



11281

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOMÉDICAS
INSTITUTO DE ECOLOGÍA

**Filogenia y evolución molecular de los
genes de la familia MADS-box en *Arabidopsis thaliana*.**

TESIS
QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS

Presenta
León Patricio Martínez Castilla

Directora de Tesis: Dra. Ma. Elena Alvarez-Buylla Roces

México, D. F., agosto 2004



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

TABLA DE CONTENIDO

Filogenia y evolución molecular de los genes de la familia MADS-box en <i>Arabidopsis thaliana</i>	1
Resumen.....	3
Abstract.....	4
Agradecimientos	5
I. INTRODUCCIÓN GENERAL.....	6
¿Porqué es importante estudiar la evolución de las familias multigénicas que codifican para reguladores transcripcionales?	6
La Familia de genes con caja MADS.	12
Estructura de la tesis.....	15
II. MÉTODOS.....	17
Inferencia Filogenética Bayesiana.	17
Polarización de la filogenia de una familia de genes por el método de árboles reconciliados.	20
Detección de la selección natural positiva a nivel de codones individuales y de codones individuales en linajes específicos.....	23
Detección de presuntos eventos de conversión génica.	26
Reconstrucción de caracteres ancestrales.	27
III. RESULTADOS	33
Genes MADS-box: desarrollo y evolución de planos corporales en plantas.....	33
Los genes MADS-box sufrieron una duplicación anterior a la divergencia de los linajes de las plantas y los animales....	34
Evolución adaptativa en la familia de genes MADS-box de <i>Arabidopsis</i> inferida a partir de la resolución de su filogenia.	35
IV. DISCUSIÓN GENERAL Y PERSPECTIVAS.....	36
Complemento completo de genes tipo MADS-box en <i>Arabidopsis thaliana</i>	37
MADS: ¿Ser o No Ser?.....	38
Historia de los genes MADS.....	42
Fuerzas Evolutivas en Acción Durante la Historia Molecular de los MADS	51
V. REFERENCIAS BIBLIOGRÁFICAS CITADAS.....	57
VI. APÉNDICES.....	68
(1) MADS-box genes: development and evolution of plant body plans. Vergara-Silva, F., Martínez-Castilla, L. y Alvarez-Buylla, E. R. 2000. <i>Journal of Phycology</i> . 36: 803-812.....	69
(2) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Alvarez-Buylla, E. R., Pelaz, S., Liljegren, S. J., Gold, S. E., Burgeff, C., Ditta, G. S., Ribas de Pouplana, Ll., Martínez-Castilla, L. y Yanofsky, M. F. 2000. <i>Proceedings of the National Academy of Sciences, USA</i> . 97: 5328-5333.	80
(3) Adaptive evolution in the <i>Arabidopsis</i> MADS-box gene family inferred from its complete resolved phylogeny. Martínez-Castilla, L. P. y Alvarez-Buylla, E. R. 2003. <i>Proceedings of the National Academy of Sciences, USA</i> . 100: 13407-13412.....	87

Resumen

Los genes MADS-box codifican para reguladores transcripcionales con diversas e importantes funciones biológicas que van desde el desarrollo del miocardio en animales hasta la respuesta a feromonas en levaduras. En las plantas, los genes MADS-box codifican para las proteínas responsables de las funciones homeóticas florales predichas por el modelo ABC de identidad de órganos. Además, regulan el tiempo de iniciación de la floración, la identidad de meristemo y varios aspectos del desarrollo de óvulo, fruto, hoja y raíz; sin embargo, aún no se ha podido dilucidar la función de más de la mitad de los genes MADS-box que se encuentran en el genoma de *Arabidopsis thaliana*. En esta planta, la familia MADS-box tiene al menos 108 miembros, lo que aunado al hecho de que muchos de sus miembros intervienen en el control del desarrollo hace que el estudio de la evolución molecular de esta familia sea especialmente interesante. Por ejemplo, el estudio de los procesos que han moldeado la diversidad de esta familia en un contexto filogenético permitiría hacer una aproximación a la cuestión de cuáles son las fuerzas evolutivas que permiten la permanencia de un par de genes parálogos inmediatamente después de su duplicación; asimismo, si se obtuviera evidencia de que las proteínas codificadas por los miembros de esta familia divergieron por selección natural, eso indicaría que las regiones codificantes de los genes que controlan el desarrollo, y no sólo sus regiones reguladoras, juegan un papel importante en la evolución fenotípica; además, si se descubriera que diferentes regiones de las proteínas con dominio MADS han estado sujetas en diferentes momentos a presiones de selección contrastantes, esto sugeriría que las rutas que siguen estas proteínas para explorar el espacio funcional son diversas puesto que las diferentes regiones están involucradas en diferentes funciones bioquímicas, por ejemplo, la unión a DNA, la unión específica a otras proteínas de esta familia para formar multímeros o el control de la transactivación.

Con el fin de estudiar las cuestiones arriba mencionadas se reconstruyó la filogenia del complemento de genes MADS-box de *Arabidopsis* y con base en esa filogenia se intentó detectar la huella de evolución por selección natural positiva, codón por codón, en los subclados de la filogenia así como en ramas específicas correspondientes a eventos de duplicación. Se encontró que los genes MADS-box se dividen en dos grupos principales derivados de una duplicación previa a la separación de los linajes de plantas, hongos y animales y que la mayoría de los genes con función conocida pertenecen a uno de estos grupos. Los análisis de detección de selección enfocados en los diferentes subclados anidados no detectaron evidencias de evolución por selección positiva en los grupos de genes con función conocida pero si en algunos clados de genes aún no caracterizados. Los análisis enfocados a ramas específicas, que no se aplicaron a todas las ramas, evidenciaron la acción de la selección positiva después de las duplicaciones que originaron a grupos de genes que controlan aspectos de historia de vida así como en un gen que interviene en el desarrollo embrionario. Curiosamente, diferentes regiones de los genes parecen estar sujetas a selección positiva en diferentes momentos evolutivos. Se discuten las implicaciones de estos hallazgos para la comprensión de los mecanismos que moldean la evolución de esta familia de genes en particular y de los genes de reguladores transcripcionales con incidencia en el desarrollo en general.

Abstract

MADS-box genes code for transcriptional regulators with diverse and important biological functions, ranging from heart muscle development in animals to pheromone response in yeast. In plants, MADS-box genes code for the proteins responsible for the floral homeotic functions predicted by the ABC model of organ identity. Moreover, plant MADS-box intervene in the regulation of flowering initiation time, meristem identity and various aspects of flower, fruit, leaf and root development; however, more than half of the MADS-box genes found in the *Arabidopsis thaliana* genome are yet to be characterized functionally. The MADS-box gene complement in the *Arabidopsis* genome contains at least 108 genes, and the fact that many of those genes regulate aspects of development makes the MADS-box family an interesting target of molecular evolution studies. For instance, studying which are the evolutionary processes that have shaped the diversity of this gene family would help to clarify the mechanisms underlying the survival of a paralogue pair after gene duplication; likewise, evidence that the proteins encoded by members of this family diverged under positive natural selection would imply that plant phenotypic evolutions does not depend solely on the evolution of the regulatory regions of genes that control development; also, if different regions of MADS-box genes were found to have evolved under different selective pressures at various times, this would suggest that there are several routes by which these genes and their encoded proteins can acquire novel functions, since different regions encode varying biochemical roles, such as DNA binding, specific binding to other MADS domain proteins, or transactivation control.

In order to explore the above mentioned questions the phylogeny of the MADS-box genes found in the *Arabidopsis* genome was reconstructed and used as a framework to test for evidences of positive natural selection, on a codon by codon basis, on all the nested subclades of the phylogeny as well as on specific branches corresponding to duplication events. Phylogenetic studies revealed that MADS-box genes group in two main subfamilies that are derived from a duplication that predates the separation of the plant, fungal and animal lineages, and that most of the plant MADS-box genes with known function, including the flower homeotic functions, belong to only one of these subfamilies. Positive selection detection analyses performed on the nested subclades did not detect evidences of positive selection on the groups of genes with known function in contrast with the uncharacterized genes, for which evolution under positive selection was detected in some of their clades. The analyses that queried specific branches were not applied to all the branches, but those tested yielded evidence of positive selection after the duplications that originated groups of genes controlling life history aspects and at the origin of a gene that intervenes in embryo development. Interestingly, different gene regions appear to evolve under positive selection at different moments in evolutionary time. The implications of these findings for the understanding of the evolutionary mechanisms that shape the MADS-box family as well as the evolution of transcriptional regulators that direct development are discussed.

Agradecimientos

La realización de esta tesis fue posible gracias al apoyo de CONACYT (número de becario 118092) y de la DGEP (proyecto 202379), así como del Instituto de Ecología de la UNAM, para León Martínez y de los proyectos CONACYT 41848 y PAPIIT IN230002 para Elena Alvarez-Buylla.

Es un placer poder reconocer la participación, en la llegada a puerto de esta *galère du roi*, de la gente que me ayudó en la travesía.

En primer lugar, mi más profundo reconocimiento a Elena Alvarez-Buylla quien siempre me apoyó y me orientó. Elena me dio libertad para explorar territorios nuevos y al mismo tiempo no fue jamás complaciente con mi trabajo ni con mi pensamiento. Verdaderamente, una *sensei* y me siento orgulloso de ser su alumno. Gracias.

Daniel Piñero y Lorenzo Segovia mostraron un entusiasmo temerario en mi trabajo y eso, junto con su orientación atinada, me dio fuerzas para seguir adelante aún en momentos de desesperación. Además, Lorenzo me dio una inaudita carta blanca para jugar con el sofisticado *cluster* computacional llamado MOJOJOJO.

Alejandra Covarrubias, Patricia León, Susana Magallón y Luis Eguiarte leyeron cuidadosamente esta tesis e hicieron críticas y comentarios que la mejoraron sustancialmente. Por supuesto, las limitaciones de la tesis son sólo mi responsabilidad. Gracias por ayudarme a mejorarla y por interesarse tanto.

Gracias a Julio Collado y a los miembros de su laboratorio por apoyarme durante la estancia que hice ahí.

Gracias a todos los integrantes del laboratorio de Genética Molecular y Evolución por su apoyo y por haber hecho de mi estancia una experiencia tan agradable. Gracias a Ángeles, América Castañeda, América Plata, Arturo, Rigo, Elizabeth, Miguel Ángel, Luis, Caroline, Alicia, Fran, Bárbara, Sol, Octavio, Nicolás, Argelia, Mario, Eduardo, Enrique, Amanda, Esther, Alma, Lupita, María, Sun, Anidia, Carlos, Mitzi, Armando, Fernán, Russely, Sara, Tania, Libertad, Julio, Carlos Lara, Alvaro, Pati, Daniel y gracias a todos aquellos que no he incluido por nombre en esta lista.

Adrián se cuece aparte, por eso tiene su propio renglón.

Muy especiales gracias a Rosalinda, quien siempre me echó porras, me ayudó con tramites y creyó en mí. No lo hubiera logrado sin ti.

Gracias a Paco. Las primeras ideas de esta tesis vienen directamente de conversaciones con él. Los proyectos que hemos armado juntos han sido disfrutables, y a veces peligrosas, aventuras.

Gracias a Ale, quien me ha mostrado que la biología evolutiva es un ejercicio placentero y que no es justo sufrirlo. Gracias por las sonrisas, las chelas, por hacernos creer en una universidad universal...

Rodolfo es otro cosmos. La ciencia mexicana, la mundial, tienen suerte de contar con él. Yo tengo el privilegio de su amistad.

René, gracias por impedirme tomarme en serio (aunque no se vea en estos agradecimientos cursis).

Gracias al Çacomixtle.

Más gente indispensable: Carolina, Alicia, Guadalupe, Carmen, Rodolfo Dirzo. Alejandra Valero, Angélica Cibrián, Raúl Cueva, Néstor, Ana Mendoza, Claudia, Andrea, Etzel, Amanda, Lev, toda la gente del laboratorio de Daniel Piñero. A Claudia Berea. A Pati Adán, perdón por desaparecer tras la tesis, prometo resurgir. Sergio, Ivette, Andrea, René, Pati, Alexis, Steven y los otros Johannson. Angélica y Humberto Macías. Reina Castillo, toda la familia Kruse y muy especialmente a mi queridísimo Garsmondo. Gracias a los guayabos.

A mis papás, que una vez más fueron los que más pusieron, no sólo a través de la fundación Castilla-Martínez para el Avance de la Evolución Molecular, sino por su cariño inquebrantable, por saber que lo lograría. Los quiero.

A Selene, lo hice gracias a ti. Por lo que vivamos juntos.

I. INTRODUCCIÓN GENERAL

¿Porqué es importante estudiar la evolución de las familias multigénicas que codifican para reguladores transcripcionales?

Una de las ideas que se han consolidado en la biología evolutiva en los últimos diez años, especialmente a la luz de la secuenciación de genomas completos de organismos modelo y del análisis comparativo de los mecanismos moleculares que subyacen a los procesos de desarrollo de organismos multicelulares, es que la generación de novedades evolutivas depende en gran medida de que un número relativamente pequeño de elementos de regulación genética adquieran nuevos papeles, e intervengan en nuevas rutas de regulación (Carroll, Grenier y Weatherbee, 2001; Gilbert, deSouza y Long, 1997). Se trata de lo que Roth (1988; citado en Wagner 1989) denominó “piratería genética” refiriéndose al hecho de que nuevos genes pueden ser reclutados para el control de procesos de desarrollo con los que previamente no estaban relacionados (ver también Van Valen, 1982). Es probable que la duplicación y la divergencia sean condiciones para que se dé la piratería de genes que intervienen en el control del desarrollo. La intuición que hay detrás de esta idea es que los genes del desarrollo están fuertemente restringidos selectivamente, es decir, están tan comprometidos con la ruta de control de la que forman parte que la exploración de nuevas capacidades les está excluida pues prácticamente siempre implicaría una disminución en la adecuación del individuo en el que se diera esta exploración. En cambio, cuando uno de estos genes es duplicado una de las copias parálogas quedaría en muchos casos relativamente libre de restricciones y podría explorar el espacio funcional, ya sea por procesos neutrales o por el efecto de la selección natural, sin que ello implicara un alto riesgo para la integridad de los procesos de desarrollo del organismo (Ohno, 1970, 1973; Ohta, 2000; Hughes, 1994; Lynch y Force, 2000). Por otro lado, también bastaría con que el gen en cuestión adquiriera elementos de control en *cis* (ver más adelante) que le permitieran actuar en otras vías de regulación. Sin embargo, no está claro cuál es el mecanismo que permite la preservación por largos períodos de tiempo de una gran proporción de los genes duplicados. El modelo clásico predice que los genes duplicados tienen inicialmente funciones completamente redundantes, de manera que una copia puede proteger a la otra del efecto de la selección natural, si el efecto de la dosis no es importante. Bajo estas condiciones, las mutaciones deletéreas ocurren mucho más frecuentemente que las mutaciones benéficas (Lynch y Walsh, 1998), y por lo tanto el modelo clásico predice que el destino más común para un par duplicado debería de ser la fijación de un alelo nulo (un alelo producido por una mutación “de nulidad” o sea, una mutación que impide la transcripción, la traducción o la función normal de la

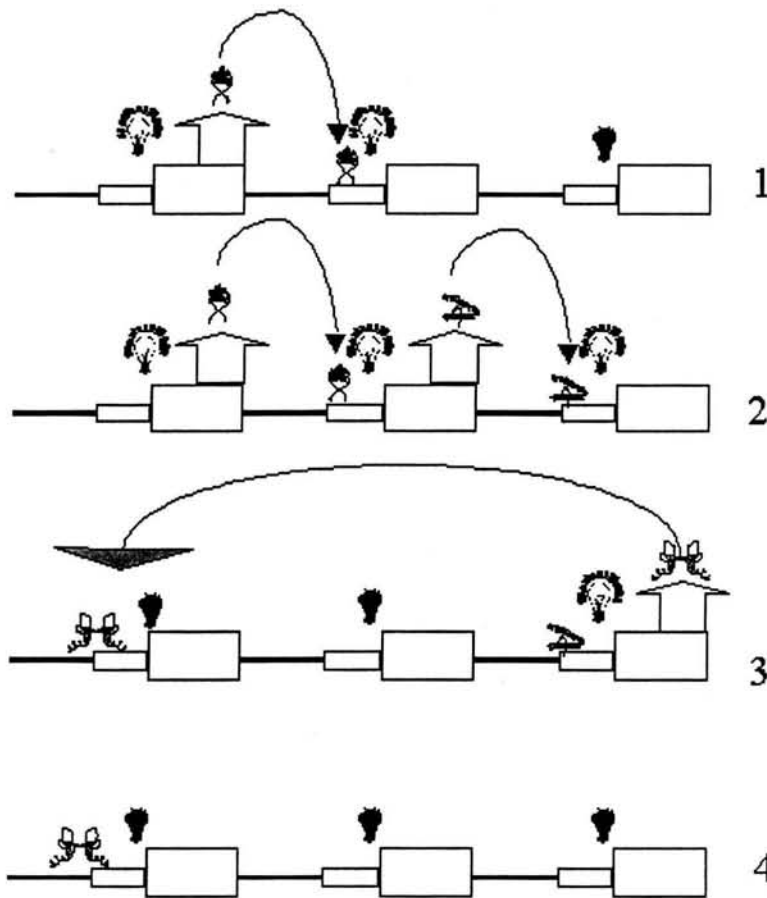
proteína). Al fijarse un alelo nulo para uno de los genes duplicados, esa secuencia se estaría convirtiendo en un pseudogen (Haldane, 1933; Nei y Roychoudhury, 1973; Bailey *et al.*, 1978; Li, 1980; Watterson, 1983). Bajo este modelo, el único mecanismo para la preservación permanente de genes duplicados es la fijación de mutaciones benéficas raras que le dan una nueva función a una de las copias (Ohno, 1970; ver Force *et al.*, 1999). Sin embargo, dado que ambas copias del gen se protegen mutuamente de los efectos de la selección natural y que la frecuencia de individuos con mutaciones de nulidad en ambos loci duplicados es negligible, las mutaciones de nulidad se comportan para fines prácticos como mutaciones neutrales. Este resultado sugiere que la mayoría de los genes duplicados deberían tener una alta probabilidad de convertirse en pseudogenes en un tiempo relativamente corto (Nei y Roychoudhury, 1973; Bailey *et al.*, 1978; Li, 1980; Watterson, 1983). Por ejemplo, si la tasa de producción de mutaciones de nulidad es de 10^{-6} por generación, entonces el tiempo medio hasta la pérdida de función es del orden de pocos millones de generaciones o menos (Force *et al.*, 1999).

No obstante, varias observaciones parecen estar en contradicción con la pérdida rápida de genes duplicados predicha por el modelo clásico. Por ejemplo, Hughes y Hughes (1993) han mostrado que en el sapo tetraploide *Xenopus laevis*, cuando ambos miembros de un par de genes duplicados se expresan, el patrón de divergencia de secuencias indica que están sujetos a selección purificadora. Igualmente, en los eventos de aloploidización del maíz, el 72% de los genes duplicados ha podido evitar la pseudogenización o pérdida de función durante 11 millones de años (Whitkus *et al.*, 1992; Ahn y Tanksley, 1993; White y Doebley, 1998). Además, en los tres principales dominios de la vida existen organismos cuyos genomas están constituidos en un porcentaje alto por genes duplicados (revisados en Zhang, 2003). Así por ejemplo, el 44% del complemento de genes de la bacteria *Mycoplasma pneumoniae* está conformado por genes duplicados, mientras que los porcentajes correspondientes en el archaea *Archaeoglobus fulgidus* y en el eucarionte *Caenorhabditis elegans* son 30 y 49 % respectivamente (Himmelreich *et al.*, 1996; Klenk *et al.*, 1997; Rubin *et al.*, 2000). En la planta *Arabidopsis thaliana* los genes duplicados representan el 65% del complemento de genes (*Arabidopsis* Genome Initiative, 2000) lo cual probablemente esté relacionado con una historia rica en duplicaciones cromosómicas o genómicas (Vision, Brown y Tanksley, 2000; Simillion *et al.*, 2002). En todo caso, estas observaciones sugieren que en lugar de perderse, las copias duplicadas se conservan con suficiente frecuencia como para llegar a constituir una parte importante de los genomas.

Resultaría entonces interesante explorar qué procesos han intervenido en la evolución molecular de genes duplicados. En particular, sería muy importante el establecer cuál es la contribución relativa de la selección natural respecto a la de procesos en principio neutrales, como la tasa de mutación, la tasa de duplicación o la deriva génica, en el mantenimiento de grupos de genes parálogos en el genoma de determinado organismo. Este análisis aportaría datos

empíricos al debate entre el neutralismo y el seleccionismo. Más aún, la detección de evidencias de que la selección natural positiva (también llamada selección Darwiniana positiva) ha jugado un papel en el mantenimiento de los genes parálogos sugeriría fuertemente que un mecanismo de conservación de genes duplicados es la adquisición de nuevas funciones.

Regresemos a nuestra inquietud inicial sobre la evolución por duplicación de los genes reguladores de la transcripción. Parte de la discusión actual acerca del impacto que tiene la evolución de los reguladores transcripcionales sobre la evolución fenotípica contrasta la importancia evolutiva de los factores de transcripción con la de las regiones reguladoras no codificantes. Para entender este debate es necesario presentar muy someramente algunas generalidades del proceso de regulación de la transcripción. Una idea muy importante es que los genes reguladores son a su vez regulados por otros factores de transcripción que modulan la manera en que los primeros producen la proteína que codifican. Dicho de otra manera, los genes que producen proteínas que funcionan como factores de transcripción, que modularán la manera en que otros genes se “prenden” o se “apagan”, son a su vez “prendidos” o “apagados” por factores de transcripción de manera que la producción de las proteínas para las que codifican está sometida a una regulación espacio-temporal muy fina (ver recuadro 1). Ahora bien, aquí es necesario introducir una nota precautoria acerca de la descripción mínima que hemos hecho del proceso de regulación transcripcional. Aunque el objetivo ha sido añadir realismo a nuestra imagen de los mecanismos de regulación de la transcripción, debe de entenderse que en realidad sigue siendo una simplificación extrema. Por ejemplo, no se ha mencionado que un gen puede poseer varias secuencias de reconocimiento diferentes, por lo que puede ser modulado por muchos factores de transcripción, simultáneamente o en secuencia; o que existe cierto grado de flexibilidad en cuanto a cuál es la secuencia específica que es reconocida, de manera que cierto factor de transcripción puede unirse —pero no con la misma afinidad— tanto a la secuencia CCATATATGG como a la secuencia CCTTATAAGG. Tampoco se ha mencionado que un gen que codifica un factor de transcripción no es necesariamente regulado por otros genes sino que puede ser regulado por su propio producto transcripcional, etc. En resumen, una gran cantidad de aspectos han sido obviados en aras de enfatizar los elementos que son relevantes para la discusión acerca del posible contraste entre el impacto evolutivo de las regiones codificantes de genes controladores del desarrollo y el de las regiones reguladoras de estos mismos genes.



Recuadro 1. Una visión esquemática del proceso de regulación de la transcripción por una serie de proteínas de factores de transcripción y por los genes que codifican para estas proteínas. Las rectas negras gruesas representan sectores del genoma que no intervienen en la transcripción de los genes que estamos observando. Los rectángulos grandes representan sectores genómicos que codifican para proteínas; en este caso esas proteínas son factores de transcripción. Las flechas anchas que apuntan hacia arriba pretenden ser un resumen del complejo proceso de transcripción y traducción por el que produce la proteína codificada en la secuencia genómica codificante. Los rectángulos pequeños representan sectores del genoma que no codifican para una proteína dada pero juegan sin embargo un papel en la producción de esa proteína, particularmente mediante la regulación del nivel de transcripción de la región codificante que le sigue inmediatamente. Estas regiones no-codificantes se han llamado regiones *cis*-reguladoras y una definición posible, pero no universalmente aceptada, de la noción de gen es que se trata de la suma de regiones *cis*-reguladoras y de regiones

transcribibles/traducibles que intervienen en la formación de una proteína (o en algunos casos, en la formación de RNAs de transferencia o ribosomales). En un primer momento (1) el gen que se encuentra a la izquierda se “enciende” (foco prendido)—es decir inicia el proceso que desemboca en la producción de la proteína codificada por este gen; en este esquema no nos interesamos sobre la naturaleza de la señal que inicia la transcripción de este gen. Se puede observar que el proceso de transcripción/traducción desemboca en un garabato con una vaga forma de tijera, esta es la proteína codificada por el gen y el hecho de que sea simétrica pretende indicar que se trata de una proteína que funciona como un dímero. La proteína viaja hasta la región *cis*-reguladora del gen que está en medio y encuentra una secuencia específica de reconocimiento a la que se adhiere iniciando así la transcripción/traducción de ese segundo gen (foco prendido). En un segundo tiempo (2) la transcripción/traducción del gen de en medio desemboca en la formación de un garabato que pretende ser la proteína codificada por ese gen. El hecho de que sea asimétrica denota que puede funcionar como monómero. Esta proteína monomérica viaja hasta la región *cis*-reguladora del gen de la derecha en donde encuentra una secuencia de reconocimiento a la que se adhiere iniciando la transcripción/traducción de ese gen. El proceso de transcripción/traducción desemboca en un garabato que parece vagamente una mariposa: la proteína codificada por ese gen y que funciona como dímero. Nótese que mientras tanto continúa la producción de la proteína codificada por el gen de la izquierda lo que garantiza que siga dándose la adherencia entre la proteína del gen de la izquierda y la secuencia de reconocimiento del gen de en medio aunque haya degradación. En un tercer tiempo (3) la proteína codificada por el gen de la derecha viaja hasta la región *cis*-reguladora del gen de la izquierda, que ha seguido produciendo su proteína, pero cuando se une a la secuencia de reconocimiento detiene la transcripción/traducción en lugar de promoverla (foco apagado). En un cuarto tiempo (4) la proteína del gen de la izquierda sigue adherida a la región *cis*-reguladora del de la derecha lo que mantiene apagado a este último y a su vez esto hace que los otros dos genes también estén apagados. Eventualmente la región reguladora del gen de la izquierda y la proteína del gen de la derecha se disociarán y el gen de la izquierda volverá a iniciar el proceso de transcripción/traducción. El énfasis sobre la naturaleza monomérica o dimérica de las formas funcionales de las proteínas tiene únicamente el fin de resaltar que el funcionamiento de algunos factores de transcripción depende no sólo de una correcta asociación a las secuencias de reconocimiento del siguiente gen sino también de la correcta asociación a la pareja de dimerización.

La mayor parte de los genes son controlados al nivel de la transcripción mediante la interacción entre factores citoplasmáticos, o que actúan en *trans* y puntos de control en el DNA, o secuencias reguladoras que actúan en *cis* (Griffiths *et al*, 1999). Los sitios reguladores en los eucariontes pueden ubicarse cerca o lejos del sitio de inicio de la transcripción y a veces se requiere de toda una batería de factores para que se dé la activación completa de ciertos genes. Todos los RNA mensajeros de los eucariontes son sintetizados por la RNA polimerasa II. Para poder alcanzar las tasas máximas de transcripción, la RNA polimerasa II requiere de varios conjuntos de secuencias de control que actúan en *cis*: promotores y elementos de secuencia adicionales llamados elementos próximos al promotor y *enhancers* (potenciadores) en plantas y en células de mamífero y secuencias activadoras corriente arriba en levaduras. Cada uno de estos elementos es reconocido por factores que actúan en *trans* que sirven como elementos de control positivos. Para el inicio de la transcripción se requieren promotores que funcionen adecuadamente; los *enhancers* y los elementos próximos al promotor maximizan la tasa de transcripción para los promotores. Originalmente estos sitios se distinguían por la distancia del sitio de inicio de la transcripción a la que operaban. Los promotores y los elementos próximos al promotor funcionan cerca del sitio de inicio, mientras que los *enhancers* pueden funcionar generalmente a grandes distancias de este punto. Los factores que actúan en *trans*, en particular los factores de transcripción, detectan una "secuencia de reconocimiento" o secuencia reguladora específica del gen que van regular y se unen a esa secuencia. Para simplificar digamos que cuando tiene lugar la asociación, el gen que va a ser regulado inicia o detiene la transcripción o modifica de alguna manera su nivel. Lo importante es que la mayor parte de las secuencias de reconocimiento o reguladoras a las que se va a unir el factor de transcripción, si bien son parte del gen que va a ser regulado (porque no podría funcionar bien sin ellas), no se encuentran en las regiones codificantes de ese gen, es decir, no se van a traducir en una proteína. Ahora bien, desde hace tres décadas, diferentes autores han postulado que la conexión entre evolución molecular y evolución fenotípica debe de buscarse en la evolución de las zonas que regulan en *cis* y no en la evolución de las zonas codificantes de los genes (para una descripción detallada de la regulación de la transcripción y una introducción a la terminología de la regulación en *cis* y *trans*, véase el capítulo 14 de Griffiths *et al.*, 1999). Una de las primeras exposiciones de esta idea fue hecha por Mary-Claire King y Allan Wilson en 1975 en su artículo clásico sobre el grado de similitud a nivel molecular entre chimpancés y seres humanos (King y Wilson, 1975). El DNA de chimpancé es 98.8% idéntico al DNA humano. Cuando la comparación se limita a los genes, la identidad es de 99.5%. En su artículo, King y Wilson sugieren que la explicación de esta discrepancia entre un alto grado de similitud genotípica y la gran diferencia fenotípica radica en que los cambios en anatomía y forma de vida se basan más

frecuentemente en cambios en los mecanismos que controlan la expresión de los genes que en cambios en la secuencia de proteínas (aunque se podría argumentar que la presuntamente enorme diferencia fenotípica entre los seres humanos y los demás primates vivientes es una mera ilusión epistemológica derivada de nuestra incapacidad de evaluarnos objetivamente -puesto que nos encontramos en una situación en la que somos a la vez observadores y observados; por otro lado, también es verdad que hay innegables diferencias cualitativas entre el ser humano y los demás animales).

