faigl, S. Werth, J. T. T. Kim, H. Saedler, and G. Theissen. 2002. Maize MADS-box genes galore. Maydica **47**:287-301.

Munster, T., J. Pahnke, A. Di Rosa, J. T. Kim, W. Martin, H. Saedler, and G. Theissen. 1997. Floral homeotic genes were recruited from homologous MADS-box genes

preexisting in the common ancestor of ferns and seed plants. Proc. Natl. Acad. Sci. USA **94**:2415-2420.

Nam, J., C. W. dePamphilis, H. Ma, and M. Nei. 2003. Antiquity and evolution of the MADS-box gene family controlling flower development in plants. Mol. Biol. Evol. **20**:1435-1447.

Nam, J., J. Kim, S. Lee, G. An, H. Ma, and M. Nei. 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc. Natl. Acad. Sci. USA **101**:1910-1915.

Nei, M. 1969a. Gene duplication and nucleotide substitution in evolution. Nature **221**:40-42.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York, New York.

Nei, M. 1969b. Gene duplication and nucleotide substitution in evolution. Nature **221**:40-42.

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

Nei, M., X. Gu, and T. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94**:7799-7806.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford Press, New York, New York.

Nei, M., P. Xu, and G. Glazko. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc. Natl. Acad. Sci. USA **98**:2497-2502.

Nesi, N., I. Debeaujon, C. Jond, A. J. Stewart, G. I. Jenkins, M. Caboche, and L. Lepiniec. 2002. The TRANSPARENT TESTA16 locus encodes the ARABIDOPSIS BSISTER MADS domain protein and is required for proper development and pigmentation of the seed coat. Plant Cell **14**:2463-2479.

Ogura, A., K. Ikeo, and T. Gojobori. 2005. Estimation of ancestral gene set of bilaterian animals and its implication to dynamic change of gene content in bilaterian evolution. Gene **345**:65-71.

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

Page, R. D., and M. A. Charleston. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. **8**:349-362.

Parenicova, L., S. De Folter, M. Kieffer, D. S. Horner, C. Favalli, J. Busscher, H. E. Cook, R. M. Ingram, M. M. Kater, B. Davies, G. C. Angenent, and L. Colombo. 2003. Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis: New Openings to the MADS World. Plant Cell **15**:1538-1551.

Pelaz, S., G. S. Ditta, E. Baumann, E. Wisman, and M. F. Yanofsky. 2000. B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature **405**:200-203.

Plouhinec, J. L., C. Granier, C. Le Mentec, K. A. Lawson, D. Saberan-Djoneidi, J. Aghion, D. L. Shi, J. Collignon, and S. Mazan. 2004. Identification of the mammalian Not gene via a phylogenomic approach. Gene Expr. Patterns **5**:11-22.

Purugganan, M. D. 1997. The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. J. Mol. Evol. **45**:392-396.

Purugganan, M. D. 1998. The molecular evolution of development. Bioessays **20**:700-711.

Rao, P. V. 1998. Statistical research methods in the life sciences. The Brooks/Cole Publishing Company, Pacific Grove, CA.

Rasmussen, B., S. Bengtson, I. R. Fletcher, and N. J. McNaughton. 2002. Discoidal impressions and trace-like fossils more than 1200 million years old. Science **296**:1112-1115.

Riechmann, J. L., and E. M. Meyerowitz. 1997. Determination of floral organ identity by Arabidopsis MADS domain homeotic proteins AP1, AP3, PI, and AG is independent of their DNA-binding specificity. Mol. Biol. Cell. **8**:1243-1259.

Riechmann, J. L., M. Wang, and E. M. Meyerowitz. 1996. DNA-binding properties of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. Nucleic Acids Res. **24**:3134-3141.

Ronshaugen, M., N. McGinnis, and W. McGinnis. 2002. Hox protein mutation and macroevolution of the insect body plan. Nature **415**:914-917.

Ruddle, F. H., K. L. Bentley, M. T. Murtha, and N. Risch. 1994. Gene loss and gain in the evolution of the vertebrates. Development Suppl.:155-161.

Russo, C. A., N. Takezaki, and M. Nei. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. **13**:525-536.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406-425.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**:301-302.

Savard, L., P. Li, S. H. Strauss, M. W. Chase, M. Michaud, and J. Bousquet. 1994. Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. Proc. Natl. Acad. Sci. USA **91**:5163-5167.

Scott, M. P., and A. J. Weiner. 1984. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proc. Natl. Acad. Sci. USA **81**:4115-4119.

Scott, M. P., A. J. Weiner, T. I. Hazelrigg, B. A. Polisky, V. Pirrotta, F. Scalenghe, and T. C. Kaufman. 1983. The molecular organization of the Antennapedia locus of Drosophila. Cell **35**:763-776.