Variantes de la hipótesis de King y Wilson han sido formuladas en otras ocasiones y quizá la manera de expresarla que alcanza más generalidad sea la siguiente: **los cambios en los sistemas *cis*-reguladores impactan en forma crucial en la regulación de la morfología y la fisiología de un organismo** (Britten y Davidson, 1969, 1971; Arnone y Davidson, 1997; Doebley y Lukens, 1998; Carroll, 2000; Carroll, Grenier y Weatherbee, 2001). Es importante notar, sin embargo, que esta hipótesis tiene versiones menos generales, o si se prefiere, más extremas; en particular, la versión de Doebley y Lukens (1998) se enfoca muy precisamente en la parte no-codificante de los genes de factores de transcripción. Concretamente, Doebley y Lukens escriben: “(nuestra conclusión) es que la evolución de la forma vegetal se llevará a cabo más fácilmente por cambios en las regiones *cis*-reguladoras de los reguladores transcripcionales”. Esta conclusión es correcta, sin embargo la manera como está formulada la ubica cerca de la postura extrema que consistiría en decir que sólo los cambios en las regiones *cis*-reguladoras de los genes reguladores de la transcripción son responsables de la evolución de los fenotipos, al menos en la medida en que la responsabilidad de los cambios fenotípicos se adjudique a los elementos de regulación en *cis* en detrimento del papel que puedan jugar los factores de regulación en *trans*. Uno de los temas centrales de este trabajo es poner a prueba esta idea. Para ello planteamos que si existen evidencias de que la parte codificante de un gen de factor de transcripción evolucionó bajo el efecto de la selección natural positiva, eso implica, por definición, que diferentes formas alélicas de la parte codificante de ese gen tuvieron diferentes efectos en la adecuación de los individuos que las portaban. Si el gen en cuestión interviene en la especificación de la morfología, las diferentes variantes deben de haber producido fenotipos diferentes (si no hubiesen producido fenotipos diferentes no habría evidencia de selección natural positiva). Entonces, si hay evidencia de que la región codificante de un gen regulador de la transcripción evolucionó bajo selección natural positiva eso implica que los cambios en las regiones codificantes, y no solo los cambios en la regiones reguladoras de esos genes, pueden jugar un papel en la evolución de la forma. En esta tesis se mostrará que la parte codificante de por lo menos algunos miembros de una importante familia de reguladores transcripcionales de plantas evolucionó bajo el efecto de la selección natural positiva.

La Familia de genes con caja MADS.

El grupo de secuencias evolutivamente relacionadas entre sí que comprende la mayoría de los *loci* con funciones homeóticas en el desarrollo de las flores es la familia multigénica MADS-box (Shore y Sharrock, 1995). El acrónimo de la familia se deriva de las iniciales de sus primeros cuatro miembros descritos (Norman *et al.*, 1988; Jarvis *et al.*, 1989; Sommer *et al.*, 1990; Yanofsky *et al.*, 1990) mientras que la caja MADS se refiere a una secuencia conservada de aproximadamente 180 nucleótidos. Igual que la “caja homeótica” u *homeobox* (Gehring *et al.*, 1994), la caja MADS codifica para un dominio de unión a DNA que permite que los productos proteínicos que tienen el dominio MADS se comporten como factores de transcripción. Algunas de las proteínas con dominio MADS de plantas (ver apéndices 2 y 3) tienen un segundo dominio altamente conservado que guarda similitud con la estructura secundaria de la proteína keratina del citoesqueleto de animales (el dominio K). El dominio K se une con la caja MADS por una región intermedia I, moderadamente conservada. Finalmente, la región C o carboxilo-terminal, que presuntamente contiene dominios de transactivación, está relativamente menos conservada entre secuencias (Riechman y Meyerowitz, 1997) (ver figura 1 del apéndice 1). Esta es la estructura de las proteínas MADS de plantas mejor caracterizadas funcionalmente hasta ahora. Sin embargo, estudios recientes han descrito nuevos grupos de genes MADS-box de plantas que no tienen la estructura canónica de dominios MADS, I, K y carboxilo-terminal (estructura MIKC) descrita previamente (Alvarez-Buylla *et al.*, 2000a; ver apéndice III). Estos genes más recientemente descubiertos están más cercanamente relacionados con los genes de animales y hongos tipo factor de respuesta sérica (*SRF-like*) que con los genes MADS-box de plantas con estructura MIKC.

Los análisis genéticos y moleculares de los mecanismos que controlan la morfogénesis floral en *Arabidopsis thaliana* (L.) Heynh y en *Antirrhinum majus* L. han permitido desarrollar un elegante modelo genético: el modelo ABC de especificación de la identidad de los órganos florales. En este modelo, las actividades combinadas de un pequeño número de *loci* es responsable de un fenotipo complejo a través de la orquestación del desarrollo (Coen y Meyerowitz, 1991) (ver figura 1B del apéndice 1). Las características principales de este modelo, revisado a profundidad, por ejemplo en Lawton-Rauh *et al.* (2000) se resumen en tres puntos. Primero, el modelo comprende tres campos espaciales de actividad génica parcialmente superpuestos entre ellos (llamados A, B y C) con genes que funcionan exclusivamente en un dominio particular pero no se expresan únicamente en ese dominio. Segundo, el modelo define y predice la identidad de los órganos florales con base en la actividad combinada de los genes de las funciones A, B o C. De acuerdo con esta idea, la

determinación del primer verticilo floral (en donde se encuentran los sépalos) depende de la presencia únicamente de los genes de la función A. La especificación de los pétalos es consecuencia de la participación simultánea de genes de las funciones A y B, y la de los estambres depende de las funciones B y C. Finalmente, la identidad del cuarto verticilo floral (en donde se encuentran los carpelos) es especificada por los genes de la función C actuando solos. El tercer punto que caracteriza al modelo consiste en una relación de antagonismo entre las funciones A y C, de tal forma que los genes de la función A se expresan también en el dominio de los genes C cuando éstos no están y viceversa.

Se puede tomar a los genes MADS-box de *A. thaliana* como guía para el establecimiento de una nomenclatura para el conjunto de los genes del modelo ABC. Todos los miembros canónicos del sistema ABC, excepto uno (*APETALA2*, un gen de la función A) son miembros de la familia MADS-box. El otro miembro de la función A que sí es un MADS-box es *APETALA1*. Los genes de la función B son *APETALA3* y *PISTILLATA*, y *AGAMOUS* es el gen de la función C. Todos los genes MADS-box mencionados aquí se muestran en la figura 2 del apéndice 1. Estudios similares de genética molecular en *A. majus* han demostrado la conservación funcional a los niveles genético y de control del desarrollo entre los genes MADS-box de *A. thaliana* y *A. majus* (Irish y Yamamoto, 1995). Esta conservación también se ha podido observar en otros sistemas modelo de plantas como maíz o petunia, y como se verá, la conservación molecular parece extenderse a todo el clado de las plantas vasculares (Henschel *et al.*, 2002).

Experimentos recientes hechos con *A. thaliana* revelan la existencia de una actividad adicional que hace más compleja la imagen del modelo ABC (Pelaz *et al.*, 2000). Pelaz y colaboradores encontraron que miembros de la subfamilia de genes MADS-box relacionados con el gen *AGL2*, concretamente los ahora llamados genes *SEPALLATA1*, (*SEP1* antes llamado *AGL2*), *SEPALLATA2* (*SEP2* ó *AGL4*) y *SEPALLATA3* (*SEP3* ó *AGL9*) actúan en forma redundante en los tres verticilos internos de las flores para determinar las identidades de pétalo, estambre y carpelo. Las mutaciones en estos genes no afectan los patrones de expresión de mRNA de los genes MADS-box de las funciones ABC sugiriendo que ellos solos no son determinantes de los patrones espacio-temporales de expresión de los genes ABC, pero su función es crucial para las funciones B y C al participar en un tetrámero que controla la transcripción de *AP3*, *PI* y *AG*. Las proteínas codificadas por estos genes forman heterodímeros con las proteínas *PI*, *AP3* y *AG*. Aún existe discusión si estos genes podrían considerarse como una cuarta función "D". Sin embargo, esta función no está activa en dos, sino en tres verticilos adyacentes. Se han hecho análisis funcionales basados en fenotipos de ganancia o pérdida de función para otros genes MADS-box. Estos estudios muestran que los genes de esta familia están involucrados en diversos aspectos de la ontogenia de las plantas,

además del desarrollo floral. Varios genes MADS-box son importantes en la determinación de la identidad del meristemo floral y del tiempo de floración de los meristemas floral y de inflorescencia. *APETALA1*, *CAULIFLOWER*, y *FRUITFUL* controlan redundantemente la arquitectura de la inflorescencia alterando la expresión y la actividad de dos genes que no pertenecen a la familia MADS-box, *LEAFY* y *TERMINAL FLOWER1*, que están involucrados en la transición de las etapas de desarrollo vegetativo a reproductivo del vástago aéreo (Ferrándiz *et al.*, 2000). El gen *SUPPRESSION OF OVEREXPRESSION OF CONSTANS1* (*SOCI*, previamente llamado *AGL20*) es uno de los blancos del gen *CONSTANS*, que promueve la floración en *Arabidopsis* en respuesta a la duración del día. *SOCI*, un gen MADS-box, se expresa en los meristemas de inflorescencia, después se “apaga” en los meristemas tempranos de flor y vuelve a activarse más tarde durante el desarrollo de la flor, lo que sugiere que tiene papeles adicionales que no son evidentes en el fenotipo del mutante sencillo (Samach *et al.*, 2000).

Otros genes MADS-box están involucrados en la determinación del tipo celular. Este es el caso, por ejemplo, de *SHATTERPROOF1* y *SHATTERPROOF2* que determinan redundantemente el desarrollo adecuado de la zona de dehiscencia de los frutos (Liljegren *et al.*, 2000). *FRUITFULL* es requerido para que se dé el patrón normal de división celular, expansión y diferenciación de las valvas de la silicua, para el desarrollo normal de la hoja y de la inflorescencia (Gu *et al.*, 1998). Se ha demostrado que varios otros genes MADS-box están involucrados en el tiempo de floración. *FLC* está cercanamente relacionado con *AGL27* y *AGL31* y los tres tienen amplios patrones de expresión y podrían compartir funciones, pero *FLC* parece tener funciones específicas pues el mutante nulo sencillo *flc* tiene un fenotipo obvio que sugiere que este gen es un represor de la floración (Michaels y Amasino, 1999).

Varios genes MADS-box se expresan predominante o exclusivamente en la raíz de *Arabidopsis thaliana*, lo que sugiere que estos genes también podrían estar involucrados en la diferenciación celular y la morfogénesis de este órgano. A pesar de que aún existe poca información funcional de estos últimos genes durante desarrollo de las raíces, las líneas de cosupresión para uno de los genes MADS-box que se expresan en raíz, *ANR1*, sugieren que este gen es importante en el control de la formación de raíces laterales en respuesta a la disponibilidad local de nitrógeno (Zhang y Forde, 1998). Muchos otros genes MADS-box han sido clonados y sus patrones de expresión han sido caracterizados (Alvarez-Buylla *et al.*, 2000b) (ver figura 2 del apéndice 1) pero no se han publicado análisis funcionales basados en fenotipos mutantes o líneas de sobreexpresión. Sin embargo, sus patrones de expresión proporcionan una primera guía para la caracterización funcional de estos genes y son un punto de partida para estudios más profundos que finalmente nos permitan entender la

evolución de la función de este importante grupo de genes reguladores. Un primer marco de referencia para avanzar en esta dirección es contar con una historia evolutiva de la familia de genes MADS-box. Nuestro grupo de investigación ha contribuido en este campo mediante análisis de reconstrucción filogenética de la familia de genes MADS-box en *Arabidopsis thaliana*.

Estructura de la tesis.

En esta tesis se presentan contribuciones encaminadas justamente a comprender la historia evolutiva de los genes MADS-box y la importancia relativa de las fuerzas evolutivas durante esta historia. La primera contribución constituye un artículo de revisión sobre aspectos generales de los factores de transcripción MADS-box e incluye información pertinente para la discusión acerca del papel que pudo jugar la diversificación de la familia MADS-box en la evolución de la complejidad morfológica de las plantas terrestres (apéndice 1).

En la segunda contribución (apéndice 2) se reconstruye la historia evolutiva de los genes de esta familia en *Arabidopsis thaliana* en el contexto de la evolución de los eucariontes. En esta segunda contribución se exploran eventos en la evolución temprana de los genes MADS-box y se determina una polarización óptima del orden de ramificación. Esta polarización revela la existencia de dos linajes principales en los genes MADS-box de los eucariontes y es la base para plantear escenarios acerca del orden de aparición de los diferentes clados de genes MADS-box de plantas con funciones conocidas y por ello también permite postular hipótesis acerca del origen algunas novedades evolutivas en plantas.

En la parte restante y última de esta tesis (apéndice 3) se exploran los fenómenos que han moldeado la evolución de los representantes de la familia de factores de transcripción MADS-box que se encuentran en el genoma de la planta *Arabidopsis thaliana*. En esta tercera contribución, se aborda la resolución de la filogenia del complemento total de genes MADS-box que son ostensiblemente funcionales en el genoma de *Arabidopsis thaliana*. La estrategia de búsqueda que se utilizó para obtener los árboles que se mencionan en ese artículo es la maximización de la probabilidad posterior en un marco estadístico bayesiano. *Grosso modo*, las filogenias completas obtenidas en esas búsquedas rescatan nuevamente la estructura más profunda de dos linajes principales existentes desde antes de la separación de plantas, hongos y animales, pero también muestran, con alto apoyo estadístico, una estructura bien resuelta en el interior de cada uno de esos dos linajes. Esta estructura incluye clados que, por lo menos en el caso de los genes MADS-box tipo II, que son los que están mejor

caracterizados funcionalmente, agrupan a genes que tienen dominios similares de expresión, o bien son redundantes entre sí, o intervienen en funciones similares. Esta estructura filogenética, junto con la alineación de las regiones más conservadas de los genes se usó para analizar las fuerzas evolutivas que moldearon la secuencia de aminoácidos de las proteínas codificadas por esos genes. Concretamente, se trató de detectar la acción de la selección natural positiva sobre codones individuales suponiendo uniformidad de la presión selectiva a través de diferentes linajes a diferentes niveles de anidamiento filogenético y también se trataron de detectar eventos de selección positiva que hayan actuado sobre codones individuales en linajes específicos seleccionados por criterios biológicos independientes. Finalmente, la sección de conclusiones pone en perspectiva los resultados obtenidos en los artículos de los apéndices, a la luz de otros artículos recientes en los que también se analiza la evolución de los genes MADS-box.

A continuación resumimos los aspectos metodológicos y los conceptos y teorías que los sustentan y que fueron usados durante la investigación doctoral de la que se deriva esta tesis. Estos métodos incluyen aquellos que se utilizaron para detectar los genes MADS-box en las bases de datos públicas, para inferir sus relaciones filogenéticas, para determinar la polaridad de los árboles filogenéticos y para detectar qué codones y qué linajes mostraban evidencia de haber evolucionado bajo el efecto de la selección natural positiva.

II. MÉTODOS

Inferencia Filogenética Bayesiana.

El método que hemos usado para inferir la filogenia de la familia de genes MADS-box es el basado en métodos de máxima verosimilitud e inferencia bayesiana. La inferencia bayesiana de filogenias (tanto de genes como de organismos) está basada en la estimación de la probabilidad posterior de un árbol (Fig. 2). El teorema de Bayes

$$\Pr[\text{árbol} | \text{datos}] = \frac{\Pr[\text{datos} | \text{árbol}] \times \Pr[\text{árbol}]}{\Pr[\text{datos}]}$$

(donde $\Pr[]$ denota la probabilidad de algo y la barra vertical debe de leerse como “dado”) se usa para combinar la probabilidad *a priori* (también llamada probabilidad previa o anterior) de una filogenia ($\Pr[\text{árbol}]$ –“la probabilidad de un árbol”) con la verosimilitud ($\Pr[\text{datos} | \text{árbol}]$ –“la probabilidad de los datos dado el árbol”) para producir una distribución de probabilidades posteriores sobre los árboles posibles ($\Pr[\text{árbol} | \text{datos}]$ –“probabilidad del árbol dados los datos”). Es importante notar que en este contexto, la noción de la probabilidad una filogenia se refiere no únicamente a una topología postulada sino, también, a las longitudes de ramas postuladas para esa topología y a los parámetros del modelo de sustitución de nucleótidos o de aminoácidos que se supone subyace al proceso que generó el árbol. La probabilidad posterior de un árbol en este contexto puede interpretarse como la probabilidad de que ese árbol haya sido producido por los datos que se usan para hacer la inferencia y dados los parámetros del modelo de sustitución elegido; dicho de otra manera, la probabilidad posterior indica que tan creíble es que una filogenia esté reflejando la historia de ramificaciones que realmente ocurrió *siempre y cuando admitamos que la información previa sobre el modelo de sustitución es correcta*. Se pueden hacer entonces inferencias sobre el grupo que se está estudiando basándose en la probabilidad posterior de los árboles. Por ejemplo, el árbol con la más alta probabilidad posterior puede escogerse como la mejor estimación de la filogenia (Rannala y Yang, 1996). Generalmente, todos los árboles posibles se consideran *a priori* como igualmente probables y la verosimilitud se calcula bajo un modelo markoviano estándar de evolución de los caracteres estudiados, como sería, para el caso de los caracteres de secuencias nucleotídicas, el modelo de sustitución de Jukes y Cantor (1969), el de dos parámetros de Kimura (1980), el modelo general reversible en el tiempo (Rodríguez *et al.*, 1990) o modelos más complejos como los

que toman en cuenta la dependencia evolutiva de nucleótidos que están dentro de un mismo codón (Muse y Gaut, 1994; Goldman y Yang, 1994).

Aunque es fácil formular la probabilidad posterior, calcularla involucra integrar para cada árbol todas las combinaciones de longitudes de ramas y de los valores de los parámetros de los modelos de sustitución y sumar estas medidas entre todos los árboles. Esto es imposible de llevar a cabo analíticamente (Huelsenbeck *et al.*, 2001) pero están disponibles una serie de métodos numéricos que permiten aproximar la probabilidad posterior de un árbol. Quizá el más útil de estos métodos sea el método Monte Carlo de cadenas de Markov (MCMC por sus siglas en inglés) (Metropolis *et al.*, 1953; Larget y Simon, 1999; Gilks *et al.*, 1996). La idea básica del MCMC es construir una cadena de Markov (es decir, una matriz de probabilidades de transiciones) que tenga por espacio de fase los parámetros del modelo estadístico (es decir, la topología, las longitudes de ramas y los parámetros del modelo de sustitución) y una distribución estacionaria que sea la distribución de las probabilidades posteriores de los parámetros. Para el problema de la inferencia filogenética, el MCMC involucra dos pasos: (i) se “propone” un nuevo árbol perturbando estocásticamente el árbol actual después de calcular su medida de optimización, por ejemplo su probabilidad posterior, y (ii) este nuevo árbol se acepta o se rechaza con una probabilidad descrita por Metropolis *et al.* (1953) y Hastings (1970) bajo un esquema de decisión en el que es más probable que el nuevo árbol se acepte si su medida de optimización es superior a la del árbol anterior. Si el nuevo árbol es aceptado entonces se le somete a nuevas perturbaciones. Dicho de otra manera, los pasos del MCMC forman algo que se podría conceptualizar como una cadena. En cada paso, se propone una nueva ubicación en el espacio de parámetros como si fuera el siguiente eslabón de la cadena. Habitualmente, esta nueva ubicación propuesta es similar a la actual porque es generada mediante la perturbación aleatoria de sólo algunos de los parámetros en el estado actual de la cadena. Después se calcula la densidad de probabilidad posterior en la nueva ubicación. Si la nueva ubicación tiene una densidad de probabilidad posterior más alta que la de la ubicación actual de la cadena, entonces el movimiento es aceptado, la nueva ubicación se convierte en el nuevo eslabón de la cadena y el ciclo se repite. Si la ubicación propuesta tiene una densidad de probabilidad posterior menor, el movimiento se aceptará solamente durante una proporción (p) del tiempo, donde p es la razón de la probabilidad posterior de la ubicación propuesta comparada con la probabilidad posterior de la ubicación actual (como quién dice, pequeñas disminuciones en la probabilidad posterior se aceptan frecuentemente, mientras que grandes bajones en la probabilidad posterior son penalizados). Si se rechaza la ubicación propuesta, la ubicación actual es la que queda como siguiente eslabón en la cadena (de esa forma, los últimos dos eslabones de la cadena serán idénticos) y se repite el ciclo. Se puede crear una larga cadena de ubicaciones en el espacio de parámetros repitiendo este procedimiento millones de veces.

La cadena tiende a quedarse en regiones de alta probabilidad posterior; en estas regiones, casi cualquier movimiento propuesto representa una disminución en la probabilidad posterior y raramente es aceptado. La proporción del tiempo que una cadena pasa en una región dada del espacio de parámetros puede usarse como una estimación de la probabilidad posterior de esa región. En el caso de la búsqueda de árboles filogenéticos, el algoritmo está diseñado de tal manera que para una cadena de Markov bien construida y ejecutada adecuadamente, la proporción de tiempo que cualquier árbol del espacio matemático de árboles posibles es “visitado”, es decir, evaluados sus parámetros, es una aproximación válida de la probabilidad posterior de ese árbol (Tierney, 1994). Este método de estimación puede hacerse arbitrariamente certero construyendo cadenas suficientemente largas. Es importante hacer notar que aunque el MCMC ha hecho posible el análisis de muchos modelos complejos, no es una panacea, puesto que las cadenas pueden no convergir en las distribuciones estacionarias por varias razones (por ejemplo, un mal mecanismo de propuesta de nuevos estados o que las cadenas no se ejecuten por un tiempo suficientemente largo [Huelsenbeck *et al.*, 2001]).

La inferencia filogenética es un problema difícil sobre todo por el gran número de árboles que pueden describir las relaciones de un grupo de unidades taxonómicas y por las peculiaridades del proceso de sustitución. Por ejemplo, cuando las tasas de sustitución del DNA son altas, las sustituciones múltiples en una posición dada (*multiple hits*) pueden oscurecer la historia de un carácter (de un sitio). De hecho, bajo ciertas condiciones de longitudes de ramas los métodos filogenéticos pueden convergir en un árbol erróneo (Felsenstein, 1978). Los métodos que modelan explícitamente el proceso de sustitución, y por lo tanto corrigen para las sustituciones múltiples, pueden a menudo evitar este problema. Desafortunadamente, los métodos más poderosos (p. ej., los de máxima verosimilitud) sólo pueden usarse en conjuntos de datos pequeños, y los métodos más rápidos (p. ej., muchos métodos de distancia) no aprovechan completamente la información contenida en las secuencias de DNA.

La inferencia bayesiana brinda la posibilidad de abordar el análisis de conjuntos de datos grandes: en lugar de buscar el árbol óptimo propiamente dicho se muestrean los árboles en función de su probabilidad posterior. Una vez que se tiene una muestra así, se pueden discernir características comunes a los árboles de esa muestra. Por ejemplo, dicha muestra puede usarse para construir un árbol de consenso, indicando dentro del árbol la probabilidad posterior de los clados individuales. Esto es a grandes rasgos equivalente a llevar a cabo un análisis de máxima verosimilitud con remuestreo de *bootstrap* (Larget y Simon, 1999) pero es mucho más rápido. Esta es la estrategia seguida por el programa de cómputo “MrBayes” (Huelsenbeck y Ronquist, 2001) con el que se obtuvieron los árboles de los que trata el apéndice 3. Otros programas de cómputo, como BAMBE (www.mathcs.duq.edu/larget/bambe.html) o MAC5

(www.agapow.net/software/mac5) también usan variantes del algoritmo de Monte Carlo de cadenas de Markov y tienen capacidades ligeramente distintas a las de MrBayes, por ejemplo, MAC5 incluye un modelo de sustituciones que prevee la posibilidad de incluir a las inserciones y las deleciones como elementos que contienen información válida que puede ser aprovechada para dilucidar las relaciones filogenéticas, a diferencia de los otros programas, que sólo usan la información contenida en las posiciones que no contienen inserciones o deleciones.

Polarización de la filogenia de una familia de genes por el método de árboles reconciliados.

Cuando se usan filogenias moleculares para inferir las relaciones entre organismos, a menudo se supone implícitamente que los árboles de genes son isomórficos con los árboles de especies, es decir, que aquellos pueden convertirse en estos simplemente sustituyendo el nombre de una secuencia con el nombre del organismo de donde se obtuvo. Sin embargo, la acumulación de datos de secuencias ha mostrado con claridad que la relación entre árboles de genes y árboles de especies es más compleja que una simple correspondencia biunívoca. La siguiente discusión sobre la manera de reconciliar discrepancias se basa en la de Page y Holmes (1998).

La asociación entre dos o más linajes a lo largo de la evolución es un tema recurrente que se extiende sobre diferentes campos de la biología, desde el nivel molecular hasta el macroevolutivo. En todos los niveles ocurren asociaciones históricas en las que linajes de genes o de organismos se encuentran ligados unos a otros y se puede ver a un linaje como si estuviera siguiendo el rastro de otro a lo largo del tiempo con cierto grado de fidelidad. En el nivel molecular cada familia multigénica tiene una historia filogenética que está íntimamente conectada, pero no es necesariamente idéntica, a la historia del linaje del organismo en que residen los genes. Procesos tales como la duplicación, la pérdida de linajes (o el muestreo inadecuado de estos) y la transferencia horizontal de genes pueden producir árboles complejos de genes que difieren de la filogenia de los organismos. Curiosamente, en otro tipo de asociaciones históricas, tanto ecológicas como evolutivas, se pueden encontrar patrones que muestran mucho paralelismo con los que acabamos de mencionar. Por ejemplo, algunos hospederos y sus parásitos (incluidos los virus) pueden tener una larga historia evolutiva de rastreo o seguimiento mutuo que se refleja en similitudes entre sus árboles filogenéticos. Se pueden observar patrones similares en otras asociaciones, como entre insectos y sus plantas hospederas, o entre animales y sus bacterias endosimbiontes, o, a una escala aún mayor, los organismos pueden rastrear la historia geológica de manera que fenómenos como la tectónica de placas se ven reflejados en la filogenia de los organismos.

Uno de los primeros intentos de lidiar con el problema de la complejidad de la relación entre árboles de genes y árboles de organismos fue el concepto de árbol reconciliado introducido por Goodman y colaboradores (1979) para abordar el problema de la discordancia entre el árbol de genes de la hemoglobina de mamíferos y las filogenias comúnmente aceptadas de este mismo grupo taxonómico. Si tenemos un árbol de genes y un árbol de especies que sean incongruentes (Fig. 1.1) pero tenemos razones para pensar que ambos árboles son correctos, podemos preguntarnos bajo que circunstancias pueden ambos árboles ser verdad. Si visualizamos a los genes como si “rastrearán” a las especies (lo cual es intuitivo, pues finalmente los organismos son las “naves” en las que se mueven los genes) entonces podemos hacer encajar al árbol de genes en el árbol de especies (como se muestra en la figura 4 del apéndice 2). La incongruencia entre ambos árboles puede explicarse postulando duplicaciones de genes que originaron juegos de genes parálogos de los cuales sólo algunos han sobrevivido hasta el presente (Fig. 1.2). En este caso los genes a y b son ortólogos entre sí, como también lo son los genes c y d. Dada la duplicación δ en la base del árbol de genes, se hubiera esperado encontrar dos copias de este gen en cada uno de los taxa, de A a D. La presencia de una sola copia de cada uno de estos taxa requiere postular por lo menos tres pérdidas independientes de genes.

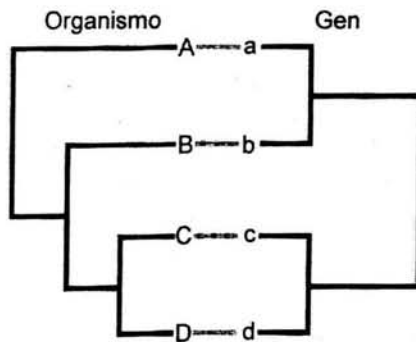


Figura 1.1 Incongruencias entre filogenias de genes y organismos.

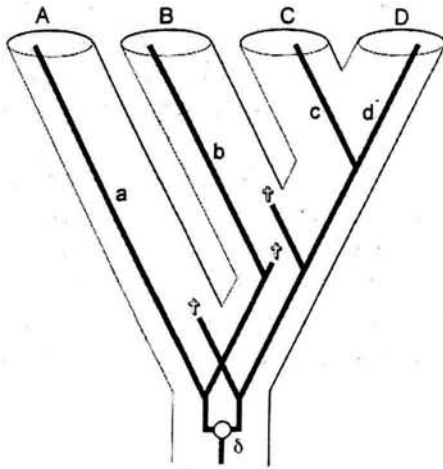


Figura 1.2 La incongruencia entre los árboles de la figura 1.1 puede explicarse postulando una duplicación génica (δ) en la base del árbol de genes, con los árboles *a* y *b* siendo parálogos con los genes *c* y *d*. La presencia de un único gen en cada especie presente nos obliga a postular tres eventos de pérdida de genes (ϕ).

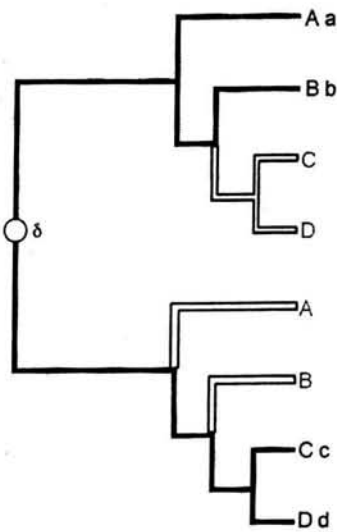


Figura 1.3 Árbol reconciliado para los árboles de genes y especies mostrados en la figura 1.2. El árbol tiene una duplicación de genes (δ) y tres pérdidas (representadas por las ramas huecas).

La figura 1.3 muestra el árbol reconciliado calculado para los árboles de la figura 1.2. Es como si este árbol se hubiera obtenido “desdoblado” el árbol de genes que estaba incrustado en el árbol de especies de la figura 1.2 y aplanándolo sobre la página. Este árbol reconcilia los árboles incongruentes de genes y especies al postular que el árbol de genes observado es en realidad un residuo de un árbol de genes más grande que se originó a partir de la duplicación de

genes en δ . Este árbol más grande es el árbol reconciliado y es el árbol que obtendríamos si no hubieran ocurrido pérdidas de genes o extinciones y si nuestro muestreo de los miembros de las familias multigénicas fuera perfecto. Aquí también, dada la duplicación δ , deberíamos de esperar dos copias del gen en cada una de las especies. El que no observemos estas copias nos obliga a postular pérdidas de genes en las especies A y B y en el ancestro de las especies C y D. Nótese que también es posible que estos genes sí estén presentes pero que no hayan sido detectados. El número total de eventos que postula el árbol reconciliado (una duplicación más tres pérdidas) es el costo de ese árbol. Uno de los ejercicios de que se habla en el apéndice 2 consistió en evaluar el costo de reconciliar un árbol de genes de la familia MADS-box en el que había secuencias provenientes de plantas, hongos y animales, con un árbol de especies generalmente considerado correcto y que también incluía plantas, hongos y animales. Más específicamente, se comparó el costo que tendría reconciliar el árbol de especies con diferentes enraizamientos posibles del árbol de genes con el fin de seleccionar el enraizamiento que fuera menos oneroso en términos de obligarnos a postular eventos de duplicación o pérdida de genes. De esta manera se encontró el enraizamiento óptimo del árbol de genes y, por lo tanto, el orden temporal de los eventos de duplicación que han generado a la familia MADS-box. Esto a su vez llevó a postular la existencia de, por lo menos, dos grandes linajes de genes MADS-box desde antes de que se separaran los linajes de las plantas, los hongos y los animales, como se describe en el apéndice 2.

Detección de la selección natural positiva a nivel de codones individuales y de codones individuales en linajes específicos.

La comparación de las tasas de mutación sinónima (silenciosa) y no sinónima (es decir, que altera la secuencia de aminoácidos) proporciona un importante instrumento para estudiar la evolución de secuencias de DNA (Kimura, 1983; Gillespie, 1991; Yang *et al.*, 2000). Puesto que, en principio, las mutaciones silenciosas son en gran medida invisibles a la selección natural (pero véase Akashi, 1995), mientras que las mutaciones no sinónimas pueden estar bajo fuerte presión selectiva, la comparación de las tasas de fijación entre estas dos clases de mutación proporciona una poderosa herramienta para entender el efecto de la selección natural sobre la evolución de las secuencias moleculares. Una medida que ha destacado en esta clase de estudios es la proporción (razón) entre las tasas de sustitución no sinónima y sinónima ($\omega = d_N/d_S$), llamada también “tasa de aceptación” por Miyata y Yasunaga (1980). Aquí las tasas d_N y d_S se definen como los números de sustituciones no sinónimas y sinónimas por sitio y su proporción, ω , es una medida de la presión de selección al

nivel proteínico. Una $\omega > 1$ significa que las mutaciones no sinónimas ofrecen ventajas de adecuación a la proteína (al individuo que lleva esa proteína) y tienen por lo tanto una mayor probabilidad de fijación en la población que las mutaciones sinónimas. Esta es la definición de trabajo de selección natural positiva (evolución molecular adaptativa o selección Darwiniana positiva) que se ha usado en esta tesis.

La proporción ω casi siempre se ha calculado haciendo un promedio sobre todos los sitios de codones (aminoácidos) y sobre todo el tiempo evolutivo que separa a las secuencias estudiadas. El que esta ω promedio sea mayor que uno es probablemente un criterio excesivamente severo para detectar selección positiva (Yang *et al.*, 2000; véase por ejemplo, Akashi, 1999; Crandall *et al.*, 1999). Las consideraciones biológicas sugieren muchos de los aminoácidos de las proteínas se encuentran bajo fuertes restricciones funcionales impuestas sobre su secuencia (con una ω cercana a cero en las regiones génicas que codifican para dichos aminoácidos). Por ejemplo, en algunos genes de la inmunoglobulina, la tasa de sustituciones no sinónimas en la región que determina la complementariedad (CDR por sus siglas en inglés; esta región también se conoce como región hipervariable) es efectivamente más alta que la tasa de sustituciones sinónimas. Este sesgo a favor de las sustituciones no sinónimas ha sido atribuido a selección por ventaja del heterócigo por un aumento en la diversidad de anticuerpos (Tanaka y Nei, 1989; Li y Graur, 1991). Sin embargo, cuando se toma en cuenta la totalidad del gen de inmunoglobulina la tasa de sustituciones no sinónimas es considerablemente menor que la de sustituciones sinónimas, lo que indica que aún en las inmunoglobulinas, la mayor parte de las mutaciones no sinónimas son desventajosas y se eliminan de la población (Li y Graur, 1991). Hughes y Nei (1989, 1998) han reportado una situación similar para ciertas regiones de los genes del complejo mayor de histoincompatibilidad, en las que la tasa de sustituciones no sinónimas excede la de sinónimas, pero la situación se invierte si se considera la longitud total de los genes. Además, la mayoría de las proteínas parecen encontrarse bajo selección purificadora (negativa) la mayor parte del tiempo (Li, 1997). Probablemente, la evolución adaptativa ocurre en pocos puntos en el tiempo e involucra a pocos aminoácidos de una proteína. En casos así, la proporción ω , promediada sobre el tiempo y sobre los sitios de aminoácidos, no será significativamente mayor que uno, incluso si realmente tuvo lugar la evolución molecular adaptativa. La relativa escasez de casos bien establecidos de adaptación molecular puede deberse en parte a la falta de poder de los métodos de detección que promedian la proporción ω sobre todos los sitios de aminoácidos y sobre todo el tiempo que separa a las secuencias estudiadas.

Un enfoque alternativo consiste en examinar la proporción ω sobre un tiempo evolutivo corto (por ejemplo, sobre linajes específicos en la filogenia) o sobre regiones funcionalmente bien definidas del gen (correspondientemente de la proteína). Messier y Stewart (1997) utilizaron secuencias ancestrales inferidas para identificar dos linajes de una filogenia

de primates que probablemente están experimentando selección diversificadora del gen de la lisozima. Como se mencionó, Hughes y Nei (1998) encontraron que la proporción ω es mayor que uno en una región del complejo mayor de histoincompatibilidad de humanos (MHC por sus siglas en inglés) que codifica para el sitio de reconocimiento del antígeno, mientras que en otras regiones del gen la proporción ω es menor que uno. Cuando no se dispone de información sobre los dominios funcionales de las proteínas o cuando se espera que sólo unos cuantos sitios estén experimentando selección positiva, un enfoque que parece prometedor consiste en idear modelos estadísticos que permitan que la proporción ω varíe entre sitios (Nielsen y Yang, 1998). Modelos con esas características pueden usarse para identificar aminoácidos individuales críticos en una proteína bajo selección positiva.

El modelo de sustitución de codones de Goldman y Yang (1994; ver también Muse y Gaut, 1994) proporciona un marco conceptual para estudiar la evolución de secuencias mediante la comparación de tasas de sustitución no sinónimas y sinónimas. El modelo original presupone una ω única para todos los linajes y todos los sitios, pero se ha extendido recientemente para incorporar la posibilidad de que ω varíe entre los linajes o entre los sitios. Los modelos específicos de linajes (Yang, 1998; Yang y Nielsen, 1998; ver también Muse y Gaut, 1994) permiten proporciones ω variables entre linajes, y por lo tanto son adecuadas para detectar selección positiva a lo largo de linajes específicos (para contestar, por ejemplo, a la pregunta ¿hubo evolución adaptativa después de determinado evento de duplicación o de divergencia?). Estos modelos no presuponen la posibilidad de variación en la proporción ω entre los sitios de las secuencias por lo que detectan selección positiva sólo si la d_N , promediada sobre todos los sitios es mayor que la d_S , promediada sobre todos los sitios. Para detectar sitios individuales bajo selección se han ideado modelos (distribuciones estadísticas) “específicos de sitios” (Nielsen y Yang, 1998; Yang *et al.*, 2000) que permiten que la proporción ω varíe entre sitios pero no entre linajes. Usando esos modelos se detecta selección positiva en los sitios individuales sólo si la d_N promedio sobre todos los linajes es mayor que la d_S promedio. Algunos de los modelos específicos de sitios (explicados detalladamente en Yang *et al.*, 2000) se utilizaron para analizar los genes MADS-box de *Arabidopsis thaliana* en un esquema en el que cada nivel de anidamiento de clados era sometido al análisis mediante varias de las distribuciones de ω posibles. Un resultado interesante es que con estos modelos específicos para sitios sólo se encontró evidencia de sitios bajo selección positiva entre los clados representativos del linaje tipo I que se había definido originalmente en el apéndice 2 (ver apéndices 2 y 3).

Tanto los modelos específicos de linaje como los específicos de sitio pueden carecer de suficiente poder para detectar evolución adaptativa, si esta ocurre únicamente en ciertos puntos temporales o afecta únicamente a algunos de los aminoácidos de una proteína, aún cuando son más poderosos que los modelos que promedian la proporción ω sobre los

sitios y sobre el tiempo. A la luz de estas limitaciones ha surgido el interés en desarrollar modelos que permitan que la proporción ω varíe tanto entre sitios como entre linajes. Uno de estos modelos “de sitio por rama” desarrollados por Yang y Nielsen (2002) se utilizó en el apéndice 3 para detectar selección en linajes de genes MADS-box de *Arabidopsis* que, *a priori*, podrían considerarse con funciones suficientemente diferentes a las de sus clados hermanos como para esperar que en ellos hubiera actuado la selección natural positiva después de la duplicación que les dió origen.

Detección de presuntos eventos de conversión génica.

La conversión génica es un proceso que hace que un segmento de DNA se copie en otro segmento o por lo menos aparezca como ha hecho eso. El segmento “diana” o “blanco” puede encontrarse en el mismo cromosoma que el segmento movilizado, o en otro cromosoma o incluso en un organismo diferente. La conversión génica por segmentos cortos es una fuerza importante en la evolución y a menudo puede ocurrir con frecuencias más altas que las de la mutación puntual (Lehrman *et al.*, 1987; Gyllenstein *et al.*, 1991, 1994; Guttman y Dykhuizen, 1994; Sawyer, 2000). Ahora bien, los patrones de variabilidad genética creados por la recombinación en general y por la conversión génica en particular pueden ser muy parecidos a los efectos de la adaptación molecular (ver, p.ej. McVean, 2001). Cuando hay recombinación, los sitios nucleotídicos en una secuencia no evolucionan a lo largo de un árbol único sino a lo largo de un conjunto de árboles correlacionados (Hudson, 1983). La recombinación da por resultado una aparente heterogeneidad en las tasas de sustitución (Worobey, 2001) y se sabe que, en la reconstrucción filogenética, puede conducir a filogenias en estrella, no resueltas, así como a sesgos en las pruebas de reloj molecular (Schierup y Hein, 2000a, b). Los modelos de heterogeneidad intercodón para las razones de tasas de mutación sinónima y no sinónima usados aquí (ver sección anterior) funcionan bajo el presupuesto de que no hay recombinación, por lo que es legítima una preocupación acerca de que las pruebas estadísticas en las que se usan esos modelos interpreten erróneamente los efectos de la recombinación como evidencia de selección positiva, puesto que la inserción de un fragmento extrínseco puede aparecer bajo estos modelos como una aceleración repentina en la tasa de sustitución. Curiosamente, si la conversión génica por recombinación interlocus sucede recurrentemente, el resultado en las secuencias afectadas por este fenómeno puede ser más bien la homogenización de las secuencias, lo que podría interpretarse, bajo los modelos de variación intercodón en la tasa dN/dS, como evidencia de selección negativa (Nei *et al.*, 2000).

Con el fin de controlar para el efecto de la conversión génica como fuente de error en la detección de selección positiva o negativa en nuestros análisis, sometimos las secuencias alineadas de genes MADS-box a pruebas de detección de conversión génica mediante los programas GENECONV 1.81 (Sawyer, 1999) y MEGA 2.1 (Kumar *et al.*, 2001). Dada una alineación de secuencias de DNA o de aminoácidos, GENECONV busca segmentos alineados para los cuales hay un par de secuencias que son lo suficientemente similares como para sugerir conversión génica en el pasado (Sawyer, 1999). El procedimiento básico es el siguiente: primero, se excluyen de la alineación los sitios monomórficos como un control para sitios constantes o bajo selección purificadora. Entonces se buscan pares alineados de segmentos entre las secuencias que sean (1) idénticos uno a uno e inusualmente largos para ese par de secuencias, o bien (2) tengan una puntuación inusualmente alta para ese par de secuencias, en donde las posiciones que concuerdan uno a uno (*matches*) cuentan como un punto a favor y existe una penalidad para las posiciones que no concuerdan (*mismatches*). La penalidad para *mismatches* depende de la densidad de sitios polimórficos entre las dos secuencias y de un parámetro de intensidad de *mismatch* especificado por el usuario. Las posiciones que no concuerdan podrían ser debidas a mutaciones ocurridas posteriormente dentro de un segmento de conversión génica o posiblemente a reparación incompleta de *mismatches* durante el evento de conversión génica. Los segmentos con altas calificaciones o inusualmente largos son candidatos a ser posibles eventos de conversión génica. A estos eventos se les asignan valores *P* de confianza estadística. Los valores *P* por pares comparan cada fragmento con el máximo que se podría esperar para ese par de secuencias en ausencia de conversión génica. Los valores *P* globales comparan cada fragmento con todos los posibles fragmentos de toda la alineación. GENECONV asigna valores *P* por dos métodos, tanto para comparaciones globales como pareadas. El primer método se basa en permutaciones y es más acertado pero computacionalmente más lento. El segundo método encuentra valores *P* aproximados por un método de Karlin y Altschul (1990, 1993). Los valores *P* de Karlin-Altschul son la base del popular método BLAST para encontrar secuencias coincidentes en bases de datos.

Reconstrucción de caracteres ancestrales.

En el artículo del apéndice 1 se hace una reconstrucción de los tiempos de aparición de ciertas características críticas en la historia de las plantas terrestres. Asimismo, en la sección de perspectivas se menciona la posibilidad de mapear el dominio de expresión hipotético de un gen ancestral para inferir si la evolución de los genes que descendieron por duplicación del gen ancestral evolucionaron por subfuncionalización. Aquí se presenta una somera introducción a las ideas relacionadas con la reconstrucción filogenética de estados ancestrales.

Existen algoritmos de optimización de estados de caracteres diseñados para ser usados bajo casos especiales de suposiciones acerca de la evolución de los caracteres estudiados (Farris, 1970; Fitch, 1971; Hartigan, 1973; Maddison *et al.*, 1984; Swofford y Maddison, 1987). Pero es posible replantear estos algoritmos en un marco conceptual más general usando un método basado en la estrategia de programación dinámica de Sankoff y sus colaboradores (Sankoff, 1975; Sankoff y Rousseau, 1975; Sankoff y Cedergren, 1983; ver también Maddison y Maddison, 1987; Williams y Fitch, 1990). Para caracteres discretos, esta aproximación de “parsimonia generalizada” puede contener las características de casi todas las otras variantes de la reconstrucción por parsimonia (Camin y Sokal, 1965; Farris, 1970, 1977; Fitch 1971). Usado en su forma “cruda”, este algoritmo siempre es computacionalmente más exigente que un algoritmo especializado diseñado para una variante particular de suposiciones sobre la evolución del carácter, pero también puede usarse para dar cabida a suposiciones para las que no hay ningún otro algoritmo disponible. Además, la aproximación generalizada es esencialmente un método de “fuerza bruta” que es más intuitivo que los algoritmos especializados. Aunque nuestra discusión se limitará a caracteres discretos, el método de parsimonia generalizada puede extenderse a caracteres continuos (Sankoff y Rousseau, 1975; Maddison, 1991). En el recuadro 2 se puede ver una versión de este algoritmo, simplificada para las suposiciones de caracteres igualmente ponderados y no ordenados, con el fin de tener una presentación más intuitiva.

El primer paso en la aplicación del algoritmo de parsimonia generalizada es la construcción de una matriz de costos que especifique el costo mínimo de una transformación de cada estado del carácter a cualquier otro estado. Si las suposiciones que subyacen la evolución del carácter se han modelado usando un grafo de estados del carácter es fácil derivar una matriz de costos a partir de ese grafo (matemáticamente un grafo es una colección de puntos y de líneas que conectan a un subconjunto de dichos puntos; en este caso, “grafo” se refiere al árbol filogenético que constituye nuestra hipótesis acerca de las relaciones entre los taxa). El costo entre cualquier par de estados es igual a la suma de los pesos de las ramas que los conectan. Cuando sólo existe una ruta posible entre un par de estados dado, el costo es simplemente la longitud de esa ruta. Cuando existe más de una ruta posible se usa la longitud de la ruta más corta. Si, debido a la presencia de ramas unidireccionales (transformaciones no permitidas), no existe una secuencia de transformaciones por las que determinado estado se pueda transformar en otro, ese costo se define como infinito. Aunque las matrices de costo a menudo son simétricas, no es obligatorio que lo sean; no existe un requerimiento de que el costo de transformar del estado i al estado j sea igual al costo de j a i . Como las matrices de transformación generalmente se especifican en unidades enteras arbitrarias llamadas “pasos” la literatura de muchos programas como MacClade (Maddison y Maddison, 1992) o PAUP* (Swofford, 1991) se refieren a las matrices de costos como *stepmatrices* o “matrices de pasos”.

Cuando se ha especificado la matriz de costos, un algoritmo comienza con los estados observados en los taxa terminales y se mueve a lo largo del árbol para encontrar la reconstrucción más parsimoniosa de estados en los nodos internos. Los diferentes algoritmos de parsimonia, aunque difieren en detalles, se mueven todos en las ramas usando el mismo “ritmo” hacia arriba y hacia abajo y van llevando con ellos la información concerniente a los estados preferidos para los ancestros a medida que se mueven de nodo a nodo (ver recuadro 2). Cuando el algoritmo llega a un nodo y está “listo” para tomar su decisión final sobre el estado ancestral en ese nodo, toma en cuenta la información sobre los nodos por arriba y por abajo para estimar el estado ancestral.

En particular, el algoritmo de programación dinámica opera haciendo dos “pases” sobre el árbol en cuestión. En el primer pase, que se mueve de las puntas hacia la raíz del árbol (“pase hacia abajo”, ver recuadro 2), se toman en cuenta las implicaciones de asignar cada estado del carácter a cada nodo interno (ancestro hipotético). Específicamente, para cada posible asignación i de estados del carácter a un nodo interno p , determinamos la mínima longitud posible para el clado del cual p es la raíz *dada esa asignación*. A esa longitud le llamamos *longitud condicional del clado de p dada la asignación i* y se representa como $L_{p|i}$. La información almacenada en cada nodo en este primer pase consiste, por lo tanto, en una lista de longitudes. En el segundo pase, que se mueve de la raíz hacia las puntas (“pase hacia arriba”, ver recuadro 2) se determina un conjunto de caracteres ancestrales reconstruidos óptimamente usando la información almacenada durante el primer pase e información calculada previamente durante el segundo pase. El resultado final será un árbol tendrán conjuntos de estados asignados para cada carácter. Es posible que siga habiendo ambigüedades, es decir, que algunos nodos tengan más de un estado posible para algunos caracteres. Estas ambigüedades se pueden resolver con varios métodos diferentes (Swofford y Maddison, 1987). El algoritmo de transformación acelerada (ACCTRAN) presupone que los cambios ocurren tan pronto como es posible a medida que nos movemos hacia los nodos terminales de un árbol, maximizando la proporción de homoplasia explicada por reversiones de cambios que ocurrieron con anterioridad. El algoritmo de transformación retrasada (DELTRAN) presupone que los cambios de estado ocurren tan tarde como sea posible a medida que nos movemos hacia los nodos terminales, maximizando la proporción de la homoplasia explicada por cambios paralelos en diferentes ramas. El algoritmo de minimización del valor F (MINF) intenta decidir entre estados ambiguos de manera que las longitudes de las rutas reconstruidas entre taxa existentes en el árbol se ajuste tanto como sea posible a los números observados de diferencias de caracteres. Los métodos ACCTRAN y DELTRAN son sensibles a la ubicación de la raíz, mientras que el algoritmo MINF no lo es (Swofford y Maddison, 1987). Esta es una versión extremadamente simplificada del funcionamiento de los algoritmos por parsimonia; para una visión más detallada, ver Swofford y Maddison (1992).

La máxima parsimonia es quizá el método más usado para reconstruir estados ancestrales de caracteres. Como la parsimonia intenta minimizar el número de eventos evolutivos, existen al menos dos condiciones bajo las cuales la parsimonia puede ser engañosa: cuando las tasas de evolución son rápidas y cuando las probabilidades de ganancias o pérdidas no son iguales. Para lidiar con este problema se ha desarrollado recientemente un conjunto de métodos de reconstrucción de estados ancestrales de caracteres tanto discretos como continuos basados en aproximaciones de máxima verosimilitud (Schulter, 1995; Yang, Kumar y Nei, 1995; Koshi y Goldstein, 1996; Zhang y Nei, 1997; ver también Maddison y Maddison, 2004). Estas aproximaciones usan un modelo explícito de evolución de caracteres para estimar las probabilidades de todas las posibles reconstrucciones de estados de carácter en cada nodo del árbol. Además del modelo de evolución, estas probabilidades se determinan por la distribución de estados de carácter en los taxa terminales, por la tasa de evolución del carácter y por la longitud de las ramas internodales.

Estas características hacen que las reconstrucciones por máxima verosimilitud sean diferentes de las que se obtienen por parsimonia. Mientras que la parsimonia minimiza el número de cambios de estados de carácter, la máxima verosimilitud puede preferir reconstrucciones menos parsimoniosas. A diferencia de la parsimonia, la máxima verosimilitud toma en cuenta la longitud de las ramas. Pero más importante, como la máxima verosimilitud considera toda posible reconstrucción, puede estimar la probabilidad relativa de cada estado del carácter en cada nodo. Estimar la probabilidad de las reconstrucciones ancestrales permite que se tenga más rigor en la prueba de hipótesis.

Para proporcionar una idea de los métodos de máxima verosimilitud, aquí se presentan las características del método de Schluter *et al.* (1997) para caracteres discretos. Schluter y colaboradores han extendido el modelo de Pagel (1994) de un proceso markoviano en tiempo continuo que describe evolución aleatoria de estados del carácter. Estos modelos tienen características importantes: (1) la probabilidad de cambio en un punto temporal en cualquier rama del árbol depende únicamente del estado del carácter en ese momento, no de estados previos (es decir, se trata de un proceso markoviano); (2) las transiciones a lo largo de cada rama son independientes de los cambios en otras partes del árbol (lo que permite que haya cambios no parsimoniosos); (3) las tasas de cambio son constantes a lo largo del tiempo y a lo largo de todas las ramas. Las tasas de cambio entre cualesquiera dos caracteres dados se puede asumir como iguales ($0 \rightarrow 1 = 1 \rightarrow 0$) o desiguales ($0 \rightarrow 1 \neq 1 \rightarrow 0$) y estas tasas se estiman maximizando su verosimilitud respecto a la distribución de caracteres observados en el árbol. Cuando las tasas son desiguales el número de parámetros a ser estimados aumenta rápidamente con el número de estados de carácter. Debido a la dificultad de estimar con precisión múltiples parámetros, Schluter *et al.* (1997) recomiendan usar el modelo de tasas iguales.

Para inferir nucleótidos o aminoácidos ancestrales también existen métodos de reconstrucción basados en estadística bayesiana que usan modelos de evolución molecular para estimar la distribución de las probabilidades *a priori*. Estos modelos pueden incorporar información sobre la secuencia como un todo; así como conocimiento independiente, tal como frecuencias de cambios entre aminoácidos derivadas empíricamente o la información estructural de las proteínas (Koshi y Goldstein, 1996). Estudios preliminares de simulación sugieren que para datos moleculares los métodos de máxima verosimilitud generalmente tienen un mejor desempeño que los de parsimonia, especialmente cuando las secuencias son muy divergentes y los árboles incluyen ramas largas (Cunningham *et al.*, 1998; Zhang y Nei, 1997), pero aún son necesarios estudios que comparen el desempeño de estos métodos con el de los métodos bayesianos (ver, por ejemplo, Huelsenbeck y Bollback, [2001]).

Recuadro 2. Reconstrucción de caracteres ancestrales usando parsimonia (tomado de Cunningham *et al.*, 1998)

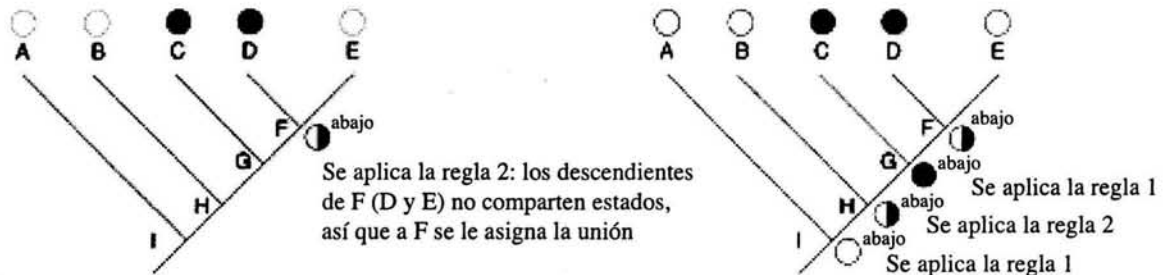
Los estados ancestrales se reconstruyen a menudo usando el criterio de parsimonia. El algoritmo que se ilustra aquí identifica todas las reconstrucciones no ambiguas para caracteres igualmente ponderados y no ordenados y está implementado en sistemas usados habitualmente, como MacClade.