Seilacher, A., P. K. Bose, and F. Pfluger. 1998. Triploblastic animals more than 1 billion years ago: trace fossil evidence from India. Science **282**:80-83.

Sheldon, C. C., J. E. Burn, P. P. Perez, J. Metzger, J. A. Edwards, W. J. Peacock, and E. S. Dennis. 1999. The FLF MADS box gene: a repressor of flowering in Arabidopsis regulated by vernalization and methylation. Plant Cell **11**:445-458.

Shin, C. H., Z. P. Liu, R. Passier, C. L. Zhang, D. Z. Wang, T. M. Harris, H. Yamagishi, J. A. Richardson, G. Childs, and E. Olson. 2002. Modulation of cardiac growth and development by HOP, an unusual homeodomain protein. Cell **110**:725-735.

Shore, P., and A. D. Sharrocks. 1995. The MADS-box family of transcription factors. Eur. J. Biochem. **229**:1-13.

Shu, D. G., S. C. Morris, J. Han, L. Chen, X. L. Zhang, Z. F. Zhang, H. Q. Liu, Y. Li, and J. N. Liu. 2001. Primitive deuterostomes from the Chengjiang Lagerstatte (Lower Cambrian, China). Nature **414**:419-424.

Soltis, P. S., D. E. Soltis, V. Savolainen, P. R. Crane, and T. G. Barraclough. 2002. Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. Proc. Natl. Acad. Sci. USA **99**:4430-4435.

Sommer, H., J. P. Beltran, P. Huijser, H. Pape, W. E. Lonnig, H. Saedler, and Z. Schwarz-Sommer. 1990. Deficiens, a homeotic gene involved in the control of flower morphogenesis in Antirrhinum majus: the protein shows homology to transcription factors. EMBO J. **9**:605-613.

Stellwag, E. J. 1999. Hox gene duplication in fish. Semin. Cell. Dev. Biol. **10**:531-540.

Stewart, W. N., and G. W. Rothwell. 1993. Paleobotany and the evolution of plants. Cambridge University Press, New York, New York.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA **99**:16138-16143.

Svensson, M. E., and P. Engstrom. 2002. Closely related MADS-box genes in club moss (Lycopodium) show broad expression patterns and are structurally similar to, but

phylogenetically distinct from, typical seed plant MADS-box genes. New Phytol. **154**:439-450.

Svensson, M. E., H. Johannesson, and P. Engstrom. 2000. The LAMB1 gene from the clubmoss, Lycopodium annotinum, is a divergent MADS-box gene, expressed specifically in sporogenic structures. Gene **253**:31-43.

Swofford, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.

Takahashi, K., and M. Nei. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. **17**:1251-1258.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. **12**:823-833.

Talbot, W. S., B. Trevarrow, M. E. Halpern, A. E. Melby, G. Farr, J. H. Postlethwait, T. Jowett, C. B. Kimmel, and D. Kimelman. 1995. A homeobox gene essential for zebrafish notochord development. Nature **378**:150-157.

Theissen, G. 2002. Secret life of genes. Nature **415**:741.

Theissen, G. 2001. Development of floral organ identity: stories from the MADS house. Curr. Opin. Plant Biol. **4**:75-85.

Theissen, G., A. Becker, A. Di Rosa, A. Kanno, J. T. Kim, T. Munster, K. U. Winter, and H. Saedler. 2000. A short history of MADS-box genes in plants. Plant Mol. Biol. **42**:115-149.

Theissen, G., and H. Saedler. 2001. Plant biology. Floral quartets. Nature **409**:469-471.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface, flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**:4876-4882.

Tzeng, T. Y., H. C. Liu, and C. H. Yang. 2004. The C-terminal sequence of LMADS1 is essential for the formation of homodimers for B function proteins. J. Biol. Chem. **279**:10747-10755.

Vandenbussche, M., G. Theissen, Y. Van de Peer, and T. Gerats. 2003. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Res. **31**:4401-4409.

Wada, S., M. Tokuoka, E. Shoguchi, K. Kobayashi, A. Di Gregorio, A. Spagnuolo, M. Branno, Y. Kohara, D. Rokhsar, M. Levine, H. Saiga, N. Satoh, and Y. Satou. 2003. A genomewide survey of developmentally relevant genes in Ciona intestinalis. II. Genes for homeobox transcription factors. Dev. Genes Evol. **213**:222-234.

Wagner, A. 2005. Distributed robustness versus redundancy as causes of mutational robustness. Bioessays **27**:176-188.

Wagner, G. P., C. Amemiya, and F. Ruddle. 2003. Hox cluster duplications and the opportunity for evolutionary novelties. Proc. Natl. Acad. Sci. USA **100**:14603-14606.

Wang, D. Y., S. Kumar, and S. B. Hedges. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals, and fungi. Proc. R. Soc. Lond. B. Biol. Sci. **266**:163-171.