El algoritmo usa un recorrido de "pase hacia abajo" y "pase hacia arriba" (ver figuras) para optimizar los estados ancestrales usando dos reglas:

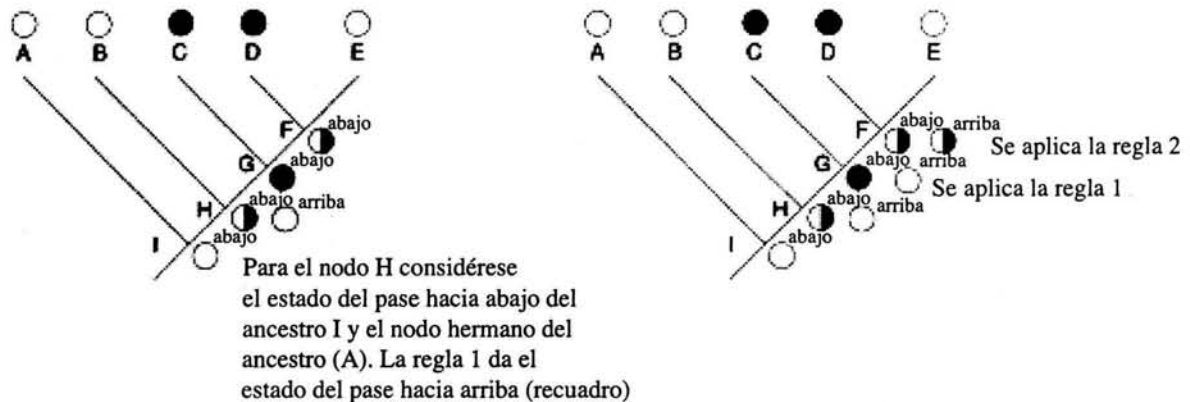
REGLA 1: Si los nodos descendientes comparten estados en común, asigna el conjunto de los estados compartidos al ancestro.

REGLA 2: Si no hay estados compartidos en los nodos descendientes, asigna la union de los estados del descendiente al ancestro.

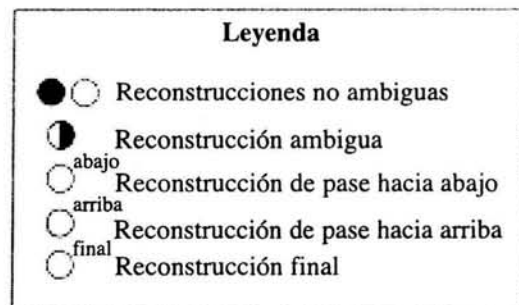
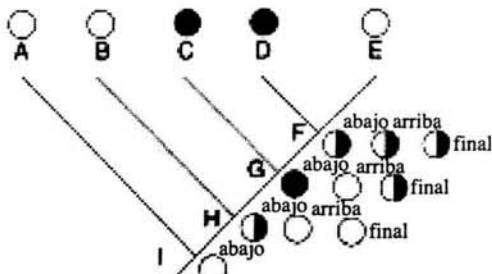
- (1) **Optimización de pase hacia abajo:** Recorre el árbol de arriba a abajo, hacia la raíz, optimizando en cada nodo ancestral.



- (2) **Optimización de pase hacia arriba:** procede "hacia arriba" del árbol, alejándose de la raíz, optimizando en cada nodo ancestral.



- (3) **Optimización final**



Para el estado final de cada nodo(p. ej. el nodo H) considérese el conjunto de pase hacia arriba de ese nodo y los conjuntos de los pases hacia abajo de sus dos nodos descendientes (B y G). Escoge el estado que tenga el mayor número en los tres conjuntos. Si ninguno es mayoría el estado se mantiene ambiguo.

III. RESULTADOS

Genes MADS-box: desarrollo y evolución de planos corporales en plantas.

Los primeros resultados se encuentran detallados en el apéndice 1, *MADS-box genes: development and evolution of plant body plans* (Genes MADS-box: desarrollo y evolución de planos corporales en plantas) por Francisco Vergara-Silva, León Martínez-Castilla y Elena R. Alvarez-Buylla, publicado en *Journal of Phycology* 36: 803-812 (2000). En este artículo revisamos datos funcionales sobre los genes MADS-box, análisis filogenéticos de estos genes y su papel en el desarrollo y evolución de innovaciones morfológicas claves de las plantas. Mapeamos el origen de estructuras morfológicas importantes en diversas etapas del ciclo de vida de diferentes grupos de plantas sobre filogenias de organismos y presentamos aspectos de la genética molecular relevantes para el desarrollo que están relacionados con los genes MADS-box. Nos enfocamos en las estructuras reproductivas del esporofito porque la mayor parte de las caracterizaciones funcionales de genes MADS-box que se han hecho son de genes que tienen que ver con el desarrollo de la flor. Discutimos la evolución de los genes MADS-box en las plantas con flor pero también revisamos estudios en las plantas vasculares sin flor, las gimnospermas (coníferas y gnetales) y los helechos, así como datos preliminares sobre las algas. Sugerimos que los genes MADS-box de plantas, tanto los florales (por ejemplo, los que tienen que ver con el tiempo de floración y la identidad de los meristemas de flor e inflorescencia), como los no florales deberían de estudiarse en forma comparativa. Son necesarios la clonación y los análisis funcionales de genes MADS-box en briofitas, particularmente en el musgo *Physcomitrella patens* (Hedw.) B. S. G. *Physcomitrella patens* es una briofita, es decir, pertenece a uno de los grupos de plantas que más tempranamente en la evolución se separaron del linaje de las angiospermas, al que pertenece *Arabidopsis thaliana*, por lo que la caracterización de sus genes MADS-box arrojaría interesante información acerca de la evolución de esta familia multigénica y acerca del papel que jugó en la evolución fenotípica de las plantas. Además, se ha logrado convertir a *Physcomitrella* en un sistema experimental en las que es posible hacer manipulaciones análogas a las que se hacen en *Arabidopsis*. Por supuesto, también es importante la caracterización de los genes MADS-box en otros grupos de plantas, filogenéticamente más cercanas a las angiospermas. El modelo ABC de especificación de órganos florales es una excelente representación general de una importante red de genes; sin embargo, se requieren herramientas analíticas formales para integrar los datos sobre interacciones complejas de genes en los análisis comparativos. Este y otros enfoques analíticos ayudarán a situar las hipótesis de homología en un marco conceptual evolutivo y de biología del desarrollo.

Los genes MADS-box sufrieron una duplicación anterior a la divergencia de los linajes de las plantas y los animales.

El apéndice 2 contiene los resultados relativos a nuestros primeros análisis filogenéticos y de polarización de la historia evolutiva de los genes MADS-box. Estos resultados se publicaron en el artículo *An ancestral MADS-box gene duplication occurred before the divergence of plants and animals* (Los genes MADS-box sufrieron una duplicación anterior a la divergencia de los linajes de las plantas y los animales) por Elena R. Alvarez-Buylla, Soraya Pelaz, Sarah J. Liljegen, Scott E. Gold, Caroline Burgeff, Gary S. Ditta, Lluís Ribas de Pouplana, León Martínez-Castilla y Martín F. Yanofsky, aparecido en los *Proceedings of the National Academy of Sciences USA* 97: 5328-5333 (2000). En este artículo señalamos que los cambios en los genes que codifican para reguladores de la transcripción son componentes importantes de los mecanismos moleculares de la evolución morfológica. Los genes MADS-box codifican para reguladores transcripcionales con importantes y diversas funciones biológicas. En las plantas, los genes MADS-box intervienen en la regulación del desarrollo de la flor, los frutos, las hojas y la raíz. Los esfuerzos recientes de secuenciación del genoma de *Arabidopsis* han permitido un muestreo casi completo de los representantes de la familia MADS-box en una única especie de planta, algo que no estaba disponible en estudios filogenéticos previos. Para poner a prueba la sospecha existente desde hace algún tiempo de que existe un paralelo entre la evolución de la familia de genes MADS-box y la evolución de la forma en las plantas, se necesita una filogenia de genes que esté polarizada. En este artículo sugerimos que una duplicación de genes previa a la divergencia de plantas, hongos y animales dió origen a dos linajes principales de genes MADS-box: el Tipo I y el Tipo II. Localizamos la raíz de la familia de genes MADS-box de eucariontes entre estos dos linajes. Un nuevo grupo monofilético de dominios MADS de plantas (dominios parecidos a AGL34) parece estar más cercanamente relacionado con los dominios MADS de animales y hongos del tipo de la proteína SRF con quienes forman el linaje de genes tipo I. La mayor parte de las otras secuencias MADS de plantas forman un grupo claramente monofilético junto con los dominios tipo MEF2 de animales y hongos para formar el linaje tipo II. Sólo las proteínas MADS tipo II de plantas tienen un dominio K corriente abajo del dominio MADS. Esto sugiere que el dominio K evolucionó después de la duplicación que dio origen a los dos linajes. Finalmente, un grupo de secuencias con características intermedias y que provienen de plantas podrían revelarse como producto de eventos de recombinación. Estos análisis pueden servir para guiar la búsqueda de secuencias MADS-box en eucariontes basales y para resolver la ubicación filogenética de nuevos genes provenientes de otras plantas.

Evolución adaptativa en la familia de genes MADS-box de Arabidopsis inferida a partir de la resolución de su filogenia.

El apéndice 3 contiene los resultados relativos a la continuación del análisis filogenético y bioinformático del complemento de genes MADS-box de *Arabidopsis* y de las pruebas de detección de evidencias de selección natural positiva en codones y eventos de duplicación específicos en la evolución de estos genes. La duplicación de genes es uno de los substratos de la evolución. Sin embargo, la importancia relativa de la selección positiva por oposición a la relajación de las restricciones en el molde de la divergencia funcional de las copias de genes, producto de duplicaciones, es todavía objeto de controversias. Los genes MADS-box, que codifican para reguladores transcripcionales que son clave en varios aspectos del desarrollo han tenido una historia de sucesivas duplicaciones que los ha llevado a constituir una familia grande de genes. Recuperamos 104 secuencias MADS del genoma de *Arabidopsis*. Nuestras reconstrucciones filogenéticas hechas con métodos bayesianos indican que el linaje tipo II es un grupo monofilético y resuelven una secuencia de ramificaciones de grupos monofiléticos dentro de este linaje. Por otra parte, el linaje tipo I se compone de varios grupos divergentes. Sin embargo, la estructura contrastante de los genes y las diferencias en los patrones de distribución cromosómica entre las secuencias tipo I y tipo II sugieren que han tenido diferentes historias evolutivas y apoyan la colocación de la raíz del árbol de genes en medio de estos dos grupos. Hicimos análisis para detectar selección darwiniana positiva con modelos de sustitución específicos para sitios de codones (que detectan codones sujetos a selección positiva recurrente) y de “rama por sitio de codón” (que detectan codones sujetos a selección positiva durante eventos específicos de divergencia) y los resultados sugieren que los diferentes linajes pudieron haber estado sujetos a diferentes regímenes selectivos puesto que diferentes posiciones codificantes mostraron huellas de haber evolucionado bajo selección natural positiva en diferentes grupos de genes y, además, en algunos grupos de genes no se detectó evidencia de la acción de la selección natural positiva.. Encontramos evidencia de selección darwiniana positiva en la rama que dio origen a los genes que intervienen en el tiempo de floración, los cuales pueden tener un impacto directo en la adecuación de la planta. Encontramos sitios con altas probabilidades de haber evolucionado bajo selección darwiniana positiva tanto en el dominio MADS como en el dominio K, lo cual sugiere que estos dominios jugaron papeles importantes en la adquisición de nuevas funciones durante la diversificación de la familia MADS-box. Los sitios detectados con estos métodos pueden ser objeto de análisis experimentales. Argumentamos que los cambios adaptativos en las secuencias de proteínas con dominio MADS han sido importantes para su divergencia funcional, lo que a su vez sugiere que los cambios en las regiones codificantes de los reguladores transcripcionales han influido en la evolución fenotípica de las plantas.

IV. DISCUSIÓN GENERAL Y PERSPECTIVAS

La actual acumulación exponencial de información sobre el contenido de los genomas de una muestra cada vez más representativa de los seres vivos implica que poco a poco podremos tener una imagen más completa y detallada de la distribución de los genes y de los procesos que moldean esa distribución en los seres vivos. Por ejemplo, para el caso de las plantas actualmente están disponibles, además del genoma anotado de *Arabidopsis thaliana* (*Arabidopsis* Genome Initiative, 2000), versiones en borrador de los genomas del maíz y del arroz (Bennetzen *et al.*, 2001; Yu *et al.*, 2002; Goff *et al.*, 2002), y están en progreso proyectos de secuenciación de los genomas completos de otras plantas, como el jitomate (Tanskley *et al.*, 1992; ver también la página del *National Center for Biotechnology Information* de presentación del proyecto sobre el genoma de *Lycopersicon* en http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4081) y la alfalfa (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3880). Asimismo, existen proyectos para dilucidar el transcriptoma de las plantas modelo ya mencionadas y de otras. La obtención de estos genomas y transcriptomas permitirá obtener información valiosa sobre los procesos que moldean a las familias multigénicas.

Por la importancia de los genes MADS-box durante el desarrollo floral y los fenotipos tan vistosos que producen sus mutantes, esta familia ha atraído la atención de muchos científicos y se está volviendo una familia paradigmática para analizar los procesos evolutivos que dan lugar a las familias multigénicas y el papel de la diversificación de sus miembros en la evolución de los organismos, en particular de las plantas. Durante los años 2003 y 2004 fueron publicados varios estudios (Parenicová *et al.*, 2003; Nam *et al.*, 2003 y 2004; Kofuji *et al.*, 2003; de Bodt *et al.*, 2003; Becker y Theissen, 2003) que tocan aspectos parecidos a las investigaciones que sustentan esta tesis (Martínez-Castilla y Alvarez-Buylla, 2003). En ellos también se pretendía obtener un recuento completo de las secuencias MADS-box de *Arabidopsis* e inferir sus relaciones filogenéticas. Algunos de estos estudios también reportan datos relativos a los genes MADS-box del arroz. Un aspecto interesante de esta multiplicación de esfuerzos es que ha ocurrido una convergencia casi total en los diferentes recuentos de genes MADS-box reportados para el genoma de *Arabidopsis*, de manera que ahora podemos afirmar que se han descubierto y catalogado prácticamente todas las secuencias tipo MADS para esta especie vegetal modelo.

Complemento completo de genes tipo MADS-box en Arabidopsis thaliana

Probablemente el listado de presuntos genes MADS-box de *Arabidopsis* más completo sea el que se encuentra en el artículo de Parenicová, Colombo y colaboradores (Parenicová *et al.*, 2003). En esta lista se reportan 108 secuencias MADS-box que son presuntamente genes funcionales. Esos autores encontraron cuatro secuencias más que las encontradas por Martínez-Castilla y Alvarez-Buylla (2003; apéndice 3), sin embargo, en este último artículo se encontró una secuencia (AGL101) que no fue encontrada por el equipo de Colombo. Por otra parte, el grupo de Nam, Nei y colaboradores (Nam *et al.*, 2003 y 2004), usando una técnica que consiste en explorar mediante PSI-Blast (Altschul *et al.*, 1997) bases de datos de la totalidad de las proteínas anotadas de *Arabidopsis* y de arroz, encontró 107 secuencias MADS-box presuntamente funcionales que incluyen dos nuevas secuencias no detectadas por Parenicová y colaboradores o por Martínez-Castilla y Alvarez-Buylla (*AGL64*, *AGL88* y *AGL105*). Cabe señalar que la exploración de Nam y colaboradores involucra también el primer esfuerzo sistemático para encontrar pseudogenes de la familia MADS-box en el genoma de *Arabidopsis* y que estos esfuerzos han detectado 41 secuencias presuntamente pseudogénicas de las cuales 37 pertenecen al grupo de genes MADS-box tipo I y 4 al tipo II (ver apéndices). Una de estas secuencias (At4g14530) había sido considerada como gen por de Bodt *et al.* (2003) y fue detectada por Martínez-Castilla y Alvarez-Buylla pero fue considerada pseudogen tanto por nosotros como por Nam *et al.* (2004) porque, si bien es muy similar al gen *AGL84*, no parece contener una región de caja MADS. Una conclusión es que el método de Colombo y colaboradores es ligeramente más poderoso para encontrar miembros de una familia multigénica que el de Nei o el de Martínez-Castilla y Alvarez-Buylla. El método de Parenicová y colaboradores (2003) está basado en usar como *query* para las búsquedas de TBLASTN un perfil construido con un modelo markoviano oculto (Eddy, 1998). Concretamente, Parenicová y colaboradores recuperaron de la base de datos Swissprot secuencias provenientes de plantas y que habían sido descritas como pertenecientes a la familia MADS y las usaron como *queries* para recuperar secuencias en una colección no curada de proteínas. Las secuencias así obtenidas fueron alineadas con ClustalW (Thompson *et al.*, 1994) y con las regiones más conservadas se usó el programa HMMER 2.1.1 (<http://hmmmer.wust.edu>) para construir un perfil estadístico de la caja MADS. Este perfil (el "modelo markoviano oculto") se usó posteriormente para encontrar nuevos miembros de la familia en la base de datos del genoma de *Arabidopsis*. Pero cabe notar que el método del laboratorio de Colombo es muy eficaz para detectar los pseudogenes de una familia multigénica y resalta el hecho de que hay muchos más pseudogenes en la subfamilia tipo I que en la tipo II. Este aspecto se retoma más adelante.

MADS: ¿Ser o No Ser?

Es importante considerar que el número relativamente grande de secuencias MADS-box que se consideran genes funcionales en el genoma de *Arabidopsis* podría estar sobreestimado debido a la inclusión de secuencias que en realidad sean pseudogenes o que no pertenezcan a esta familia. De hecho, Kofuji y colaboradores (Kofuji *et al.*, 2003) argumentan que la mayoría de las secuencias MADS-box de tipo I (o de “tipo M” según la nomenclatura propuesta por ellos) en el genoma de *Arabidopsis* serían pseudogenes, puesto que con el método de macroarreglos de mRNA montado por esos autores la mayoría de estas secuencias no se expresó en ninguno de los siete tipos de tejidos que pusieron a prueba o en todo caso se expresó muy debilmente. Esto, junto con la observación de que las ramas filogenéticas que llevan a los genes tipo I son relativamente más largas, lleva a los autores de ese estudio a sugerir que las secuencias tipo I se encuentran bajo restricciones funcionales más débiles que las de los genes tipo II (llamados genes MIKC en su estudio). Si esto fuera cierto, le daría más credibilidad a la idea de que los genes tipo I fueran pseudogenes. Kofuji y colaboradores parecen sentirse atraídos por esa posibilidad, puesto que plantean varios “escenarios” que explicarían tanto un origen pseudogénico para estas secuencias como su presunto mantenimiento por largos períodos de tiempo a pesar de que, siendo pseudogenes, deberían de degradarse bastante rápidamente en la evolución como para que no fuera posible detectarlos con los métodos basados en homología posicional. Por ejemplo, puesto que muchas secuencias tipo I no tienen intrones o bien tienen solamente uno, es posible pensar que se originaron por retrotranscripción de RNAs mensajeros. Si esto fuera cierto sería común encontrar secuencias poly-A en la región flanqueante 3’ de las secuencias tipo I, pero solamente en At5g49490 (*AGL83*) Kofuji y colaboradores (2003) encontraron secuencia poly-A a menos de 1000 pares de bases del codón de parada del marco abierto de lectura, por lo que concluyen que es probable que estas secuencias no tengan un origen retrotranscripcional. Si realmente hubiera un gran número de genes tipo I que no fueran funcionales, una posible explicación sería que estos pseudogenes se hubieran originado a partir de los genes tipo II. Kofuji y sus colaboradores admiten que a primera vista esto contradice la ausencia de motivos de caja K en las secuencias tipo I, pero de acuerdo con ellos, la caja K debería de degradarse rápidamente y ser irreconocible en un tiempo mucho más corto que la caja MADS y, entonces, sería normal encontrar secuencias que tuvieran una caja MADS relativamente bien conservada sin caja K alguna. Sin embargo, las reconstrucciones filogenéticas de esta familia de genes hechas únicamente con la región de la caja MADS (p. ej., Parenicová *et al.*, 2003; Nam *et al.*, 2003 y 2004) contradicen esta idea, puesto que en estas reconstrucciones las secuencias tipo II (incluyendo las que Parenicová *et al.* llaman MIKC y Mδ) tienden a agruparse y a excluir a las secuencias tipo I, lo

que no debería de suceder si las secuencias tipo I fuesen derivadas recientes de las tipo II. Kofuji y colaboradores sugieren que la agrupación en las filogenias de las secuencias tipo I, a pesar de no ser derivadas de un ancestro común, podría explicarse por el fenómeno de “atracción de ramas largas” (Felsenstein, 1978). Sin embargo, los métodos filogenéticos basados en modelos explícitos de evolución molecular y cuya función de optimización se basa en la máxima versimilitud o en estadísticos derivados como la probabilidad posterior bayesiana son mucho más robustos frente al efecto de atracción de ramas largas que los métodos de parsimonia o los métodos de distancia (Swofford, Olsen, Waddell y Hillis, 1996). Ahora bien, tanto Parenicová *et al.* (2003) como Martínez-Castilla y Alvarez-Buylla (2003, ver apéndice 3), recurrieron a la filogenética bayesiana para reconstruir el árbol de las secuencias MADS-box de *Arabidopsis* y encontraron que los genes tipo II formaban un grupo monofilético separado de los genes tipo I, los cuales formaban entre ellos varios grupos filogenéticos. Finalmente, tanto Parenicová y colaboradores (2003) como Nam y colaboradores (2004) encontraron que al incluir algunas de las secuencias MADS-box del genoma de arroz en las reconstrucciones, algunas se agrupaban con secuencias tipo II de *Arabidopsis* y otras con diferentes subgrupos de las secuencias tipo I. Estas observaciones contradicen la idea de que las secuencias tipo I de *Arabidopsis* son derivadas recientes de genes tipo II, por el contrario, esta información indica que ambos grupos estaban bien constituidos al menos de desde antes de la separación de los linajes del arroz y de *Arabidopsis*.

Sin embargo, aún queda abierta la posibilidad de que las secuencias tipo I sean pseudogenes. Los autores que muestran reservas para otorgarle estatus de genes a las secuencias MADS-box de tipo I recientemente encontradas en *Arabidopsis* se basaron principalmente en estudios de expresión. Por ejemplo, el grupo de Kofuji (Kofuji *et al.*, 2003) reporta haber detectado señal de expresión para 49 de las 105 secuencias MADS-box encontradas por ellos en *Arabidopsis*. Pero es importante notar que este equipo no pudo encontrar señal de expresión para los genes tipo I (*AGL39* y *AGL40*) cuya expresión estaba reportada en las bases de datos de *expressed sequence tags* (ESTs). Sin embargo el grupo de Kofuji (2003) menciona también (sin mostrar datos) que con métodos más sensibles (RT-PCR) pudo observar señal positiva para 42 secuencias adicionales para las cuales no se había detectado expresión con el método basado en mRNA. Por lo tanto, este grupo observó expresión, de una manera u otra, para 97 de las 105 secuencias detectadas por ellos. Curiosamente, entre las secuencias para las que se detectó expresión por alguno de los dos métodos están cuatro de los genes encontrados por ese equipo que no habían sido previamente considerados como genes funcionales (*AtMADS1*, *AtMADS3*, *AtMADS4*, *AtMADS5*). Entonces podemos concluir que el hecho de que con un método u otro se haya detectado expresión para la

mayor parte de las secuencias estudiadas de ambos grupos apoya la idea de que la mayor parte de las secuencias tipo MADS son genes funcionales.

El estudio de De Bodt y colaboradores (2003) proporciona evidencia adicional que apoya la propuesta de que los genes tipo I no son pseudogenes. Estos autores encuentran que varios de los subgrupos de secuencias MADS-box tipo I incluyen representantes provenientes tanto de *Arabidopsis* como de arroz y que los motivos característicos de estos subgrupos están conservados entre estas secuencias separadas hace alrededor de 200 millones de años (Wilkstrom *et al.*, 2001), lo que indicaría restricciones funcionales relativamente fuertes. De todas formas, estos autores no pierden de vista que las ramas más largas que conectan a las secuencias tipo I y el hecho de que los genes tipo I parecen estar ausentes casi por completo de las bases de datos de secuencias expresadas podrían indicar que este tipo de secuencias se encuentra relativamente más libre de restricciones funcionales. Sin embargo es importante notar que la presencia de ramas largas en sus clados también podría sugerir que las secuencias tipo I han estado sujetas a mayores presiones de selección positiva (diversificadora) en lugar de estar libres de restricciones.

En el estudio de Parenicová y colaboradores (2003) también se discute si las secuencias tipo I son pseudogenes, pero sus datos de expresión y conservación de motivos, así como la formación de grupos monofiléticos dentro del linaje de genes MADS tipo I sugieren que ese no es el caso. Sin embargo, los genes tipo I son intrigantes porque casi no se han descrito mutantes de este tipo de genes MADS-box y, en general, no se han caracterizado funcionalmente (hasta ahora, sólo el gen *PHRES1*, antes llamado *AGL37*, muestra un fenotipo obvio cuando está mutado). Parenicová y colaboradores (2003) consideran hipótesis para explicar esta observación. La primera es que los mutantes de pérdida de función entre los miembros de este grupo son letales en etapas tempranas del desarrollo. Sin embargo, estos autores mencionan que han obtenido mutantes de inserción en varios de estos genes, varios de los cuales son viables por lo que, en una primera aproximación, los genes tipo I no parecerían ser más indispensables que los tipo II. Otra posibilidad es que estos genes sean redundantes entre sí, como de hecho ocurre entre varios genes MADS-box tipo II como los del grupo *API/CAL/FUL* (Ferrándiz *et al.*, 2000), los genes del grupo *SEP* (Pelaz *et al.*, 2000) o los genes *SHP* (Liljegren *et al.*, 2000). Esta posibilidad sería compatible con la observación hecha por Martínez-Castilla y Alvarez-Buylla (2003, ver apéndice 3) de que la selección natural positiva no parece ser prevalente entre los clados de aparición más reciente (clados menos incluyentes, cuyos nodos basales son menos profundos) de los genes tipo I y, en cambio, fue posible detectarla en algunos nodos más profundos, lo cual podría indicar un proceso de divergencia funcional después de algunas de las duplicaciones ancestrales que originaron a los subclados de este grupo de secuencias, y que quizá duplicaciones más recientes involucraron procesos

neutrales que quizá no conllevaron divergencia funcional. Finalmente, el grupo de Colombo considera la posibilidad de que los genes MADS-box tipo I estén involucrados, en contraste con los genes tipo II estudiados hasta ahora, en funciones más sutiles que las de los tipo II, posiblemente incluso no relacionadas con el desarrollo, y que requieran un examen fenotípico más estricto y cuantitativo para ser detectados o que incluya pruebas bajo diversos ambientes, la construcción de combinaciones múltiples de mutantes, perfiles de transcritos y análisis proteómicos y de metabolitos.

En el artículo de Nam y colaboradores (2004) se toma una aproximación más pragmática a la cuestión de si los genes tipo I son funcionales o no: simplemente se definen como pseudogenes todas aquellas secuencias halladas por este equipo que tuvieran por lo menos un codón de parada dentro de la caja MADS. Esta es una definición inequívoca de pseudogen. Pero de esta forma, prácticamente todas las secuencias consideradas en los otros estudios son consideradas por Nei y colaboradores como genes presuntamente funcionales. Sin embargo, otro de sus resultados interesantes es que estos genes parecen estar evolucionando bajo un régimen de selección purificadora más débil que el de los genes tipo II. Aún así, los genes tipo I podrían ser funcionalmente importantes, como indica la disminución en las tasas de mutaciones no sinónimas y sinónimas para pares de genes tipo I observada por Nam y colaboradores (2004).

Nuestros propios análisis (Martínez-Castilla y Alvarez-Buylla, 2003, ver apéndice 3) señalan que los miembros de algunos de los grupos de genes tipo I de *Arabidopsis* han sufrido selección natural positiva durante su divergencia posterior a las duplicaciones. Esto es un fuerte indicio de que al menos los genes que forman parte de estos clados han adquirido funciones suficientemente importantes como para que hayan afectado sensiblemente la adecuación de los organismos que llevaban las formas divergentes de los productos de las duplicaciones. Es interesante señalar que un estudio del transcriptoma de distintas etapas reproductivas de *Arabidopsis* encontró expresión significativa varios genes MADS tipo I en las etapas de desarrollo embrionario y de desarrollo de los gametofitos, tanto antes como inmediatamente después de la reproducción (Hennig *et al.*, 2004, en prensa). Estos datos recientes apoyan por un lado que una proporción importante de los genes MADS tipo I son funcionales y, por otro, sugieren que el hecho de que sean importantes durante la reproducción podría explicar por qué en su evolución molecular la selección natural positiva parece haber jugado un papel más importante.

Armados con toda la información arriba mencionada, ahora podemos preguntarnos cuáles elementos constituyen a un gen MADS-box funcional. Aparentemente el requerimiento estrictamente mínimo es que exista una secuencia capaz de codificar un dominio MADS completo. Si bien nuestro conocimiento actual de las proteínas con dominio MADS sugiere que este dominio interviene en la unión con DNA y en la dimerización (Pellegrini *et al.*, 1995) no se ha demostrado aún que

este dominio intervenga en el control de la transactivación, por lo que probablemente la sola presencia del dominio MADS no sea suficiente para conferir a una proteína la función de activador. Además no podemos descartar que la función de las proteínas MADS también puede depender de que establezcan complejos con otras proteínas o que, en lugar de activadores, también funcionen como represores de la transcripción y en este caso podrían tal vez establecer complejos con otras proteínas.

Como hemos visto, algunos de genes tipo I parecen transcribirse a niveles muy bajos, al menos en los tejidos puestos a prueba por Parenicová y colaboradores (2003) y por Kofuji y colaboradores (2003) (pero ver Hennig *et al.*, en prensa), pero este dato no necesariamente implica que estas secuencias no sean funcionales ya que, por ejemplo, el gen *PHRES1* (antes llamado *AGL37*), el cual se ha demostrado que está asociado con el aborto de semillas en cierto fondo mutante (Kohler *et al.*, 2003), no dio, en el estudio de Kofuji y colaboradores, ninguna señal de expresión, tanto con el método de macroarreglos de mRNA, como con el de RT-PCR (Kofuji *et al.*, 2003). Por otro lado, desarrollos bioinformáticos recientes permitirán analizar las secuencias reguladoras de las secuencias en cuestión y así establecer si estas tienen sitios conservados con respecto a otros genes estudiados desde el punto de vista funcional. Este enfoque complementará los análisis de las secuencias codificantes para explorar la posible regulación espacio-temporal de éstas y así evaluar si las secuencias en cuestión son funcionales o no, antes de embarcarse en estudios funcionales más detallados.

Historia de los genes MADS

Alvarez-Buylla y colaboradores (2000a, ver apéndice 2) han planteado que los genes MADS-box sufrieron al menos una duplicación de su gen más ancestral antes de que ocurriera la separación de los linajes de plantas, hongos y animales, y que esta duplicación ancestral dio origen a dos linajes de genes distintos: los genes MADS-box tipo I (o *SRF-like*) y los tipo II (o *MEF2-like*). Cuando se planteó esta idea, el genoma de *Arabidopsis* aún no estaba completamente secuenciado y otros genomas de plantas estaban aún más lejos en el horizonte. Reconstrucciones filogenéticas más recientes, hechas después de que se completó la secuenciación del genoma de *Arabidopsis*, y que seguramente incluyen a prácticamente todas las secuencias MADS-box presuntamente funcionales, muestran con mayor claridad algunos de los patrones que se vislumbraban en el artículo de Alvarez-Buylla y colaboradores y permiten corregir ciertas interpretaciones.

El primer resultado notable del trabajo de Alvarez-Buylla y colaboradores es que las secuencias tipo *MEF2* de plantas, animales y hongos (incluyendo los genes *SMP1* y *RLM1* de *Saccharomyces cerevisiae*) forman un grupo

monofilético. La monofilia de este grupo está sustentada en primer lugar por varios aminoácidos conservados en el dominio MADS (ver apéndice 2). Este grupo ha sido denominado tipo II. Adicionalmente, los miembros de este grupo provenientes de plantas tienen una estructura llamada IKC que incluye un dominio conservado llamado caja K o dominio K. El dominio K fue definido originalmente como un dominio que “tiene una similitud baja pero significativa con una porción de las secuencias de keratina” (Ma, Yanofsky y Meyerowitz, 1991). Puesto que la región de la keratina que tiene similitud con el dominio K de proteínas MADS de plantas forma parte de la estructura con forma de *coiled coil* que constituye el dominio central en forma de bastón de la keratina, y como en muchos de los representantes de este dominio hay un espaciamiento regular de aminoácidos hidrofóbicos que son los responsables de conferir la estructura tridimensional, los criterios para identificar dominios K han sido: la similitud de secuencia con otros dominios K identificados tempranamente, el espaciamiento regular de aminoácidos hidrofóbicos y, derivado de este criterio, la predicción de la formación de *coiled coils*. En el estudio de Alvarez-Buylla *et al.* (2000a, ver apéndice 2) se usó el criterio de identificar un dominio K con base en la formación predicha de *coiled coils*, pero este criterio se revelaría posteriormente como demasiado riguroso. En efecto, si bien la monofilia de los genes tipo II está bien sustentada por sinapomorfías que se encuentran en el dominio MADS, así como por otras características como el número de exones de los genes de ese tipo, o la estructura MIKC de sus productos, la presencia de *coiled coils* en el dominio K no fue detectada por Alvarez-Buylla y colaboradores en todas las proteínas MADS del tipo II provenientes de plantas, por lo que se postuló que la aparición de esta característica estructural había ocurrido relativamente tarde en la evolución de la familia MADS, después de la separación del linaje de las plantas de los otros reinos eucariontes. Hay que señalar que la baja resolución del orden de ramificación de ese grupo de genes sugería que, en principio, no era descabellada la hipótesis de que los *coiled coils* del dominio K se distribuyeran entre las secuencias tipo II de manera tal que fuera posible formar con ellas un grupo monofilético que fuera un subgrupo de las secuencias tipo II. De hecho, si este fuera el caso, sería un ejemplo interesante de aparición gradual de una característica estructural con posible relevancia funcional a partir de una secuencia que no poseía esa característica. Sin embargo, estudios posteriores que contaron con información más completa acerca de las características de las secuencias MADS tipo II en plantas con un origen más antiguo, como el musgo *Physcomitrella patens* (Henschel *et al.*, 2002) y acerca de la resolución interna del grupo de genes tipo II (incluyendo el trabajo posterior de Martínez-Castilla y Alvarez-Buylla, 2003; ver apéndice 3) mostraron que la formación de estructuras de *coiled coil* es probablemente una característica lábil y que se podría obtener más información filogenética si se toma como criterio para identificar dominios K la similitud con secuencias identificadas previamente. De hecho, si se usa este criterio resulta que puede ser considerado una sinapomorfía adicional que define al

grupo monofilético de todas las secuencias MADS tipo II de *Arabidopsis* y tal vez de las plantas vasculares en general. En todo caso, esa exploración inicial sirvió para descubrir que algunas proteínas MADS tipo II de plantas probablemente no forman *coiled coils*, lo que puede tener relevancia para entender sus funciones.

Cuando el criterio para identificar el dominio K es la similitud de secuencia, se aclara el por qué el clado de los genes tipo II se extiende para incluir a todas las secuencias con estructura MIKC, incluyendo a *AGL30*, que en el estudio de Alvarez-Buylla y colaboradores del 2000 (Alvarez-Buylla *et al.*, 2000a) no había podido ser asignado sin ambigüedad a ninguno de los dos tipos de genes MADS. Entre los estudios donde se puede apreciar este agrupamiento, que implica una definición más amplia del clado tipo II, está el de Martínez-Castilla y Alvarez-Buylla (2003, ver apéndice 3), el de Parenicová *et al.* (2003) y los de Nam *et al.* (2003 y 2004).

La monofilia del grupo aún mayor de genes tipo II que incluye tanto a secuencias de plantas como de animales y hongos todavía es objeto de debate. Por ejemplo, Kofuji y colaboradores (2003) afirman que los genes tipo MEF2 de animales y hongos no constituyen el grupo hermano de los genes MIKC de plantas debido a que, en sus reconstrucciones, los genes MEF2 y los MIKC no forman un grupo monofilético (pero el soporte de las ramas que separarían a estos grupos es menor del 50% de *bootstrap*). Sin embargo, los genes MEF2 habían sido agrupados junto con los genes MIKC de plantas por Alvarez-Buylla y colaboradores (2000a) bajo el criterio de sinapomorfias existentes en la región MADS, formando así el grupo tipo II. Esta contradicción se explica porque el árbol al que recurren Kofuji y colaboradores (2003) para hacer esta afirmación no muestra un buen respaldo estadístico en las ramas que estarían separando a los genes tipo II de plantas (los MIKC) y a los MEF2 de animales y hongos. Kofuji y colaboradores (2003) señalan que las sinapomorfias usadas por Alvarez-Buylla y colaboradores (2000a) para unir en un clado a los genes tipo II de plantas y a los genes tipo MEF2 están moderadamente conservadas entre los genes MEF2 y el subgrupo de genes tipo II de plantas que Kofuji y colaboradores llaman MIKC^C (“MIKC clásico”) pero no en el subgrupo que han llamado MIKC*, sin embargo no está claro cuál es la interpretación que esos autores pretenden darle a ese dato, puesto que por otro lado defienden la monofilia del agrupamiento de los genes MIKC^C y los MIKC*.

Otros autores aceptan la monofilia del agrupamiento de los genes MEF2 de animales y hongos con los genes MIKC de plantas suplementados con los genes del grupo de *AGL30* (es decir, los genes MIKC^C suplementados con los MIKC*) para formar el grupo de los genes tipo II. En particular, el estudio de Nei y colaboradores (Nam *et al.*, 2003), que comparó la caja MADS de un set de genes más diverso que el analizado por Alvarez-Buylla *et al.* (2000a), encuentra

fundamentalmente que los genes MADS-box tipo II, que incluyen a los genes MIKC^C y MIKC* de un muestreo diverso de plantas y los genes MEF2 de hongos y animales, forman un grupo monofilético con buen soporte de *bootstrap*.

La monofilia de los genes tipo I también es controversial. De entrada hay que señalar que el artículo donde se planteó originalmente la existencia de por lo menos una duplicación en los genes MADS-box anterior a la divergencia de los principales reinos eucariontes no se afirmaba definitivamente que los genes tipo I formaran un único linaje monofilético, aunque algunos de los resultados mostrados en ese artículo (Alvarez-Buylla *et al.*, 2000a; ver apéndice 2) sugerían que ese bien podría ser el caso. Los artículos recientes de Kofuji *et al.* (2003) y de Becker y Theissen (2003) hacen señalamientos un tanto ambiguos acerca del carácter monofilético del grupo de genes tipo I. Por ejemplo, el grupo de Hasebe señala que el grupo de genes tipo I probablemente no sea monofilético, pero su afirmación se basa fundamentalmente en el hecho de que los genes del grupo de *AGL30* (que en el artículo de Alvarez-Buylla y colaboradores no había sido asignado claramente a alguno de los dos grupos aunque los árboles filogenéticos lo colocaban entre los tipo I) pertenecería más bien al tipo II (como de hecho lo ubican otras filogenias más recientes [Martínez-Castilla y Alvarez-Buylla, 2003; Parenicová *et al.*, 2003; y menos claramente, De Bodt *et al.*, 2003 y Nam *et al.*, 2004] y a que en sus reconstrucciones los genes tipo MEF2 parecen agruparse, aunque con bajo soporte estadístico, con los genes de plantas que no son del tipo II.

El problema de la monofilia de los genes tipo I es difícil porque la región MADS, que por lo menos a primera vista es el único grupo de caracteres compartido por todas las secuencias en cuestión, parece contener poca información que permita agrupar sin ambigüedad a los genes que no son tipo II. El único estudio en el que se reporta la cantidad de sinapomorfias para el tipo I presentes en la caja MADS es el de Alvarez-Buylla y colaboradores (2000a) y sólo involucraba una fracción de las secuencias MADS presuntamente tipo I disponibles hoy en día, por lo que valdría la pena repetir ese estudio con un muestreo más amplio, tanto de especies como de secuencias. A la espera de ese trabajo hay que notar que todavía hay información cladística a la que se le ha prestado poca atención y que es relevante para el problema de la monofilia de los genes tipo I. En primer lugar hay que recordar que los genes que en el trabajo de Alvarez-Buylla y colaboradores (2000a) no se pudieron asignar ni al tipo I ni al II quedaron en esa posición ambigua no porque no tuvieran las sinapomorfias propias de uno y otro grupo sino porque tenían al menos algunas de *ambos* grupos. Es notorio que en este grupo de asignación incierta se encontraban secuencias que estudios posteriores han colocado como grupo hermano del resto de los genes tipo II de plantas (como *AGL30*); secuencias que después han sido asignadas al subgrupo M α de los genes tipo I, al cual tres estudios diferentes (Martínez-Castilla y Alvarez-Buylla, 2003; Parenicová *et al.*, 2003; Nam *et al.*, 2004) ubican como el más cercano a los genes tipo II (MIKC^C+MIKC*+MEF2) de todos los subgrupos de genes tipo I, aunque no

necesariamente con alto soporte; y una secuencia (*AGL33*) que dos trabajos (Parenicová *et al.*, 2003 y Nam *et al.*, 2004) colocan en una rama que es basal tanto al subgrupo M α como a los subgrupos M β y M γ . (Nuestros análisis de detección de eventos de recombinación o conversión génica nunca captaron señal sugerente de estos fenómenos en relación con la secuencia *AGL33*, Es decir, las observaciones preliminares sugieren que los miembros del grupo de asignación incierta realmente tienen características intermedias. También es notable que en el estudio de Alvarez-Buylla y colaboradores (2000a) están ausentes genes de lo que más tarde se denominaría el subgrupo M β , por lo que sería interesante ver si comparten apomorfias con miembros del subgrupo M γ o --menos probablemente-- del grupo M α . Si se confirmara la monofilia del agrupamiento de los genes M β y M γ (que, por los resultados de Alvarez-Buylla *et al.* [2000a] cabría esperar que también formaran un clado con los genes tipo SRF de animales y hongos) el escenario más parsimonioso para explicar la distribución de los genes tipo I en hongos, animales y plantas postularía al menos un origen único para los genes tipo SRF de animales y hongos y los genes tipo M β y M γ de plantas y entonces el problema de la monofilia de ambos tipos de genes MADS-box se reduciría a dilucidar el origen de los genes M α de plantas o bien a cambiar su afiliación, que en todo caso era incierta desde la propuesta original de dos grandes linajes.

Existen al menos dos estrategias con las que es posible abordar el problema de la monofilia de los genes tipo I. La primera es, por supuesto, repetir el análisis que se reporta en el trabajo de Alvarez-Buylla y colaboradores (2000a; ver apéndice 2) ahora con el conjunto aparentemente completo de genes MADS-box de *Arabidopsis* y un muestreo más diverso de genes provenientes de otras plantas así como de hongos y animales para interrogarse de nuevo sobre cuál es el punto óptimo para ubicar la raíz de la familia MADS-box. El método utilizado en el apéndice 2 está basado en parsimonia pues trata de minimizar los eventos postulados de duplicación y *gene sorting*, pero desde la aparición de ese artículo se han desarrollado métodos que abordan el problema de la ubicación óptima de la raíz de un árbol filogenético en un marco conceptual bayesiano (Huelsenbeck, Bollback y Levine, 2002) y sería interesante explorar el desempeño relativo de ambas aproximaciones en este problema.

La otra estrategia para estudiar la monofilia de los genes tipo I es hacer pruebas estadísticas de monofilia. Estas pruebas nos ayudan a determinar si las diferencias en los métricos de optimización (por ejemplo, número de pasos o verosimilitud) de hipótesis filogenéticas alternativas en las que un grupo en cuestión es o no es monofilético, son significativas. De esta forma podríamos preguntarnos, si es verdad que la evolución ocurrió bajo determinado modelo de sustituciones (por ejemplo, el modelo GTR + SS, ver apéndice 3) con determinados valores de parámetros y a lo largo de un árbol que se está sometiendo a prueba (por ejemplo, nuestro árbol de la figura 1 del apéndice 3 contra el de alguno de los

otros autores mencionados) entonces, ¿cuál sería la diferencia en número de pasos entre el árbol más parsimonioso obligado a tener a los genes $M\alpha$ formando un grupo monofilético con los demás genes tipo I (o, alternativamente, con los genes tipo II, o bien formando su propio grupo) y el árbol más parsimonioso que no tuviera restricciones? Supongamos que al manipular el árbol para ver el efecto de forzar la monofilia obtenemos una diferencia en número de pasos de, digamos, nueve pasos entre el árbol forzado y el árbol sin forzar, entonces, ¿la diferencia simulada sería similar a la observada de nueve pasos o es nueve un valor inesperado? Específicamente, ¿es nueve un valor que esperaríamos observar menos del 0.05% de las veces? Si es así, podríamos rechazar la hipótesis de que los genes $M\alpha$ forman un clado monofilético con los demás genes tipo I.

El procedimiento que acabamos de mencionar es un ejemplo del uso del *bootstrapping* paramétrico (Huelsenbeck *et al.*, 1995; Swofford *et al.*, 1996; Goldman *et al.*, 2000). Esta prueba es casi idéntica al test de monofilia propuesto por Huelsenbeck *et al.* (1996) excepto que en esa otra prueba se usa la diferencia en las verosimilitudes como estadístico de prueba, en lugar de las longitudes de los árboles. El ejemplo se refiere a la medida cladística con el fin de hacerlo más intuitivo, pero ambos métodos se pueden llevar a cabo en un sistema de análisis de evolución de caracteres como Mesquite (Maddison y Maddison, 2004) o MacClade (Maddison y Maddison, 1992).

Existe otro aspecto importante de la estructura filogenética de los genes tipo I: algunos de los análisis recientes muestran una estructura resuelta de los genes tipo I aunque sea con bajos valores de soporte de *bootstrap* (Parenicová *et al.*, 2003; Martínez-Castilla y Alvarez-Buylla, 2003; Nam *et al.*, 2004), mientras que otros prefieren mostrar árboles colapsados cuando el soporte es bajo (p. ej., De Bodt *et al.*, 2003). Ahora bien, entre los trabajos donde sí se muestra la estructura interna de los genes tipo I, los genes $M\beta$ y $M\gamma$ son grupos hermanos que forman un grupo monofilético. De hecho, en uno de los estudios este resultado aparece con buen soporte de *bootstrap* (Parenicová *et al.*, 2003) y en otro (Martínez-Castilla y Alvarez-Buylla, 2003) con muy buen soporte por probabilidad posterior bayesiana (aunque esta medida ha sido criticada como estadístico de credibilidad de clados [Suzuki, Glazko y Nei, 2002; pero contrástese con Douady *et al.*, 2003]). En los tres estudios se utilizaron métodos diferentes (máxima parsimonia y *neighbor joining* en el estudio de Nam *et al.*, reconstrucción bayesiana por Monte Carlo de cadena de Markov en los estudios de Parenicová *et al.* y Martínez-Castilla y Alvarez-Buylla) y organizaciones de los datos similares pero no idénticas (alineaciones de aminoácidos en los estudios de Nam *et al.* y Parenicová *et al.*, alineaciones de nucleótidos analizados bajo un modelo de sustitución de bases dependiente de la posición intra-codon, no de bases que pueden variar independientemente de su posición, en el caso de Martínez-Castilla y Alvarez-Buylla. Además, en este último estudio se utilizó no sólo las partes más conservadas de las secuencias

sino también las de alineación más ambigua, una estrategia que parece temeraria pero que, como estamos viendo, arrojó resultados similares a los de estrategias más conservadoras aunque con mejor soporte estadístico (ver descripción suplementaria de métodos del apéndice 3). En resumen, la imagen que se va formando de la estructura interna de los genes tipo I a través de varias líneas de evidencia es que los genes que Martínez-Castilla y Alvarez-Buylla llaman tipo *AGL26* (grupo casi idéntico a los genes *Mβ* de Parenicová *et al.*) y los tipo *PHERES1* (idénticos a los *Mγ* de Parenicová *et al.*) forman un clado en el que ambos grupos son hermanos y que probablemente también incluya a los genes tipo SRF de animales y hongos, mientras que los genes tipo *AGL23* (idénticos a los *Mα*) constituyen un grupo externo al grupo anterior, cuya afiliación a los genes tipo I o tipo II, como ya vimos, puede resolverse con relativa facilidad. La estructura interna de los grupos es casi idéntica en el estudio de Parenicová *et al.* y en el de Martínez-Castilla y Alvarez-Buylla, con la excepción de que en el primer caso los genes *AGL82* y *AGL47* son asignados con bajo soporte al grupo *Mβ*, mientras que en el segundo caso ambos genes forman un grupo hermano tanto a los genes tipo *AGL23* como a los tipo *PHERES*.

Por otro lado, los distintos análisis filogenéticos muestran algunos grupos monofiléticos bien definidos dentro de los tipo II, mientras que el orden de ramificación de otros no está completamente resuelto. Por ejemplo, muchos de los estudios recientes (Parenicová *et al.* 2003, Nam *et al.* 2003, Martínez-Castilla y Alvarez-Buylla 2003, Becker y Theissen 2003, Kofuji *et al.* 2003) ubican a los genes tipo *SEPALLATA* (*SEP1*, *SEP2* y *SEP3*) en un grupo monofilético junto con los genes tipo *API/SQUA* (*API*, *CAL*, *FUL* y *AGL79*) y los genes *AGL3*, *AGL6* y *AGL13*. De hecho, los análisis que muestran mejor soporte estadístico para la estructura interna de este grupo (Becker y Theissen, Martínez-Castilla y Alvarez-Buylla) muestran a *AGL3*, *AGL6* y *AGL13* como grupo hermano de los genes *SEPALLATA* y a los genes tipo *API* como grupo hermano de los otros dos clados. En el artículo de Parenicová *et al.* estos tres grupos tienen una estructura interna diferente pero con bajo soporte, lo cual se puede deber a la inclusión del gen *AGL12*, que casi todos los otros análisis ubican como hermano de los genes tipo *AG/PLENA*. Esto último es interesante pues *AGL12* se expresa primordialmente en raíz, aunque también se ve a bajos niveles en ciertos tejidos de la flor (Tapia-López *et al.* com. personal) mientras que los otros genes son específicos de flor.

Dos estudios (Becker y Theissen, 2003 y Martínez-Castilla y Alvarez-Buylla, 2003) ubican a los genes del grupo de la función C o *AG/PLENA*, incluyendo a *AGL12*, como grupo hermano del de todos los genes que se mencionaron en el párrafo anterior, mientras que en otros dos estudios (Nam *et al.* y Parenicová *et al.*) el grupo hermano de esos genes es el grupo constituido por los genes *SOC1*, *AGL71*, *AGL72*, *AGL14*, *AGL19* y *AGL42*. Este último grupo también se rescata como clado en los otros dos estudios pero asociado a otro grupo. Otros grupos monofiléticos que también se rescatan en

todos los estudios son el grupo de los genes de la función B (*AP3* y *PISTILLATA*), que en todos los estudios forma un clado con el grupo llamado B-sister (*AGL63* y *TT16*), el grupo de los genes tipo *FLF* (*FLF*, *AGL70*, *MAF1*, *MAF2*, *MAF4* y *MAF5*), el grupo de los genes *AGL16*, *AGL17*, *AGL21*, y *ANR1*, el grupo de *AGL24* y *SVP* y el grupo de *AGL15* y *AGL18*. Todos estos grupos aparecen con buen soporte en la mayoría de los estudios pero sus afiliaciones relativas varían en todos los artículos y en la mayoría de éstos el soporte estadístico de las agrupaciones profundas entre estos grupos es bajo. Curiosamente, existe mucha mayor concordancia entre los resultados de los diferentes estudios sobre la resolución interna de los genes tipo I que sobre los genes tipo II, a pesar de que en estos últimos existen más caracteres no ambiguos para resolver las relaciones entre los diferentes sub-grupos.

Hay una cuestión adicional relativa a la composición del grupo de los genes tipo II. En los artículos de Vergara-Silva y colaboradores (2000; ver apéndice 1) y Alvarez-Buylla y colaboradores (2000a; ver apéndice 2) se plantea que en los genes MADS-box tipo II, el dominio K (definido como una región capaz de formar una estructura de *coiled coil*) evolucionó después de la duplicación que llevó a la formación de los linajes de genes de animales tipo MEF2 y tipo SRF. También se plantea que, con base en la filogenia de la figura 3b del apéndice 2, no se puede distinguir si esta conformación de la región correspondiente al dominio K apareció en el linaje de las plantas después de que divergiera del de hongos y animales o si estaba presente en el gen tipo II ancestral y después se perdió en los linajes de hongos y animales y también en algunos genes tipo II de plantas. El grado de resolución de la filogenia y la amplitud del muestreo de genes MADS-box de plantas disponibles cuando se publicó el apéndice 2 sugerían que una hipótesis plausible para la evolución del dominio K y por lo tanto de la estructura MIKC de los genes tipo II de plantas era que el dominio K había aparecido en los genes tipo II de plantas después de la separación del linaje de las plantas del de hongos y animales, e incluso después de que los genes tipo II de plantas ya habían sufrido algunos eventos de duplicación y diversificación. Como ya se ha mencionado, esta visión se deriva en parte de una definición del dominio K basado en la presencia predicha de la estructura *coiled coil*, pero otros autores definen al dominio K (y, por implicación, a la estructura MIKC) con base en la similitud de secuencia de este dominio con el de una porción de las keratinas (Ma, Yanofsky y Meyerowitz, 1991; Henschel *et al.*, 2002), independientemente de la conformación tridimensional que adquiriera esta región. Por otro lado, la visión de un origen único y relativamente tardío del *coiled coil* en el dominio K depende también de que ciertos genes MADS-box (como *AGL12* o los genes tipo *FLF*) tengan una ubicación basal en la filogenia de los genes tipo II de plantas, lo cual no resulta claro por ahora. En las filogenias recientes más completas (incluyendo las del apéndice 3) se ubican a estos genes en posiciones derivadas

por lo que hay que concluir que el dominio K, o al menos su conformación como *coiled coil*, sufrió repetidas pérdidas y reparaciones durante la evolución de los genes tipo II de plantas.

Ahora bien, la estructura MIKC definida con base en la similitud de secuencia relata una historia más sencilla de la evolución de los genes tipo II de plantas. En las reconstrucciones que se muestran en el apéndice 3 los genes tipo II incluyen a los genes *AGL30*, *AGL33*, *AGL65*, *AGL66*, *AGL67*, *AGL94* y *AGL104* que tienen una estructura MIKC divergente pero reconocible a nivel de similitud de secuencia, mientras que en otros estudios recientes (p. ej. Parenicová *et al.*, 2003) estos mismos genes (llamados tipo Mδ) forman un grupo hermano al de los genes con estructura MIKC más clásica, si se define a ésta por similitud de secuencia. Recientemente han sido clonados en el musgo *Physcomitrella patens* genes MADS-box con una estructura MIKC residual similar a la de algunos de los genes de *Arabidopsis* del clado que Parenicová *et al.* (2002) han llamado Mδ (Henschel *et al.*, 2002). Además, estos genes muestran una estructura intron-exon muy similar a la de los genes MIKC clásicos, con más de cinco exones, mientras que las secuencias de todos los subgrupos de los genes tipo I están estructuradas en un máximo de tres exones. Tomados en conjunto, estos datos indican que la estructura MIKC es una estructura característica de los genes MADS-box tipo II de plantas y que probablemente fue adquirida por estos genes muy pronto después de la separación del linaje de las plantas del de los hongos y los animales.