Weigel, D., and E. M. Meyerowitz. 1994. The ABCs of floral homeotic genes. Cell **78**:203-209.

Winter, K. U., A. Becker, T. Munster, J. T. Kim, H. Saedler, and G. Theissen. 1999. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. Proc. Natl. Acad. Sci. USA **96**:7342-7347.

Wolf, Y. I., I. B. Rogozin, and E. V. Koonin. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. **14**:29-36.

Wolfe, K. H., M. Gouy, Y. W. Yang, P. M. Sharp, and W. H. Li. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA **86**:6201-6205.

Xiao, S., Y. Zhang, and A. H. Knoll. 1998. Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. Nature **391**:553-558.

Yang, Y., L. Fanning, and T. Jack. 2003. The K domain mediates heterodimerization of the Arabidopsis floral organ identity proteins, APETALA3 and PISTILLATA. Plant J. **33**:47-59.

Yang, Z. 2002. PAML: a program package for phylogenetic analysis by maximum likelihood, London, England.

Yanofsky, M. F., H. Ma, J. L. Bowman, G. N. Drews, K. A. Feldmann, and E. M. Meyerowitz. 1990. The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature **346**:35-39.

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17**:1081-1090.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296**:79-92.

Zhang, H., and B. G. Forde. 1998. An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. Science **279**:407-409.

Zhang, J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J. Mol. Evol. **50**:56-68.

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

Zhang, J., and M. Nei. 1996. Evolution of Antennapedia-class homeobox genes. Genetics **142**:295-303.

VITA

Jongmin Nam was born in Kyungbook, Republic of Korea, on February 21st, 1973. He received a degree of Bachelor of Sciences in Biology from Hanyang University in February 1997 and a degree of Master of Life Sciences from Pohang University of Science and Technology (POSTECH) in February 1999. He also served as a sergeant in the 998 field artillery battalion from December 1992 to February 1995. In August 2000, he entered the Ph.D. program in Biology at the Pennsylvania State University. Jongmin Nam is a student member of the Society for Molecular Biology and Evolution.

Recent Publications

Nam, J., K. Kaufmann, G. Theissen, M. Nei (2005) A simple method for predicting the functional differentiation of duplicate genes and its application to MIKC-type MADS-box genes. *Nucleic Acids Res*. 33:e12.

Nam, J., J. Kim, S. Lee, G. An, H. Ma, and M. Nei (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proc. Natl. Acad. Sci. USA.* 101:1910-1915.

Nam, J., C. W. dePamphilis, H. Ma, and M. Nei (2003) Antiquity and evolution of the MADS-box gene family controlling flower development in plants. *Mol. Biol. Evol.* 20:1435-1447.

Lee, S., J. Kim, J. S. Son, **J. Nam**, D. H. Jeong, K. Lee, S. Jang, J. Yoo, J. Lee, D. Y. Lee, H. G. Kang, and G. An (2003) Systematic reverse genetic screening of T-DNA tagged genes in rice for functional genomic anaylses: MADS-box genes as a test case. *Plant Cell Physiol.* 44:1403-1411.

An, S., S. Park, D. H. Jeong, D. Y. Lee, H. G. Kang, J. H. Yu, J. Hur, S. R. Kim, Y. H. Kim, M. Lee, S. Han, S. J. Kim, J. Yang, E. Kim, S. J. Wi, H. S. Chung, J. P. Hong, V. Choe, H. K. Lee, J.-H. Choi, **J. Nam**, S. R. Kim, P. B. Park, K. Y. Park, W. T. Kim, S. Choe, C. B. Lee, and G. An (2003) Generation and analysis of end sequence database for T-DNA tagging lines in rice. *Plant Physiol.* 133:2040-2047.

Jeon J. S., S. Lee, K. H. Jung, S. H. Jun, D. H. Jeong, J. Lee, C. Kim, S. Jang, K. Yang, **J. Nam**, K. An, M. J. Han, R. J. Sung, H. S. Choi, J. H. Yu, J. H. Choi, S. Y. Cho, S. S. Cha, S. I. Kim, G. An (2000) T-DNA insertional mutagenesis for functional genomics in rice. *Plant J.* 22:561-70.

Jeon J. S., S. Jang, S. Lee, **J. Nam**, C. Kim, S. H. Lee, Y. Y. Chung, S. R. Kim, Y. H. Lee, Y. G. Cho, G. An (2000) leafy hull sterile1 is a homeotic mutation in a rice MADS box gene affecting rice flower development. *Plant Cell.* 12:871-84.

Sung S. K., G. H. Yu, **J. Nam**, D. H. Jeong, G. An (2000) Developmentally regulated expression of two MADS-box genes, MdMADS3 and MdMADS4, in the morphogenesis of flower buds and fruits in apple. *Planta.* 210:519-28.