¿Es la filogenia de los genes MADS-box de *Arabidopsis* que se muestra en el apéndice 3 una buena representación de la evolución de los genes MADS-box de plantas en general? La secuenciación de genomas y transcriptomas completos o parciales de otros representantes de diferentes ramas de la filogenia de las plantas permitirán contar con respuestas más certeras a esta pregunta. Por ejemplo, el equipo de Lucia Colombo (Parenicová *et al.*, 2002) encontró que casi todos los subgrupos mayores de genes MADS-box de *Arabidopsis* tienen homólogos en el genoma de arroz con los que forman grupos monofiléticos. La excepción es el grupo Mβ para el que no encontraron genes correspondientes en el genoma de arroz pero la estructura de los árboles que se reporta en ese artículo implica que el grupo Mβ ya existía antes de la divergencia entre las eudicotiledóneas y las monocotiledóneas.

En contraste con esa observación, Parenicová *et al.* también reportan que en el interior del clado de los genes MIKC existen algunos subclados privativos de arroz y otros de *Arabidopsis*. Esto es similar a resultados preliminares de Martínez-Castilla y Alvarez-Buylla, no mostrados aquí, en los que también se puede ver que la expansión de algunas de las subfamilias de genes MADS-box son específicas de linajes particulares de especies de plantas. Para explorar el posible papel de los distintos clados de genes en los distintos linajes de plantas será interesante analizar, por ejemplo, si los patrones de selección natural difieren entre los grupos de genes compartidos y los no compartidos entre los linajes de organismos.

Fuerzas Evolutivas en Acción Durante la Historia Molecular de los MADS

La teoría sugiere por lo menos cuatro posibles destinos para los genes que resultan de una duplicación: el primero es que ambos miembros del par duplicado mantengan su función original y sean completamente redundantes entre sí; la segunda posibilidad es que la relajación de la presión de selección permita que haya divergencia entre los duplicados que podría resultar en la adquisición de una nueva función (neofuncionalización); la tercera posibilidad es que un miembro del par duplicado sea silenciado y finalmente degenera como pseudogen (pseudogenización). Recientemente se ha propuesto una cuarta posibilidad que consiste en que el patrón espacio-temporal de expresión que tenía el gen antes de ser duplicado se reparta entre los dos duplicados; entonces cada gen queda con un subconjunto de las funciones iniciales. A este proceso se le ha llamado subfuncionalización (Lynch y Force, 2000).

Ahora bien, la subfuncionalización ha sido presentada como un proceso que se lleva a cabo bajo condiciones de neutralidad selectiva. La neofuncionalización, por otro lado, puede ocurrir tanto bajo condiciones de neutralidad como por efecto de la selección natural positiva. Por ejemplo, bajo el supuesto de neutralidad, las restricciones selectivas pueden relajarse después de una duplicación, lo que conduce a una elevación en la tasa de sustituciones neutrales, y algunas de estas sustituciones eventualmente llevan a nuevas funciones, cuando el ambiente o el contexto genético cambian (el “efecto Dykhuizen-Hartl” [Dykhuizen y Hartl, 1980; Kimura, 1983; Zhang, Rosenberg y Nei, 1998]). En cambio, la selección natural positiva siempre está relacionada con la adquisición de nuevas funciones y si observáramos evidencia de que ocurrió selección positiva podríamos postular que por lo menos uno de los genes duplicados adquirió nuevas funciones.

Para el caso de la familia de genes MADS, los análisis que se detallan en el apéndice 3 mostraron evidencias de que la selección natural positiva jugó un papel en la evolución de por lo menos algunos de los genes de esta familia en *Arabidopsis* después de las duplicaciones que les dieron origen. Para disminuir la posibilidad de cometer errores tipo I, se evitaron las zonas de las secuencias de los genes MADS-box que son más divergentes y cuya alineación con otras secuencias es ambigua. Así, para los genes tipo II sólo se analizó las regiones MADS, K y una parte de la I, mientras que para los genes tipo I sólo se analizó la región MADS. Con este criterio conservador los análisis de sitio por sitio detectaron clados en los genes tipo I con huellas de haber evolucionado bajo selección natural positiva, pero esto contrasta con los genes de tipo II, para los que no se detectó señal de selección positiva.

En cambio, usando los análisis llamados de rama por sitio, que tienen un mayor grado de sensibilidad que los análisis por sitio, se pudo evidenciar un papel para la selección natural positiva en el origen de dos grupos de genes MADS-box tipo II y en la divergencia del único gen tipo I cuya función se ha caracterizado (ver apéndice 3). Nuevamente, con el fin de evitar errores de tipo I, se adoptó la estrategia conservadora de aplicar los análisis de rama por sitio únicamente a tres ramas de la filogenia de los genes MADS-box. Estas ramas se eligieron porque representaban casos en los cuales parecía más intuitivo postular divergencia funcional después de una duplicación. Estos son los grupos de genes FLC y SVP que intervienen en la ontogenia de los órganos reproductivos de *Arabidopsis* mediante un control del tiempo de aparición de estos órganos, a diferencia de otros genes tipo II cuya función tiene más que ver con la especificación de órganos y tejidos. Además, el control temporal fino de la floración puede ser una característica de la historia de vida de *Arabidopsis* que en principio afectaría sensiblemente su adecuación y por lo tanto podría ser sujeta de evolución bajo el control de la selección natural. Por otro lado, el gen *PHERESI* (*AGL37*) es el único gen tipo I de plantas que se ha caracterizado funcionalmente y por tanto resultaba atractivo explorar si un gen que presuntamente no tiene redundancia funcional con sus parálogos cercanos había alcanzado esa condición por efecto de la selección natural positiva.

Puesto que los análisis que llevamos a cabo fueron bastante conservadores, la ausencia de evidencia de selección positiva en muchas de las pruebas de los análisis por sitio deben de interpretarse cuidadosamente. En particular, no pueden interpretarse como evidencia de ausencia de selección positiva, pero, de todas formas, a la luz del debate seleccionismo-neutralismo, ésta es una posibilidad intrigante. El equipo de Masatoshi Nei (Nam *et al.*, 2004) usando métodos menos sensibles que la detección de selección por sitio y de rama por sitio (ellos usaron el número de sustituciones sinónimas y no sinónimas promediadas a lo largo de toda la secuencia codificante de los genes y recurrieron a evidencia circunstancial de mayores tasas de nacimiento y muerte entre los genes tipo I que entre los tipo II) concluye que los genes tipo I están bajo un régimen de selección purificadora ($dN/dS < 1$) más débil que los tipo II. La selección purificadora se puede deber a restricciones funcionales y estos resultados sugieren que los genes tipo I pueden ser menos importantes funcionalmente, o pueden estar menos sujetos a restricciones funcionales, que los genes tipo II. Pero esos autores también son cautelosos y agregan que eso no implica que los genes tipo I no tengan funciones importantes. De hecho, nuestros resultados sugieren que pueden ser muy importantes y que la selección positiva puede haber jugado un papel importante en su diversificación funcional.

El carácter modular de las proteínas con dominio MADS ya había sido puesto en evidencia en estudios anteriores (p. ej., Pellegrini, Tan y Richmond, 1995; Riechmann, Wang y Meyerowitz, 1996; Krizek y Meyerowitz, 1996; Krizek,

Riechmann y Meyerowitz, 1999; Santelli y Richmond, 2000; Huang *et al.*, 2000). Nuestros resultados sobre la acción puntual de la selección natural positiva enriquecen esa perspectiva al mostrar que aún dentro de los módulos parcialmente caracterizados (p. ej. las regiones MADS, I, K y COOH) se encuentran subregiones o incluso aminoácidos específicos que al mutarse pueden conferir nuevas funciones a las proteínas, probablemente mediante la posibilidad de interactuar con un conjunto nuevo de moléculas. Recientemente se publicó un estudio (Vandenbussche *et al.*, 2003) en el que también se reportan observaciones que implican evolución por neofuncionalización en los genes MADS-box tipo II de plantas. Concretamente, los autores de ese artículo reportan haber identificado mutaciones de recorrimiento del marco de lectura en los motivos COOH-terminales de genes de las subfamilias AP3/DEF, AP1/SQUA y SHP. Esta región no fue analizada en el estudio del apéndice 3 por ser demasiado divergente, y en principio no contamos con un marco conceptual que nos permita determinar si ese tipo de mutaciones fue mantenido por la acción de la selección natural. Pero es probable que la región COOH que codifica para secuencias de aminoácidos con funciones distintas a las del resto de la secuencia codificante de los genes MADS-box pudo haber estado sujeta a patrones de selección contrastantes. De cualquier manera, los datos obtenidos hasta ahora sugieren que las proteínas con dominio MADS parecen estar evolucionando por varias rutas distintas.

Como se mencionó, el mantenimiento de genes duplicados por subfuncionalización ha sido típicamente asociado a situaciones de neutralidad selectiva por lo que nuestros datos de selección positiva no parecen apoyar un *escenario* de subfuncionalización en los genes MADS-box, pero tampoco permiten descartarlo. Además los modelos originales de subfuncionalización se han formulado más en términos de una partición del dominio de expresión del gen original entre los genes duplicados que en términos de otros aspectos del funcionamiento de un gen o sus productos, pero los dominios de expresión de los genes MADS-box no fueron analizados detalladamente en el apéndice 3. Sin embargo, los estudios de Parenicová *et al.* (2003) y de Kofuji *et al.* (2003) incluyen información sobre los tipos de órganos en los que se expresan prácticamente todos los genes MADS-box de *Arabidopsis*. Esta información, junto con las reconstrucciones filogenéticas de esta familia de genes, puede ser aprovechada para explorar la contribución relativa de los procesos de subfuncionalización en la evolución de los genes MADS-box. Concretamente, se pueden mapear sobre las filogenias los órganos o tejidos, o en general, los dominios espacio-temporales en los que se expresa cada uno de los genes MADS-box en *Arabidopsis* para, con base en criterios de máxima verosimilitud o máxima parsimonia, hipotetizar sobre cuáles eran los dominios de expresión de los genes ancestrales. Si esos dominios de expresión reconstruidos son iguales a la unión de los dominios de los genes actuales, la idea de mantenimiento de los duplicados por subfuncionalización se vería reforzada, de otra forma se vería falseada, al menos en el caso de los genes MADS-box de *Arabidopsis*.

La evidencia acumulada hasta ahora para los genes MADS-box sugiere que estos genes jugaron un papel importante en la evolución del linaje de plantas al que pertenece *Arabidopsis*, aunque aún no contamos con información suficiente para explorar la manera en que estos genes afectaron la evolución de esta y otras plantas. Es difícil pensar en que se mantengan tantos miembros de una familia multigénica sin que estos hayan cumplido papeles importantes en la evolución del linaje en el que se encuentran. La detección de evidencia de que al menos algunos miembros de la familia MADS-box de *Arabidopsis* evolucionaron por selección natural positiva después de las duplicaciones que los originaron implica necesariamente que los nuevos genes tuvieron una incidencia sobre la adecuación de las plantas que los tenían. Sin embargo, también es factible que algunos miembros de esta familia multigénica hayan evolucionado bajo procesos neutrales o incluso de que la divergencia que podemos observar entre algunas secuencias no involucre cambios en las capacidades funcionales de los productos de los genes y que, por lo tanto, se trate de casos, un tanto atípicos, de evolución por selección natural negativa. Todo esto apunta hacia varios desarrollos que se podrán llevar a cabo en el futuro inmediato para abordar inquietudes surgidas del estudio de los genes MADS-box:

- El desarrollo de técnicas altamente sensibles de detección de selección positiva sobre aminoácidos específicos. Las técnicas disponibles sólo se pueden usar si se tiene una hipótesis *a priori*, y sería útil contar con técnicas estadísticas más exploratorias y menos dependientes de pruebas de hipótesis, y con las que también se pueda abordar con alta sensibilidad el papel de la selección natural a nivel de sitios individuales de las proteínas pero sin perder la posibilidad de explorarlo a lo largo de todo un árbol filogenético. La técnica de reducción de ruido estadístico desarrollada recientemente por Clegg y colaboradores (Jia, Clegg y Jiang, 2003) indica un posible camino para el desarrollo de este tipo de métodos.
- Los estudios presentados aquí apoyan fuertemente la idea de que la evolución molecular de las secuencias codificantes de los genes de factores transcripcionales ha sido importante para la evolución fenotípica, pero también es muy probable que la evolución de secuencias reguladoras no codificantes haya jugado papeles importantes en la evolución de las plantas. Hasta ahora, el único marco conceptual que permite dilucidar la acción de la selección natural sobre regiones no codificantes del genoma es la genética de poblaciones (especialmente con la prueba Hudson-Kreitman-Aguadé [Hudson *et al.*, 1987; ver también Kreitman y Hudson, 1991, Aguadé *et al.*, 1992; Begun y Aquadro, 1993; Gaut y Clegg, 1993]). Pero probablemente los próximos desarrollos bioinformáticos revelen un código, unas reglas gramaticales, subyacentes a la función de

estas secuencias y con ello se podrán desarrollar nuevos métodos para evaluar el papel de la selección natural en su evolución.

- Ya se mencionó que la probabilidad posterior de los clados de una filogenia, inferida a partir del muestreo de Monte Carlo de la superficie de verosimilitud, ha sido criticada como medida de confiabilidad estadística de las filogenias (Suzuki *et al.*, 2002; pero véase Douady *et al.*, 2003 y Alfaro *et al.*, 2003). Sin embargo, una de las conclusiones de este estudio es que la fuerza de la presión selectiva a lo largo de todos los sitios de aminoácidos de una proteína puede ser notablemente heterogénea, lo que se traduce en tasas de sustitución divergentes para diferentes zonas de una proteína o su gen codificante. Ahora bien, la medida de confiabilidad de una filogenia usada más habitualmente, el *bootstrap* no paramétrico, parte de la suposición de homogeneidad de tasas a lo largo de la secuencia analizada por lo que no parece una prueba adecuada para nuestros datos. (Esto se puede ver claramente si se considera que valores altos de *bootstrap* se obtienen si todos los conjuntos remuestreados (*bootstrap sets*) producen esencialmente la misma topología, mientras que los valores bajos de *bootstrap* ocurren cuando los diferentes conjuntos remuestreados producen diferentes topologías. Ahora bien, el que los conjuntos remuestreados den topologías similares o diferentes depende de que los diferentes caracteres, concretamente los sitios de una alineación, se comporten de manera similar, o dicho de otra forma, “cuenten la misma historia”, y una expresión de que los diferentes sitios cuentan la misma historia es que tengan tasas relativamente homogéneas de sustitución.) Además, la exploración Monte Carlo de la superficie de verosimilitud en un marco bayesiano ofrece el atractivo de resumir más información estadística sobre los árboles filogenéticos que el *bootstrap*, de manera que a la luz de nuestros datos parece que vale la pena continuar explorando las propiedades de este método estadístico.
- Los datos disponibles sugieren que tanto procesos neutrales como selectivos han jugado papeles importantes en la evolución molecular de los genes MADS-box en plantas. Es interesante observar que los modelos más difundidos de la subfuncionalización, que se ha planteado como la explicación más adecuada para la existencia de familias multigénicas grandes, dependen de la evolución neutral, mientras que nuestros datos revelan evolución adaptativa en por lo menos algunos eventos de duplicación.
- Los genes MADS-box de plantas mejor estudiados son los que intervienen en la regulación de la morfogénesis floral, pero en este proceso intervienen muchos más genes además de los MADS-box. Análisis de la propiedades dinámicas de la red de regulación de la morfogénesis de la flor de *Arabidopsis* (p. ej. Mendoza y

Alvarez-Buylla, 1998; Mendoza, Thieffry y Alvarez-Buylla, 1999) indican que esta red es robusta en el sentido de poder acomodar mutaciones puntuales en sus elementos sin salirse de sus programas implícitos de patrones de activación de genes. Esto parece contradecir la sugerencia de que los genes MADS-box que forman parte de esta red evolucionan bajo selección positiva puesto que, si es cierto que la red de interacciones entre genes es sumamente robusta, (en el sentido de que se puede bloquear por completo la acción normal de algunos de los genes que constituyen nodos en la red de regulación del desarrollo de la flor y aún así obtener fenotipos normales o cercanos a los normales) las mutaciones en estos genes no se traducirían muy frecuentemente en cambios en el fenotipo de la flor y por lo tanto, los cambios en la adecuación –tanto de disminución como de aumento-- no serían tan inmediatos como tal vez lo serían si hubiera un mapeo más lineal, menos amortiguado por interacciones con otros genes, entre el genotipo y el fenotipo. Desde esta perspectiva, resultará muy interesante explorar esta aparente tensión entre una evolución del desarrollo conducida por restricciones estructurales y una conducida por fuerzas selectivas. Concretamente, sería muy relevante explorar los patrones de evolución molecular de genes que ocupan distintos sitios en la arquitectura de las redes y evaluar si las mutaciones de las redes son equivalentes a las sustituciones en los sitios que son blanco de la selección positiva de acuerdo a los análisis de evolución molecular y de detección del papel de la selección.

Las perspectivas que se sugieren aquí sobre la evolución de los genes MADS-box plantean que esta familia de genes podría ser un modelo interesante de evolución molecular para estudiar muchos de los procesos que moldean la diversidad de las formas vivientes a muchos niveles de organización. Si sabemos aprovechar las posibilidades de aprendizaje que nos brindan estos genes quizá seamos tan exitosos como ellos.

V. REFERENCIAS BIBLIOGRÁFICAS CITADAS

- Alfaro, M. E., Zoller, S. y Lutzoni, F.** 2003. Bayes or Bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo and bootstrapping in assessing phylogenetic confidence. *Mol Biol. Evol.* 20: 255-266.
- Aguadé, M., Miyashita, M. y Langley, C. H.** 1992. Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. *Genetics* 132:755-770.
- Ahn, S. y Tanksley, S. D.** 1993. Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. USA* 90: 7980-7984.
- Akashi, H.** 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139: 1067-1076.
- Akashi, H.** 1999. Within- and between-species DNA sequence variation and the "footprint" of selection. *Gene* 238: 39-51.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. y Lipman, D. J.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
- Alvarez-Buylla, E. R., Pelaz, S., Liljegren, S. J., Gold, S. E., Burgeff, C., Ditta, G. S., Ribas de Pouplana, Ll., Martínez-Castilla, L. y Yanofsky, M. F.** 2000a. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proc. Natl. Acad. Sci. USA* 97: 5328-5333.
- Alvarez-Buylla, E. R., Liljegren, S. J., Pelaz, S., Gold, S. J., Burgeff, C. N. L., Ditta, G. S., Vergara-Silva, F. y Yanofsky, M. F.** 2000b. MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *Plant J.* 24: 457-466.
- Arabidopsis Genome Initiative.** 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 408: 796-815.
- Arnold, M. y Davidson, E.** 1997. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124: 1851-1864.
- Bailey, G. S., Poulter, R. T. M. y Stockwell, P. A.** 1978. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proc. Nat. Acad. Sci. USA* 75: 5575-5579.
- Becker, A. y Theissen, G.** 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phyl. Evol.* 29: 464-489.

- Begun, D. J. y Aquadro, C. F.** 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* 365: 548-550.
- Bennetzen, J. L., Chandler, V. L. y Schnable, P.** 2001. National Science Foundation-sponsored workshop report. Maize genome sequencing project. *Plant Physiol.* 127: 1572-1578.
- Britten, R. y Davidson, E.** 1969. Gene regulation for higher cells: a theory. *Science* 165: 349-357.
- Britten, R. y Davidson, E.** 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Quart. Rev. Biol.* 46: 111-133.
- Camin, J. H. y Sokal, R. R.** 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19: 311-326.
- Carroll, S. B.** 2000. Endless forms: the evolution of gene regulation and morphological diversity. *Cell.* 101: 577-580.
- Carroll, S. B., Grenier, J. K. y Weatherbee, S. D.** 2001. *From DNA to diversity: molecular genetics and the evolution of animal design.* Blackwell Science. Malden, Massachussets.
- Coen, E. S. y Meyerowitz, E. M.** 1991. The war of the whorls: genetic interactions controlling flower development. *Nature* 353: 31-37.
- Cooke, T. J., Poli, D. y Cohen, J. D.** 2003. Did auxin play a crucial role in the evolution of novel body plans during the Late Silurian- Early Devonian radiation of land plants? *In* Hemsley, A. y Poole, I. (eds.) *The evolution of plant physiology.* Academic Press.
- Crandall, K., Kelsey, C. R., Imamichi, H., Lane, H. C. y Salzman, N. P.** 1999. Parallel evolution of drug resistance in HIV: failure of non-synonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* 16: 372-382.
- Cunningham, C. W., Omland, K. E. y Oakley, T. H.** 1998. Reconstructing ancestral character states: A critical reappraisal. *Trends Ecol. Evol.* 13: 361-366.
- De Bodt, S., Raes, J., Florquin, K., Rombauts, S., Rouzé, P., Theissen, G. Y van de Peer, Y.** 2003. Genomewide structural annotation and evolutionary analysis of the type I MADS-box genes in plants. *J. Mol. Evol.* 56: 573-586.
- Doebly, J. y Lukens, L.** 1998. Transcriptional regulators and the evolution of plant form. *Plant Cell.* 10: 1075-1082.
- Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. y Douzery, E. J. P.** 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20: 248-254.
- Dykhuisen, D. y Hartl, D. L.** 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* 96: 801-817.
- Eddy, S. R.** 1998. Profile hidden Markov models. *Bioinformatics* 14: 755-763.
- Farris, J. S.** 1970. Methods for computing Wagner trees. *Syst. Zool.* 19: 83-92.

- Farris, J. S. 1977. Phylogenetic analysis under Dollo's law. *Syst. Zool.* 26: 77-88.
- Felsenstein, J. 1978. Cases in which parsimony compatibility methods will be positively misleading. *Syst. Zool.* 27: 401-410.
- Ferrández, C., Gu, Q., Martienssen, R. y Yanofsky, M. F. 2000. Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER. *Development* 127: 725-734.
- Fitch, W. M. 1971. Toward defining the course of evolution: Minimal change for a specific tree topology. *Syst. Zool.* 20: 406-416.
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-I. y Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531-1545.
- Gaut, B. y Clegg, M. T. 1993. Molecular evolution of the *Adh1* locus in the genus *Zea*. *Proc. Natl. Acad. Sci. USA* 90:5095-5099.
- Gehring, W. J., Affolter, M. y Bürglin, T. 1994. Homeodomain proteins. *Annu. Rev. Biochem.* 63: 487-526.
- Gilbert, W., deSouza, S. J. y Long, M. Y. 1997. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA.* 94: 7698-7703.
- Gillespie, J. H. 1991. *The causes of molecular evolution*. Oxford University Press. Oxford, Reino Unido.
- Gilks, W., Richardson, S. y Spiegelhalter, D. (eds.) 1996. *Markov chain Monte Carlo in practice*. Chapman and Hall. Londres.
- Goldman, N., Anderson, J. P. y Rodrigo, A. G. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst. Zool.* 49: 652-670.
- Goldman, N. y Yang, Z. 1994. A codon-based model of nucleotide substitution for protein coding DNA sequences. *Mol. Biol. Evol.* 11: 725-736.
- Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E. y Matsuda, G. 1979. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* 28: 132-168.
- Goff, S. A. *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L.ssp. *japonica*). *Science* 296: 92-100.
- Griffiths, A. J. F., Gelbart, W. M., Miller, J. H. y Lewontin, R. C. 1999. *Modern Genetic Analysis*. W. H. Freeman. Nueva York, Estados Unidos.
- Gu, Q., Ferrández, C., Yanofsky, M. F. y Martienssen, R. 1998. The FRUITFULL MADS-box genes mediates cell differentiation during *Arabidopsis* fruit development. *Development* 125: 1509-1517.

- Guttman, D. S. y Dykhuizen, D. E.** 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266: 1380-1383.
- Gyllensten, U. B., Sundvall, M. y Erlich, H. A.** 1991. Allelic diversity is generated by intraexon sequence exchange at the *DRB1* locus of primates. *Proc. Natl. Acad. Sci USA* 88: 3686-3690.
- Haldane, J. B. S.** 1933. The part played by recurrent mutation in evolution. *Am. Nat.* 67: 5-19.
- Hartigan, J. A.** 1973. Minimum mutation fits to a given tree. *Biometrics* 29: 53-65.
- Hastings, K. W.** 1970. MonteCarlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97-109.
- Hennig, L., Gruissen, W., Grossniklaus, U. y Köhler, C.** 2004. *En prensa*. Transcriptional programs of plant reproduction. *Plant Physiol.*
- Henschel, K., Kofuji, R., Hasebe, M., Saedler, H., Münster, T. y Theissen, G.** 2002. Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. *Mol. Biol. Evol.* 19: 801-814.
- Hilliker, A. J., Clark, S. H. y Chovnik, A.** 1991. The effect of DNA sequence polymorphisms on intragenic recombination in the *rosy* locus of *Drosophila melanogaster*. *Genetics* 129: 779-781.
- Hilliker, A. J., Harauz, G., Raume, A. G., Gray, M., Clark, S. H. y Chovnik, A.** 1994. Meiotic gene conversion track length distribution within the *rosy* locus of *Drosophila melanogaster*. *Genetics* 137: 1019-1026.
- Himmereich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. y Hermann, R.** 1996: Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24:4420-4449.
- Huang, K., Louis, J. M., Donaldson, L., Lim, F.-L., Sharrocks, A. D. y Clore, G. M.** 2000. Solution structure of the MEF2A-DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. *EMBO Journal* 29:2615-2628.
- Hudson, R. R.** 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* 23: 183-201.
- Hudson, R. R., Kreitman, M. y Aguadé, M.** 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153-159.
- Huelsenbeck, J. P. y Bolback, J. P.** 2001. Empirical and hierarchical bayesian estimation of ancestral states. *Syst. Biol.* 50: 351-366.
- Huelsenbeck, J. P., Bolback, J. P. y Levine, A. M.** 2002. Inferring the root of a phylogenetic tree. *Syst. Biol.* 51: 32-43.
- Huelsenbeck, J. P., Hillis, D. M. y Jones, R.** 1995: Parametric bootstrapping in molecular phylogenetics: applications and performance. In J. D. Ferraris y S. D. Palumbi, (eds.) *Molecular zoology: Advances, strategies and protocols. Symposium held during the annual meeting of the American Society of Zoologists, St. Louis, Missouri, USA, January 1995.* pp. 19-45. Wiley-Liss, Inc. Nueva York, Estados Unidos.

- Huelsenbeck, J.P., Hillis, D. M. y Nielsen, R.** 1996. A likelihood ratio test of monophyly. *Syst. Zool.* 45: 546-558.
- Huelsenbeck, J. P. y Ronquist, F.** 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754-755.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R. y Bolback, J. P.** 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*. 294: 2310-2314.
- Hughes, A. L.** 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R.. Soc. London Ser. B* 256: 119-124.
- Hughes, A. L., Friedman, R., Ekollu, V. y Rose, J. R.** 2003. Non-random association of transposable elements with duplicated genomic blocks in *Arabidopsis thaliana*. *Mol. Phyl. Evol.* 29: 410-416.
- Hughes, M. K. y Hughes A. L.** 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* 10:1360-1369.
- Hughes, A. L. y Nei, M.** 1989. Nucleotide substitution at major histoincompatibility complex class II loci: Evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* 86: 958-962.
- Hughes, A. L. y Nei, M.** 1998. Pattern of nucleotide substitution at major histoincompatibility complex loci reveals overdominant selection. *Nature* 335: 167-170.
- Irish, V. F. y Yamamoto, Y. T.** 1995. Conservation of floral homeotic gene function between *Arabidopsis* and *Antirrhinum*. *Plant Cell* 7: 1635-1644.
- Jarvis, E. E., Clark, K. L. y Sprague, G. F.** 1989. The yeast transcription activator PRTF, a homolog of the mammalian serum response factor, is encoded by the *MCM1* gene. *Genes Dev.* 3: 936-945.
- Jia, L., Clegg, M. T. y Jiang, T.** 2003. Excess non-synonymous substitutions suggest that positive selection episodes occurred during the evolution of DNA-binding domains in the *Arabidopsis* R2R3-MYB gene family. *Plant Mol. Biol.* 52: 627-642.
- Jukes, T. H. y Cantor, C. R.** 1969. Evolution of protein molecules. pp. 21-123. In Munro, H. N. (ed.) *Mammalian protein metabolism*. Academic Press, Nueva York.
- Karlin, S. y Altschul, S. F.** 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87: 2264-2268.
- Karlin, S. y Altschul, S. F.** 1993. Applications and statistics for multiple high-scoring segments from molecular sequences. *Proc. Natl. Acad. Sci. USA* 90: 5873-5877.
- Kimura, M.** 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.
- Kimura, M.** 1983. *The neutral theory of molecular evolution*. Cambridge University Press. Cambridge, Reino Unido.

- King, M.-C. y Wilson, A. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107-116.
- Klenk, H. P. *et al.* 1997. The complete genome sequence of the hyperthermophilic, sulphate reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390: 364-370.
- Kofuji, R., Sumikawa, N., Yamasaki, M., Kondo, K., Ueda, K., Ito, M. y Hasebe, M. 2003. Evolution and divergence of the MADS-box gene family based on genome-wide expression analyses. *Mol. Biol. Evol.* 20: 1963-1977.
- Kohler, C., Hennig, L., Spillane, C., Pien, S., Grissem, W. y Grossniklaus, U. 2003. The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene *PHERES1*. *Genes Dev.* 12: 1540-1553.
- Koshi, J. M. y Goldstein, R. A. 1996. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* 42: 313-320.
- Kreitman, M. y Hudson, R. R. 1991. Inferring the evolutionary histories of the *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* 127: 565-582.
- Krizek, B. A. y Meyerowitz, E. M. Mapping the protein regions responsible for the functional specificities of the *Arabidopsis* MADS domain organ-identity proteins. *Proc. Natl. Acad. Sci. USA* 1996: 4063-4070.
- Krizek, B. A., Riechmann, J. L. y Meyerowitz, E. M. 1999. Use of the APETALA1 promoter to assay the *in vivo* function of chimeric MADS-box genes. *Sex Plant Reprod.* 12: 14-26.
- Kumar, S., Tamura, K., Jakobsen, I. B. y Nei, M. 2001. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* 17: 1244-1245.
- Larget, B. y Simon, D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16: 750-759.
- Lawton-Rauh, A. L., Alvarez-Buylla, E. R. y Purugganan, M. D. 2000. Molecular evolution of flower development. *Trends Ecol. Evol.* 15: 144-149.
- Lehrman, M., Russell, D., Goldstein, J. y Brown, M. 1987. Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with hypercholesterolemia. *J. Biol. Chem.* 262: 3354-3351.
- Li, W.-H. 1980. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics* 95: 237-258.
- Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates. Sunderland, Massachusetts, Estados Unidos.
- Li, W.-H. y Graur, D. 1991. *Fundamentals of molecular evolution*. Sinauer Associates. Sunderland, Massachusetts, Estados Unidos.
- Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L. y Yanofsky M. F. 2000. SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* 404: 766-770.

- Lynch, M. y Force, A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459-473.
- Lynch, M. y Walsh, J. B. 1998. *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, Massachusetts.
- Ma, H., Yanofsky, M. F. y Meyerowitz, E. M. 1991. *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes Dev.* 5: 484-495.
- Maddison, W. P. 1991. Squared-change parsimony reconstructions of ancestral states for continuous valued characters on a phylogenetic tree. *Syst. Zool.* 40: 304-314.
- Maddison, W. P., Donoghue, M. J. y Maddison, D. R. 1984. Outgroup analysis and parsimony. *Syst. Zool.* 33: 83-103.
- Maddison, W. P y Maddison, D. R. 1987. *MacClade 2.1*. Programa y manual. Cambridge, Massachusetts, Estados Unidos.
- Maddison, W. P y Maddison, D. R. 1992. *MacClade Version 3: Analysis of phylogeny and character evolution*. Sinauer Associates, Inc. Sunderland, Massachusetts, Estados Unidos.
- Maddison, W. P y Maddison, D. R. 2004. Mesquite: a modular system for evolutionary analysis. Version 1.01 <http://mesquiteproject.org>
- Martínez-Castilla , L. P. y Alvarez-Buylla , E. R. 2003. Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. *Proc. Natl. Acad. Sci. USA* 100: 23407-13412.
- McVean, G. A. 2001. What do patterns of genetic variability reveal about mitochondrial recombination? *Heredity* 87: 613-620.
- Mendoza, L. y Alvarez-Buylla, E. R. 1998. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *J. Theor. Biol.* 193: 307-319.
- Mendoza, L., Thieffry, D. y Alvarez-Buylla, E. R. 1999. Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics* 15: 593-606.
- Messier, W. y Stewart, C.-B. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385: 151-154.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. y Teller, E. 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21: 1087-1091.
- Michaels, S. D. y Amasino, R. M. 1999. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* 11: 949-956.
- Miyata, T. y Yasunaga, T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.* 16: 23-36.

- Muse, S. V. y Gaut, B. S.** 1994. A likelihood approach for comparing synonymous and non-synonymous substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11: 715-724.
- Nam, J., dePamphilis, C. W., Ma, H. y Nei, M.** 2003. Antiquity and evolution of the MADS-box gene family controlling flower development in plants. 2003. *Mol. Biol. Evol.* 20: 1435-1447.
- Nam, J., Kim, J., Lee, S., An, G., Ma, H. Y Nei, M.** 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. y Roychoudhury, A. K.** 1973. Probability of fixation of nonfunctional genes at duplicate loci. *Am. Nat.* 107: 362-372.
- Nei, M., Rogozin, I. B. y Piontkivska, H.** 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl. Acad. Sci USA.* 97: 10866-10871.
- Nielsen, R y Yang, Z.** 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929-936.
- Norman, C., Runswick, M., Pollock, R. y Treisman, R.** 1988. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds the *c-fos* serum response element. *Cell* 55: 989-1003.
- Ohno, S.** 1970. *Evolution by gene duplication*. Springer, Berlín.
- Ohno, S.** 1973. Ancient linkage groups and frozen accidents. *Nature* 244: 259-262.
- Ohta, T.** 2000. Evolution of gene families. *Gene* 259: 45-52.
- Page, R. M. y Holmes, E. C.** 1998. *Molecular evolution: a phylogenetic approach*. Blackwell, Oxford, Reino Unido.
- Pagel, M.** 1994. Detecting correlated evolution in phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. London Ser. B* 255: 37-45.
- Parenicová, L., deFolter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., Cook, H. E., Ingram, R. M., Kater, M. M., Davies, B., Angenent, G. C. y Colombo, L.** 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., Ditta, G. S., Bauman, E., Wisman, E. y Yanofsky, M. F.** 2000. B and C floral identity functions require SEPALLATA MADS-box genes. *Nature* 405: 200-203.
- Pellegrini, L., Tan, S. y Richmond, T. J.** 1995. Structure of serum response factor core bound to DNA. *Nature* 376: 490-498.
- Rannala, B., y Yang, Z.** 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* 43: 304- 311.

- Riechmann, J. L. y Meyerowitz E. M. 1997. MADS domain proteins in plant development. *Biol. Chem.* 378: 1079-1101.
- Riechmann, J. L., Wang, M. y Meyerowitz E. M. 1996. DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. *Nucleic Acids Res.* 24: 3134-3141.
- Rodríguez, F., Oliver, J. L., Marin, A. y Medina, J. R. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142: 485-501.
- Rubin, G. M. *et al.* 2000. Comparative genomics of the eukaryotes. *Science* 387: 2204-2215.
- Roth, V. L. 1988. The biological basis of homology. pp. 1-26. In C. J. Humphries (ed.), *Ontogeny and sytematics*. Columbia University Press, Nueva York.
- Samach, A., Onuchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F. y Coupland, G. 2000. Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. *Science* 288: 1613-1616.
- Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* 28: 35-42.
- Sankoff, D. y Cedergren, R. J. 1983. Simultaneous comparisons of three or more sequences related by a tree. pp. 253-264. In Sankoff, D y Kruskal, J. B. (eds.) *Time warps, string edits and macromolecules; the theory and practice of sequence comparison*. Addison-Wesley, Reading, Massachusetts, Estados Unidos.
- Sankoff, D. y Rousseau, P. 1975. Locating the vertices of a Steiner tree in an arbitrary metric space. *Math. Prog.* 9: 240-246.
- Santelli, E. y Richmond, T. J. 2000. Crystal structure of MEF2A core bound to DNA at 1.5 Å resolution. *J. Mol. Biol.* 297: 437-449.
- Sawyer, S. A. 1999. GENECONV: A computer package for the statistical detection of gene conversion. *Distribuido por el autor, Departamento de Matemáticas, Washington University, Saint Louis, Estados Unidos. Disponible en <http://www.math.wustl.edu/~sawyer>.*
- Schierup, M. H. y Hein, J. 2000a. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156: 879-891.
- Schierup, M. H. y Hein, J. 2000b. Recombination and the molecular clock. *Mol. Biol. Evol.* 17: 1578-1579.
- Schluter, D. 1995. Uncertainty in ancient phylogenies. *Nature* 377: 108-109.
- Schluter, D., Price, T. D., Mooers, A. Ø., y D. Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* 51: 1699-1711.
- Shore, P. y Sharrocks, A. D. 1995. The MADS-box family of transcription factors. *Eur. J. Biochem.* 229: 1-13.
- Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. y Van de Peer, Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* . 99: 13627-13632.

- Sommer, H., Beltrán, J.-P., Huijser, P., Pape, H., Lönnig, W.-E., Saedler, H. y Schwarz-Sommer, Z.** 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.
- Suzuki, Y., Glazko, G. V. y Nei, M.** 2002. Overcredibility of molecular phylogenies obtained from bayesian phylogenetics. *Proc. Natl. Acad. Sci USA* 99: 16138-16143.
- Swofford, D. L.** 1991. PAUP: Phylogenetic analysis using parsimony, version 3.0. Illinois Natural History Survey, Cahmpaign, Illinois.
- Swofford, D. L. y Maddison, W. P.** 1987. Reconstructing ancestral character states under Wagner parsimony. *Syst. Zool.* 36: 293-325.
- Swofford, D. L. y Maddison, W. P.** 1992. Parsimony, character-state reconstructions and evolutionary inferences. pp. 186-223. In Mayden, R. L. (ed.) *Systematics, historical ecology and North American fresh water fishes*. Stanford University Press, Stanford, California, Estados Unidos.
- Swofford, D. L. y Olsen, G. J.** 1990. Phylogeny reconstruction. pp. 411-501. In: Hillis, D. M. y Moritz, C. (eds.), *Molecular systematics*. Sinauer Associates. Sunderland, Massachusetts, Estados Unidos.
- Swofford, D. L., Olsen, G. J., Waddell, P. J. y Hillis, D. M.** 1996. Phylogenetic inference. In Hillis, Moritz y Mable (eds.), *Molecular systematics* (2ª ed.) Sinauer Associates, Inc. Sunderland, Massachusetts, Estados Unidos.
- Tanaka, T. y Nei, M.** 1989. Positive Darwinian selection observed at the variable region genes of immunoglobulins. *Mol. Biol. Evol.* 6: 447-459.
- Tierney, L.** 1994. Markov chains for exploring posterior distributions. *Ann. Stat.* 22: 1701-1762.
- Thompson, J. D., Higgins, D. G. y Gibson, T. J.** 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
- Van Valen, L.** 1982. Homology and causes. *J. Morphol.* 173: 305-312.
- Vedder, E.** 1998. Do the evolution. In Pearl Jam, *Yield*. Sony Music.
- Vergara-Silva, F., Martínez-Castilla, L. y Alvarez-Buylla, E. R.** 2000. MADS-box genes: development and evolution of plant body plans. *J. Phycol.* 36: 803-812.
- Vision, T. J., Brown, D. G. y Tanksley, S. D.** 2000. The origins of genomic duplications in *Arabidopsis*. *Science.* 290: 2114-2117.
- Wagner, G. P.** 1989. The origin of morphological characters and the biological basis of homology. *Evolution* 43: 1157-1171.
- Watterson, G. A.** 1983. On the time for gene silencing at duplicate loci. *Genetics* 105: 745-766.

- Wilkstrom, N., Savolainen, V. y Chase, M. W. 2001. Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. Lond. Ser. B.* 268: 2211-2220.
- Williams, P. L. y Fitch, W. M. 1990. Phylogeny determination using dynamically weighted parsimony method. *Methods in Enzymology* 183: 615-626.
- White, S. E. y Doebley, J. 1998. Of genes and genomes and the origin of maize. *Trends Genet.* 14: 327-332.
- Whitkus, R., Doebley, J. y Lee, M. 1992. Comparative genome mapping of Sorghum and maize. *Genetics* 132: 1119-1130.
- Worobey, M. 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria and mitochondria. *Mol. Biol. Evol.* 18: 1425-1434.
- Yang, Z. 1998. Likelihood ratio test for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15: 568-573.
- Yang, Z., Kumar, S. y Nei, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141: 1641-1650.
- Yang, Z. y Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46: 409-418.
- Yang, Z. y Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908-917.
- Yang, Z., Nielsen, R., Goldman, N. y Pedersen, A.-M. K. 2000. Codon substitution models for heterogenous selection pressure at amino acid sites. *Genetics.* 155: 431-449.
- Yanofsky, M. F., Ma, H., Bowman, J. L., Drews, G. N., Feldman, K. A. y Meyerowitz, E. M. 1990. The protein encoded by the *Arabidopsis* homeotic gene *AGAMOUS* resembles transcription factors. *Nature* 346: 35-40.
- Yu, J. *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L.ssp. *indica*). *Science* 296: 79-92.
- Zhang, J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18: 292-298.
- Zhang, H. y Forde, B. G. 1998. An *Arabidopsis* MADS-box gene that controls nutrient-induced changes in root architecture. *Science* 279: 407-409.
- Zhang, J. y Nei, M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood and distance methods. *J. Mol. Evol.* 44 (Suppl.) S139-S146.
- Zhang, J., Rosenberg, H. F. Y Nei, M. 1998. Positive darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* 95: 3708-3713.

VI. APÉNDICES

(1) MADS-box genes: development and evolution of plant body plans. Vergara-Silva, F., Martínez-Castilla, L. y Alvarez-Buylla, E. R. 2000. *Journal of Phycology*. 36: 803-812.

ESTA TESIS NO SALE
DE LA BIBLIOTECA

MINIREVIEW

MADS-BOX GENES: DEVELOPMENT AND EVOLUTION OF PLANT BODY PLANS¹

Francisco Vergara-Silva, León Martínez-Castilla, and Elena R. Alvarez-Buylla²

Instituto de Ecología, UNAM, México D.F. 04510, México

We review functional data on MADS-box genes, recent phylogenetic analyses of these coding regions, and their roles in the development and evolution of key morphological innovations in plants. We map the origin of important morphological structures in particular diverse stages of the life cycle in different plant clades onto organismal phylogenies, and present relevant molecular genetic aspects of development related to the MADS-box genes. We focus on reproductive structures of the sporophyte because most functional characterizations have been done of MADS-box genes involved in flower development. We discuss MADS-box evolution in flowering plants, but we also review studies in the other nonflowering vascular plants, gymnosperms (conifers and gnetales), and ferns and preliminary data from the algae. We suggest that floral (e.g. flowering time, inflorescence, and flower meristem identity) MADS-box and nonfloral plant MADS-box genes should be the focus of future comparative research. Cloning and functional analyses of MADS-box genes in bryophytes, particularly in the experimental system *Physcomitrella patens* (Hedw.) B.S.G., are needed. The ABC model of floral organ specification is an excellent general representation of an important network of genes; however, formal analytical tools are required to integrate data on complex gene interaction in comparative analyses. This and other analytical approaches to constructing gene network models will help to frame homology hypotheses in an evolutionary and developmental framework.

Key index words: ABC model; character mapping; evolution of development; MADS-box; morphological evolution; phylogeny

The group of evolutionarily related sequences that comprises most of the floral homeotic loci is the MADS-box multigene family (Shore and Sharrocks 1995). The acronym of the family name is derived from the initials of its first four described members (Norman et al. 1988, Jarvis et al. 1989, Sommer et al. 1990, Yanofsky et al. 1990), whereas “box” refers to a conserved sequence of approximately 180 nucleotides. In a fashion analogous to the homeobox (Gehring et al. 1994), the MADS-box encodes a DNA-binding domain that allows the products that contain it to behave as transcription

factors. MADS-domain proteins have a second highly conserved segment with considerable similarity to the secondary structure of the animal cytoskeleton protein keratin (the K-domain); this segment seems to be involved in protein–protein interactions. The K-domain is connected to the MADS-box by an intermediate I region, moderately conserved among members characterized until now. Finally, the C-terminus, which encodes the putative transactivation domain, is poorly conserved among sequences (Riechmann and Meyerowitz 1997) (Fig. 1A). This is the structure of most plant MADS-box genes characterized up to now. However, recent studies have described a new group of plant MADS-box genes that lack the canonical MADS-I-K domains (MIK) structure of previously described plant genes (Alvarez-Buylla et al. 2000a). The latter are more closely related to the serum response factor (SRF)-like genes of animals than to the other plant MADS-box genes (Fig. 1A).

Molecular and genetic analyses of the mechanisms that control floral morphogenesis in *Arabidopsis thaliana* (L.) Heynh and *Antirrhinum majus* L. have provided one of the most elegant genetic models to date: the ABC model for specification of floral organ identity. In this model, the combined activities of a small number of loci are responsible for a complex phenotype through the orchestration of development (Coen and Meyerowitz 1991 and other references therein) (Fig. 1B). The main features of the model, well reviewed elsewhere (e.g. Lawton-Rauh et al. 2000), are summarized in three points. First, it comprises three partially overlapping fields of gene activity (A, B, and C; hence the name of the model) composed of genes that exclusively function in a particular domain but are not solely expressed there. Second, it defines and predicts organ identity on the basis of these combined activities. According to this idea, determination of the first floral whorl of sepals corresponds to the presence of A function alone. Similarly, petal organ specification results from the simultaneous participation of A and B functions, whereas stamens are determined from the sum of B and C functions. Finally, carpel or fourth whorl identity is specified by C function alone. A third feature of the model consists of a relationship of mutual antagonism between functions A and C, resulting in dominance of A when C is not present and vice versa.

Taking *A. thaliana* as a nomenclatural guide, the genes comprising the ABC model can be listed. All but one (*APETALA2*, an A-function gene) of the canonical ABC genes are members of the MADS-box gene fam-

¹ Received 30 March 2000. Accepted 16 August 2000.

² Author for correspondence: e-mail abuylla@servidor.unam.mx.

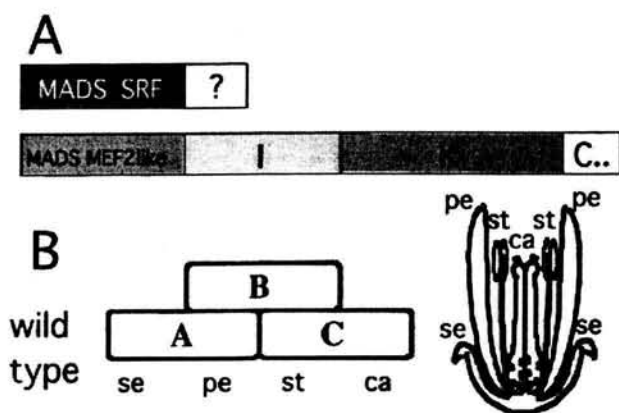


FIG. 1. (A) Schematic representation of plant MADS-domain proteins. Two types of MADS-domain proteins are found in plants: those that are more similar to the SRF-like genes from animals, called type I MADS-domain, and those more similar to the MEF2-like genes from animals, called Type II MADS-domains. Most of the latter proteins in plants have I, K, and COOH domains as shown in this figure (see text for more details). (B) The ABC combinatorial model proposes that the 4 different floral organs are determined by the specific combination of 3 different functions or activities. Activity A specifies sepals, activities A and B together specify petals, activities B and C specify stamens, and activity C alone specifies carpels. Additionally, the ABC model postulates a mutual inhibition between activities A and C, such that when function A is absent, function C takes its place and vice versa. To the right is a schematic representation of *Arabidopsis thaliana* flowers, which in wild-type plants are composed from the outside to the inside of 4 sepals (se), 4 petals (pe), 6 stamens (st), and 2 carpels (ca). The A-function genes are the MADS-box gene *APETALA1* (*AP1*) and the non-MADS-box gene *APETALA2* (*AP2*). The B-function genes are the MADS-box genes *APETALA3* (*AP3*) and *PISTILLATA* (*PI*). The C-function gene is the MADS-box gene *AGAMOUS* (*AG*). Recently, mutations in the *AGAMOUS*-like genes, *AGL2*, *AGL4*, and *AGL9*, have been identified and renamed *SEPALLATA1*, 2, and 3 (Pelaz et al. 2000). The triple mutant of these genes yields a flower with sepals in all 4 whorls, but single and double mutants of these genes do not alter the flower phenotype. Hence, these genes are functionally redundant and necessary for the B and C function genes (see text for further details).

ily. *APETALA1* is the remaining A-function gene. The B-function genes are *APETALA3* and *PISTILLATA*, and *AGAMOUS* is the C-function gene. All MADS-box genes mentioned here are depicted in Figure 2. Similar molecular genetic studies in *A. majus* have demonstrated the functional conservation at the genetic and developmental levels between the ABC MADS-box genes of *A. thaliana* and *A. majus* (Irish and Yamamoto 1995). This conservation is also seen in other model systems—plants like maize or petunia—and, as mentioned below, the molecular conservation extends at least to the entire seed plant clade.

Recent experiments with *A. thaliana* document an alternative supernumerary activity that is an interesting addition to the ABC functions (Pelaz et al. 2000). Pelaz and collaborators found that 3 members of the MADS-box gene family related to *AGAMOUS-LIKE 2* (*AGL2*), the genes *SEPALLATA1* (previously known as *AGL2*), *SEPALLATA2* (*AGL4*), and *SEPALLATA3* (*AGL9*) act

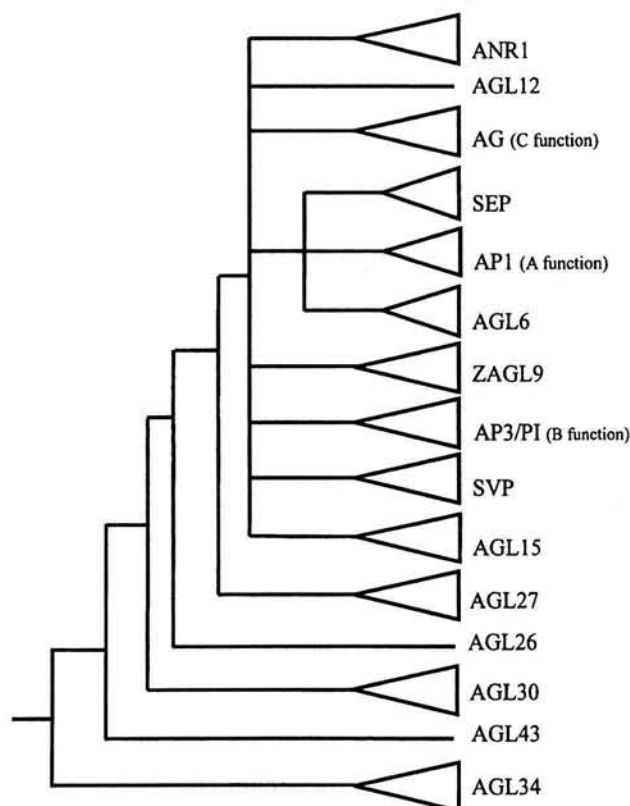


FIG. 2. Schematic representation of phylogenetic relationships among plant MADS-box genes. Groups of genes are represented by triangles and are named after *Arabidopsis thaliana* representatives except for ZAGL9 group. ANR1 group contains the *A. thaliana* genes *ANR1*, *AGL21*, *AGL17*, and *AGL16* as well as genes from *Antirrhinum*, *Medicago*, and *Gnetum*. AG group contains the *A. thaliana* genes *AGAMOUS*, *SHATTERPROOF1*, *SHATTERPROOF2* as well as genes from *Antirrhinum*, *Petunia*, *Nicotiana*, *Lycopersicon*, *Zea*, and *Oryza*. SEP group contains the *A. thaliana* genes *SEPALLATA1*, *SEPALLATA2*, *SEPALLATA3*, and *AGL3* as well as genes from *Petunia*, *Lycopersicon*, and *Pinus*. AP1 group contains the *A. thaliana* genes *APETALA1*, *CAULIFLOWER*, and *FRUITFULL* as well as genes from *Antirrhinum*, *Silene*, and *Zea*. AGL6 group contains the *A. thaliana* genes *AGL6* and *AGL13* as well as genes from *Zea*, *Pinus*, and *Picea*. ZAGL9 group contains only genes from *Zea* and *Aranda*. AP3/PI group contains the *A. thaliana* genes *APETALA3* and *PISTILLATA* as well as genes from *Antirrhinum*, *Petunia*, *Medicago*, *Nicotiana*, *Syringa*, *Lycopersicon*, *Solanum*, *Silene*, *Argyroxiphium*, *Papaver*, *Brassica*, *Dicentra*, *Caltha*, *Ranunculus*, *Delphinium*, *Michelia*, *Liriodendron*, *Peperomia*, *Piper*, *Zea*, *Oryza*, *Gnetum*, and *Ceratopteris*. SVP group contains the *A. thaliana* genes *SHORT VEGETATIVE PHASE* and *AGL24* as well as genes from *Gnetum*. AGL15 group contains the *A. thaliana* genes *AGL15* and *AGL18* as well as genes from *Zea* and *Ceratopteris*. AGL27 group contains the *A. thaliana* genes *AGL27*, *AGL31*, and *FLOWERING LOCUS F* as well as genes from *Gnetum* and *Ceratopteris*. AGL30 group contains the *A. thaliana* genes *AGL23*, *AGL28*, *AGL29*, *AGL39*, *AGL40* as well as genes from *Gnetum*. AGL34 group contains the *A. thaliana* genes *AGL30*, *AGL33*, *AGL34*, *AGL35*, *AGL36*, *AGL37*, *AGL38*, *AGL41*, and *AGL43* as well as a gene from *Ceratopteris*. This figure is based on the combined results of two maximum parsimony analyses. Nodes with less than 50% bootstrap support have been collapsed. Clades whose members' spatiotemporal expression patterns are in agreement with ABC model functions are shown. (Modified from Alvarez-Buylla et al. [2000b] and L. Martinez-Castilla, F. Vergara-Silva and E. R. Alvarez-Buylla, UNAM [unpublished data]).

redundantly in the three inner whorls of the flowers to determine petal, stamen, and carpel identity. Mutations in these genes do not affect the mRNA expression patterns of the ABC MADS-box genes, but they interfere with B- and C-functions at the protein-protein interaction level. The proteins encoded by these genes form heterodimers with PI, AP3, and AG. These three genes could therefore be considered to constitute a D-function. This function is, however, not active in 2 but rather in 3 adjacent whorls. Because the combined gene effect of the *SEPALLATA* genes is also not defined at the transcriptional but at the protein level, this interesting function has not been widely accepted as an additional activity in the ABC model.

Functional analyses based on loss-of-function and gain-of-function phenotypes have been done for other MADS-box genes. These studies show that genes of this family are involved in diverse aspects of plant ontogeny in addition to flower development. Several MADS-box genes have been shown to be important in determining the identity and time of formation of inflorescence and floral meristems. *APETALA1*, *CAULIFLOWER*, and *FRUITFULL* redundantly control inflorescence architecture by altering the expression and activity of two non-MADS-box genes, *LEAFY* and *TERMINAL FLOWER*, that are involved in the transition from vegetative to reproductive development (Ferrández et al. 2000). *SUPPRESSION OF OVEREXPRESSION OF CONSTANS1* (*SOCI1*, previously *AGL20*) is one of the targets of *CONSTANS*, which promotes flowering of *A. thaliana* in response to day length. This MADS-box gene is expressed in inflorescence meristems and then turns off in early flower meristems and comes on again later during flower development, suggesting that it might have additional roles not apparent from the single mutant (Samach et al. 2000).

Other MADS-box genes are involved in determining cell-type specification. This is the case of *SHATTERPROOF1* and *SHATTERPROOF2* that redundantly determine the proper development of the dehiscence zone of fruits (Liljegren et al. 2000). *FRUITFULL* is required for the normal pattern of cell division, expansion, and differentiation of the silique valves, for normal leaf development, and for normal inflorescence development (Gu et al. 1998). Several MADS-box genes have been shown to be involved in flowering time. *FLC* is very closely related to the *AGL27* and *AGL31* genes. They all have broad expression patterns and could share some functions, but *FLC* seems to have its own functions because the single *flc* loss-of-function mutant has a clear phenotype, suggesting that this gene is a flowering repressor (Michaels and Amasino 1999).

Several MADS-box genes are expressed predominantly or exclusively in roots, suggesting that these genes could also be involved in the morphogenesis of this plant organ. Notwithstanding the absence of mutant phenotypes that can be used to test the function of these genes in roots, cosuppression lines for one of the root genes, *ANRI*, suggests that this gene is important in controlling lateral root formation in response

to local availability of nitrogen (Zhang and Forde 1998). Many other MADS-box genes have been cloned and their expression patterns characterized (Fig. 2) (Alvarez-Buylla et al. 2000b), but functional analyses based on mutant phenotypes and/or overexpression lines have not been published for them. Nonetheless, their expression patterns provide a first guide to the functional characterization of these genes and are a starting point for further studies of functional and character evolution.

PHYLOGENETIC AND MOLECULAR EVOLUTIONARY ANALYSES OF THE MADS-BOX GENE FAMILY

MADS-box genes are not restricted to plants. Phylogenetic analyses that include representative members of all eukaryotic MADS-box genes sampled suggest that the first MADS-box-containing sequence was present in the common ancestor of the three main multicellular eukaryotic groups (fungi, plants, animals). It is possible that it was present earlier in prokaryotic lineages (Mushegian and Koonin 1996) and hence in the earliest eukaryotes. In addition to flower development in plants, these roles include regulation of muscle development in mammals and insects (Martin et al. 1993, Affolter et al. 1994, Lilly et al. 1994) and arginine metabolism in yeast (Dubois and Messenguy 1991). Recent analyses show that a duplication of an ancestral MADS-box-containing sequence probably gave rise to two main lineages of MADS-box genes before animals and plants diverged. Type I MADS-box genes in plants are more similar to the animal and fungal SRF-like sequences and include a group of recently identified *A. thaliana* sequences (*AGL34* clade) (Alvarez-Buylla et al. 2000a). Type II MADS-box genes include most plant MADS-box genes previously identified and characterized and the MEF2-like genes of animals and fungi. Only plant members of this lineage encode a K-domain downstream of the MADS-domain.

Recent phylogenetic analyses of MADS-box sequences in *A. thaliana* (Alvarez-Buylla et al. 2000b) resolve 7 new MADS-box gene clades (Fig. 2). These add to those previously identified, the largely flower-specific genes that comprise the core of the ABC model (J. J. Doyle 1994, Purugganan et al. 1995, Theissen et al. 1996; see above). Among the newly identified groups, 3 monophyletic clades of genes almost exclusively expressed in roots and leaves (*ANRI* and *AGL12*, *AGL14*) and 2 clades of widely expressed genes (*AGL15* and *FLC*) have been resolved (Fig. 2). Additionally, there are other well-supported clades (*AGL23-LIKE*, and the plant SRF-like genes) for which no expression or functional data are yet available. Phylogenetic studies have also revealed the existence of new groups of closely related and possibly functionally redundant MADS-box sequences (Alvarez-Buylla et al. 2000b). These analyses could be useful to guide further functional characterization of these genes, as it is possible that only double, triple (see Pelaz et al. 2000), or even quadruple mutants of these closely related sequences may show phenotypes amenable to further analysis.

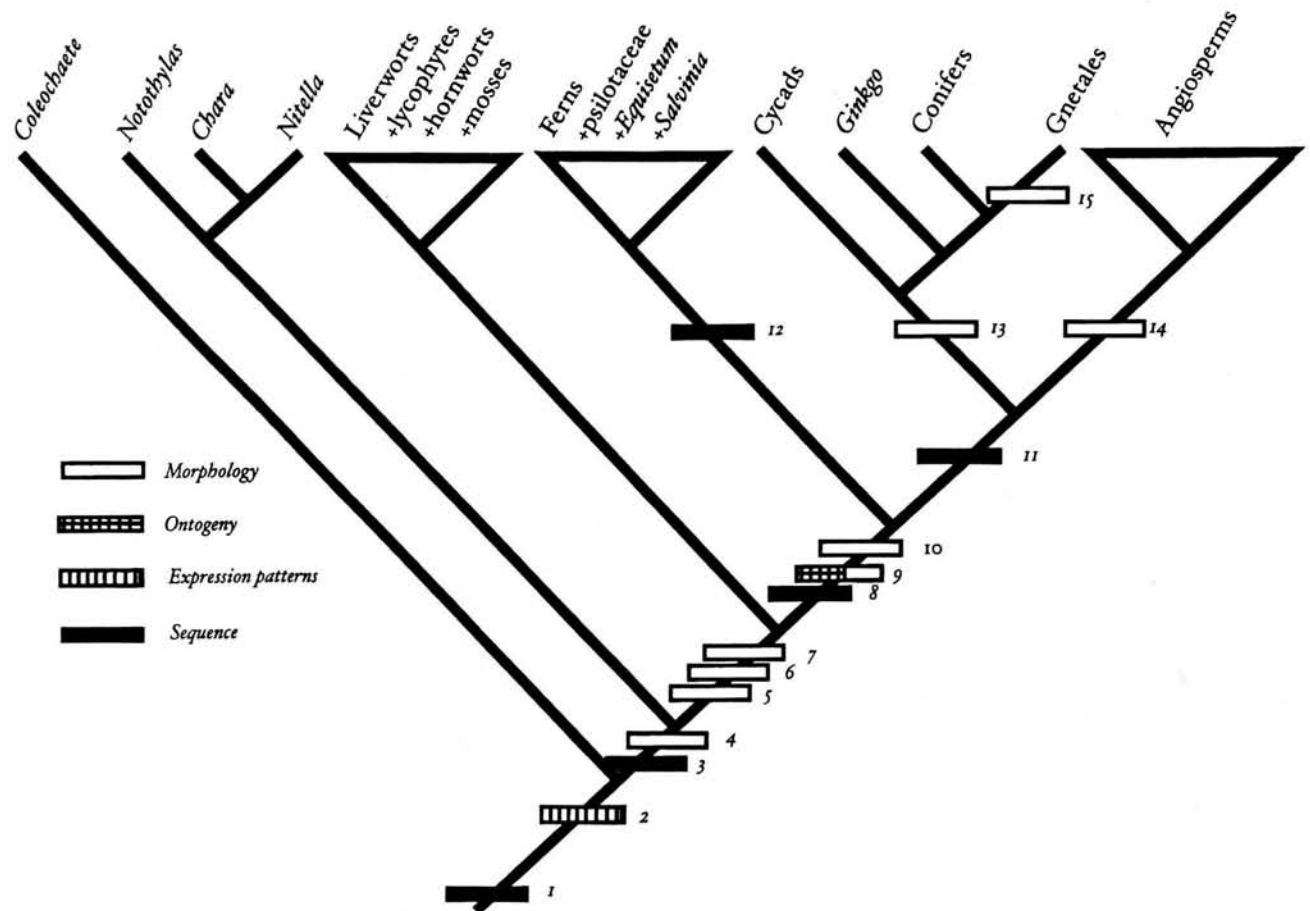


FIG. 3. Mapping of the evolutionary events leading to the establishment of the salient features of the plant body plan. Four main levels of homology defining events are depicted: appearance of protein families or subfamilies ("Sequence"), common spatiotemporal patterns of mRNA expression ("Expression patterns"), common embryological origins of plant modules and/or shifts in the relative prominence of alternating multicellular phases ("Ontogeny"), and shared morphological novelties ("Morphology"). 1, MADS-box genes; 2, Meristems; 3, MIK-type MADS-box genes; 4, Antheridia; 5, Multicellular sporophyte; 6, Sporangia; 7, Archegonia; 8, B and C classes of MIK MADS-box genes; 9, Ancestral reproductive structure of Tracheophyta (possibly, a product of B and C MADS-domain proteins acting in concert—C-class proteins would trigger the formation of either reproductive or nonreproductive structures, whereas the presence of B-class proteins would determine the sexual fate of these organs); 10, Passage from free-living or mycotrophic gametophyte to endosporic gametophyte and from dependent sporophyte to free-living sporophyte; 11, Class A of MIK MADS-box genes; 12, Non-ABC fern MADS-box genes; 13, Strobili; 14, Flowers; 15, Gnetalean "flowers." The morphological structures corresponding to 13, 14, and 15 are independent morphological elaborations based on a common genetic regulatory substrate—the ABC network. (Modified from Soltis et al. 1999.)

Arabidopsis thaliana and *A. majus* provide important models for the discovery of orthologous MADS-box sequences in a wide array of species within and outside the angiosperm clade. The successful cloning and characterization of homologous MADS-box genes in approximately 40 different plant genera have confirmed this idea. More than 30 of these taxa are angiosperms, whereas the remainder are nonflowering plants (Fig. 3). Among the latter group, *Chara* is an important taxon. Its MADS-box genes have a similar molecular structure to that of previously characterized MADS-box genes including a MADS, I, K, and COOH regions (Tanabe et al. 1999) as in Type II plant MADS-box genes. This is not surprising because phylogenetic analyses of the eukaryotic gene family suggest that both MIKC Type II and Type I (without K-box) genes should be found in green algae. This suggests

that they can be used to make inferences about the molecular basis of morphological evolution in all plant groups.

All angiosperms studied to date possess members of the canonical MADS-box ABC gene subfamilies (Theissen et al. 2000). Homologous sequences of these genes have also been cloned from conifers. The search for MADS-box genes in this seed plant group started in the Norway spruce (*Picea abies* (L.) Karst., Pinaceae). In this work, 3 homologous genes were found, one of which—named *DEFICIENS-AGAMOUS-LIKE 2* (*DAL2*)—was identified as a member of the AG/PLE gene clade (Tandre et al. 1995, 1998). Additionally, the presence of B-function homologues has been reported in the same species (Sundström et al. 1999). These findings are supported by independent research on another spruce species (*P. mariana* (Mill.) B.S.P., the black spruce) (Rutledge et al.

1998) and less directly by results on *Pinus radiata* D. Don (Mouradov et al. 1998a,b).

Representative species from the Order Gnetales also possess MADS-box genes (Shindo et al. 1999, Winter et al. 1999). These studies have found genes orthologous to the ABC subfamilies and MADS-box genes that cluster elsewhere in the phylogenies (Fig. 2). Five of the 13 homologues found by Winter et al. (1999) are grouped with relatively high bootstrap support with at least one of the conifer MADS-box genes and excluded the corresponding angiosperm orthologue. Shindo et al. (1999) found that 2 of 4 genes clustered with the angiosperm AG/PLE and the DEF/GLO sequence clades, respectively (Fig. 2). These studies suggest that the ABC genes evolved before the gymnosperm/angiosperm divergence, consistent with the estimated antiquity of the main plant MADS-box gene subfamilies (Purugganan 1997). In addition, they provide conclusive evidence against the taxonomic assignment given to the Gnetales over the last 2 decades as the sister group of angiosperms (Frohlich and Meyerowitz 1997, Frohlich 1999, Winter et al. 1999, Frohlich and Parker 2000). The latter result corroborates studies of other molecular markers (Hasebe et al. 1992, Goremykin et al. 1996, Chaw et al. 1997, Bowe et al. 2000, Chaw et al. 2000).

MADS-box genes have also been cloned from the fern genera, *Ceratopteris* and *Ophioglossum*, which belong to the leptosporangiate and eusporangiate types, respectively (Münster et al. 1997, Hasebe et al. 1998), and the lycopods (Svensson et al. 2000). In contrast to the gymnosperm MADS-box genes, these sequences do not group with high bootstrap values with any of the ABC clades, suggesting that the fern and lycopod MADS-box genes characterized up to now are not orthologous to any of the seed plant ABC genes. As mentioned above, preliminary reports on MADS-box gene cloning from charophyte algae suggest that these organisms also possess MADS-box genes (Tanabe et al. 1999). Although this is not surprising, given that the MADS-box gene family has been found in all eukaryotes, algal MADS-box genes known to date have a similar status to the fern MADS-box genes in that they cluster outside the ABC gene clades.

The three A, B, and C lineages and others that are not flower specific form a monophyletic group (J. J. Doyle 1994, Purugganan et al. 1995, Alvarez-Buylla et al. 2000b). However, this basal monophyletic group is not well resolved, and the only structure that can be discerned at this level is the sister group relationship of *FLC*-like genes to the remaining sequences (Alvarez-Buylla et al. 2000a). Rapid evolution could explain the lack of resolution in the branching order of the different clades shown in Figure 2. This is an appealing ad hoc hypothesis that should be tested with estimations of divergence times among gene clades and analyses of sequences from other species (e.g. Purugganan 1997).

The patterns of molecular evolution of the MADS-box gene family can be described by analyzing substitution rates (Nei and Gojobori 1986) of the different gene regions within and among clades. The MADS-box

proper has on average the lowest overall substitution rate, and the K-box and C-terminal regions evolve at rates as high as 3 and 10 times that of the MADS-box, respectively. Analyses also suggest that the diversification of the K-box and C-terminal regions play a greater role in amino acid sequence divergence between plant MADS-box genes than the other regions and these regions appear to be under strong purifying selection. This is because their ratios of nonsynonymous to synonymous substitutions, though higher than that of the MADS-box, are still very much lower than 1, the ratio expected for sequences free of constraint (Purugganan et al. 1995). This suggests that the evolution of MADS-box floral homeotic gene lineages, which played potentially important roles in the evolution of plant form, did not occur in a neutral fashion.

Substitution rate analyses of *APETALA1*-related genes within angiosperms suggest different selection pressures for genes performing different roles in flower and inflorescence development (paralogous genes) but not for genes performing similar roles in different organisms (orthologous genes) (Lawton-Rauh et al. 1999). These conclusions are based on analyses of orthologues of the *A. thaliana* genes *API*, *CAL*, *PI*, and *AP3* from its sister species *A. lyrata*, the confamilial *Brassica oleraceae* L., and the distantly related dicots *A. majus* and *Silene latifolia* Poiret. In these genes, no locus-by-lineage effects (Muse and Gaut 1997) were found, despite the considerable variation that these 5 species show in inflorescence and flower morphology. However, significant differences in the evolutionary dynamics of paralogous genes were found. For example, the *CAL* locus, which is only found in the Brassicaceae, seems to be evolving at a higher rate than the other paralogous genes, with an increase in the nonsynonymous substitution rate. Comparisons between the different domains of the genes also show differences in amino acid substitution rates, both among regions and paralogous loci.

A study in the well-known Hawaiian silversword alliance provides a very nice example of the possible coupling of MADS-box gene evolution and morphological diversification using *APETALA1* and *APETALA3* homologues. The ratios of nonsynonymous to synonymous substitutions in these genes for the Hawaiian silverswords were 3 times higher than the ratios observed for their ancestral species, the North American tarweeds (M. Barrier, M. D. Purugganan, NCSU; Robichaux, R. H., Univ. of Arizona, unpublished data). In fact, nearly 20% of the genes from the alliance members have substitution rates that are expected under adaptive selection. Additional tests have shown that the high levels of replacement substitutions are not due to an acceleration of the rate of neutral mutation in the island species (Barrier et al., unpublished data). Allozyme data suggest that structural protein evolution is not correspondingly high in the Hawaiian silverswords either (Witter and Carr 1988). Therefore, one possible hypothesis is that evolution at the studied regulatory MADS-box genes is responsible for the impressive morphological radiation observed among Hawaiian silverswords.

EVOLUTION OF PLANT MADS-BOX GENE FUNCTION

Mapping functional data of the plant MADS-box genes onto their corresponding gene trees allows the inference of different aspects about the evolutionary diversification of function within and among clades of the gene family. Previous studies had hypothesized that ancestral plant MADS-box genes were specific to vegetative structures and that selection for the evolution of specialized reproductive structures led to the high diversity of MADS-box genes approximately 450 million years ago, around the time of origin of the land plants (Purugganan 1997). However, more recent analyses made for all available data from *A. thaliana* suggest that basal genes, both in the global gene-family phylogeny and within each clade, have broader patterns of mRNA expression than more derived genes. These results suggest that the function of the plant MADS-box genes did not progress from vegetative to reproductive. Instead, it seems that differential and simultaneous gene recruitment was correlated with spatiotemporal restriction of expression pattern in both reproductive and vegetative structures as gene duplication events occurred (Alvarez-Buylla et al. 2000b) (Fig. 2). These latest results show that hypotheses related to the role of natural selection, which may explain the diversification of reproductive structures during early evolution of land plants, should also be applied to the diversification of vegetative structures.

As pointed out above, MADS-box genes that belong to different subfamilies outside the floral homeotic gene lineages have also been cloned from several angiosperm and nonangiosperm species (Fig. 2). If the occurrence of MADS-box genes in *A. thaliana* is a good guide, then homologues to several gene subfamilies have not been discovered yet in any other plant species. The analyses reviewed above will be important to interpret future findings of sequences homologous to MADS-box genes. Until now, only one monophyletic clade restricted to species other than *A. thaliana* has been found, the *ZAGL9* clade limited to *Zea mays* L., but nothing has yet been published about these genes. Combining the expression data of MADS-box genes with the gene tree for all sequences characterized to date shows that monophyletic clades are formed and the genes within them share expression patterns of high overall similarity. However, more expression data are needed to perform formal analyses of ancestral expression patterns with gene families that include species other than *A. thaliana*.

ROLE OF MADS-BOX GENES IN PLANT BODY PLAN EVOLUTION

Data regarding the taxonomic distribution, phylogenetic relationships, expression patterns, and functional interactions of the plant MADS-box genes can be mapped onto an organismal phylogeny. Such an analysis, performed simultaneously with information on the ontogenetic and adult morphological innovations responsible for body plan diversity in the plant

kingdom, is the ultimate objective of the recently emerged field of plant evolutionary developmental biology (Baum 1998). Investigating the molecular basis of morphological and developmental evolution in this way establishes a powerful approach to the problem of elucidating homology relationships among the entire spectrum of plant morphological structures. Although evolutionary and developmental character mapping exercises such as the one put forward in this work could be done for any cell, tissue, or structure type, our emphasis is on reproductive structures, because flower development has been most thoroughly studied in model systems.

Examples of the conservation of the ABC MADS-box gene regulatory network in angiosperms at the purely structural and expression pattern levels, along with mutant phenotypes, can be analyzed to corroborate or falsify homology relationship hypotheses among floral organs of species with contrasting morphologies. Flowering plants comprise more than a quarter million of species that exhibit an astounding diversity of ecological traits and interactions, as well as an apparently endless variation of floral morphological features. However, most share a stereotypical arrangement of floral organs (sepals, petals, stamens, and carpels). Perhaps the best example of the conservation of the ABC network and its use to assess homology relationships is provided by the study of Ambrose et al. (2000) on *Z. mays*. These authors established a correspondence between the highly derived floral organs of this monocot and those of the dicot model systems based on the analysis of the expression pattern of *SILKY1*, a B-function homologue and its mutants (Ambrose et al. 2000). The results of this study suggest that, despite their conspicuous differences, lodicules are modified petals and, possibly, palea and lemma are modified sepals.

Variations in the expression patterns of ABC genes can also be responsible for diversity in floral arrangements among angiosperms. For example, in the dioecious dicotyledon *Rumex acetosa* L. (sorrel; Polygonaceae), differential expression of RAPI, an AGAMOUS/PLENA (AG/PLE) homologue (Ainsworth et al. 1995), is responsible for sex determination. But the most outstanding flower morphological variation among angiosperms is the one found in the Mexican triurid, *Lacandonia schismatica* Martinez et Ramos (Martinez and Ramos 1989). This mycoheterotrophic monocot species is the only angiosperm with central stamens and carpels in the third whorl. The simplest hypothesis to explain this homeotic phenotype is to postulate a centripetal shift in the spatial domain of the B-function (Vergara et al. 1999).

In summary, comparative analyses of the ABC model have confirmed that it is a valid working hypothesis to consider the synergistic mode of floral organ determination as a synapomorphy of the angiosperms (Bowman 1997) (see the mapping of characters 9, 11, and 14 in Fig. 3). It can be concluded that the ABC model is an excellent guide for the interpretation of the molecular basis of homology among floral organs in angiosperms (Coen

and Meyerowitz 1991). Coding sequences of orthologous ABC genes are widely conserved among flowering plants, and most variations in flower arrangements can be explained in terms of changes in the expression domains of these genes, implying changes in their promoter sequences (Baum 1998).

As discussed above, studies of the occurrence of MADS-box genes successfully "went down" the species tree, resulting in the cloning and characterization of conifer MADS-box genes. Some of these genes cluster within the canonical ABC subfamilies (Fig. 2). The acceptance of an apparently natural seed plant group—implicit in the tree topology presented in Figure 3—has two inescapable consequences bearing directly on the interpretation of expression patterns of developmentally important genes in the gymnosperms, especially MADS-box genes. Given that one of the most recent estimations of the phylogenetic relationships between the Gnetales and other groups of gymnosperms actually considers them as derived conifers (Bowe et al. 2000), homology assessments among seed plant reproductive structures should be done and should be clarified first at this level. Thereafter, the common features found between conifers, Gnetales, angiosperms, and the yet to be described data in cycads and *Ginkgo* will allow the final estimation of the developmental genetic potential already present in the last common ancestor of the spermatophytes (seed plants).

Expression patterns of *DAL2*, that is, the AG orthologue from *Picea*, along with the phenotypes of *A. thaliana* transgenic plants overexpressing this conifer gene, suggest that this gene is the functional equivalent of the angiosperm C-function genes in reproductive organ determination (Tandre et al. 1998). In the compound female cone, this gene is exclusively expressed in the ovule-bearing scale. On the other hand, patterns of expression of B-function homologues in the same species (Sundström et al. 1999) suggest that pollen-bearing organ specification is also conserved between angiosperms and gymnosperms. Similar patterns have been found in *Gnetum*. In this gymnosperm, male reproductive structures only express these genes, whereas female reproductive axes have transcription products of both B-function and C-function homologues (Winter et al. 1999). Therefore, these authors have postulated that B-function homologues are the critical element in the developmental mechanism of sex determination in all seed plants. Clearly, this is an extraordinarily interesting conclusion because it defines a critical developmental character state of the last common ancestor of the seed plants but has yet to be tested in the cycads and *Ginkgo*.

Variations in reproductive organ molecular specification among gymnosperms would lead to possible differences in the estimation of ancestral character states in the last common ancestor of the seed plants; however, such variations are unlikely to be found. Therefore, the approach taken by Sundström et al. (1999), which consists of defining morphological homologues between gymnosperms and angiosperms in the most general level possible (e.g. microsporangia

should be considered instead of stamens or pollen cones), seems appropriate. On the other hand, Winter et al. (1999) do not address directly homology issues between conifers and the Gnetales. In contrast, Shindo et al. (1999) suggested that the gnetalean ovule is the homologue of the conifer ovule-ovuliferous scale complex. This means that the latter structure alone is homologous to the outer envelope of the *Gnetum* ovule. Interestingly, they do so in the context of an independent derivation of this structure in each corresponding lineage from a Cordaitales-like ancestor. More work is needed to refine homology hypotheses among gymnosperm reproductive structures and between these and those of angiosperms.

Comparative evolutionary and developmental analyses in the seed plants are important because they might become the basis to elucidate the mystery of the origin of the flower, Darwin's "abominable mystery." Probably, the origin of both flowers and strobili is a consequence of ontogenetic divergence (Rieppel 1993) in the action of a progressively better defined regulatory genetic network that was already present in the last common ancestor of seed plants. According to this point of view, the first event can be conceived as the result of the constitution of a gene regulatory network that allowed the formation of structures with aggregation of sporophylls. This event most likely occurred after the divergence of the fern lineages from the seed plants. This view and the aforementioned fact that the Gnetales are part of the monophyletic gymnosperm clade, rather than a sister group to the angiosperms, justify a rejection of traditional transformational hypotheses (i.e. anthophyte and neopseudanthial) (J. A. Doyle 1994) on the origin of flowers. Neither of these hypotheses can be validated if the last common ancestor of angiosperms, conifers, and the Gnetales is actually the last common ancestor of all seed plants, because both of them assume that the angiosperm flower was transformationally derived from a particular gymnospermous reproductive morphology. The above observations are summarized in the mapping of characters 9, 11, 13, 14, and 15 in Figure 3, where we show that the establishment of the ABC MADS-box gene network is a requisite for the divergent elaboration of the characteristic reproductive structures of gymnosperms and angiosperms, strobili, and flowers, respectively. Furthermore, the morphological characters that in the past were the basis for the clustering of Gnetales and the angiosperms in the anthophyte clade actually reflect convergence on the basis of the same developmental genetic potential.

Molecular evolutionary estimations of the time of divergence among the floral homeotic gene subfamilies placed their putative ancestral sequence or sequences in the Ordovician Period 478 ± 24 million years ago (Puruggannan 1997). In the species tree of Figure 3, this event falls into the node that defines the tracheophytes (ferns [gymnosperms + angiosperms]). However, as reviewed in previous sections, no estimation of the MADS-box gene genealogical relationships

has shown robust clustering of the fern sequences with any of the ABC clades. Therefore, the reconstruction of the corresponding ancestral structures at the basal node of the vascular plant clade could not incorporate ABC gene evolution. In other words, the "developmental integration of characters" (Abouheif 1997) or "true homology" (Bolker and Raff 1996) scenarios, valid for reconstructing ancestral seed plant reproductive structures, could not be applied to reconstruct ancestral states of tracheophyte reproductive structures. According to this second scenario, the origin of fern sporangia should be correlated with the presence of MADS-box gene subfamilies that are apparently exclusive to the entire fern group (Fig. 3, mapping of character 12) (Theissen et al. 2000).

PERSPECTIVES AND CONCLUSIONS: EVOLUTION OF REGULATORY GENETIC NETWORKS

The genetic–developmental–evolutionary events analyzed above are salient episodes in the origin of diversity of body plans in extant land plants. All these events depend, in turn, on the previous appearance of molecular specification mechanisms for the ontogenetic formation of multicellular sporophytes (Fig. 3, character 5) and meristems with increasing degrees of complexity (Fig. 3, character 2). However, it is clear that such large-scale morphological innovations have involved complex regulatory gene networks in which genes from other families participate (e.g. Bowman and Eshed 2000).

Most comparative analyses have been done with floral genes. Flowering time, inflorescence, and flower meristem identity MADS-boxes should be further explored to analyze the molecular basis of inflorescence architecture evolution. However, pattern and functional data on nonfloral plant MADS-box genes will be useful for addressing questions of morphological and ontogenetic components of homology outside sporophytic reproductive structures (Figs. 2 and 3) (Alvarez-Buylla et al. 2000b). For example, genes expressed in the seed plant gametophytes (Alvarez-Buylla et al. 2000b) will likely be useful to guide the cloning and characterization of homologues in plant groups outside the tracheophytes and to clarify the molecular basis of antheridial and archegonial evolution (mapped as characters 4 and 7 in Fig. 3). Other nonfloral MADS-box genes will probably be useful for studying other aspects of body plan evolution in species with less well-characterized MADS-box gene sets. Cloning and functional analyses of MADS-box genes in bryophytes, particularly in the experimental system, *Physcomitrella patens*, will certainly shed important insights into morphological evolution in plant groups from algae, bryophytes, and tracheophytes.

The ABC model of floral organ specification is an excellent general representation of a particularly important network of genes, because members of it have been found in every angiosperm species in which they have been looked for, playing indispensable homologous roles in the determination of organ identity. However, an ultimate goal, although still far from being achieved, should be to incorporate gene network mapping into the evolu-

tionary and developmental approach described here and to identify critical interactions at the transcriptional and posttranscriptional levels responsible for morphological innovations. To this end, formal analytical tools will have to be used to integrate complex gene interaction information. Preliminary trials for flower and root genes are available (Mendoza and Alvarez-Buylla 1998, Mendoza et al. 1999). This and other analytical approaches to constructing gene network models should help to develop homology hypotheses in an evolutionary and developmental framework.

We thank Barbara Ambrose and two anonymous reviewers for useful comments on previous versions of this manuscript. The authors were financed by UNAM, CONACYT, the PEW Foundation and Human Frontiers grants while completing this work.

- Abouheif, E. 1997. Developmental genetics and homology: a hierarchical approach. *Trends Ecol. Evol.* 12:405–8.
- Affolter, M., Montagne, J., Walldorf, U., Groppe, J., Kloter, U., LaRosa, M. & Gehring, W. J. 1994. The *Drosophila* SRF homolog is expressed in a subset of tracheal cells and maps within a genomic region required for tracheal development. *Development* 120:743–53.
- Ainsworth, C., Thangavelu, M., Crossley, S., Buchanan-Wollaston, V. & Parker, J. 1995. Male and female flowers from the dioecious plant *Rumex acetosa* show different patterns of MADS-box gene expression. *Plant Cell* 7:1583–98.
- Alvarez-Buylla, E. R., Pelaz, S., Liljegren, S. L., Gold, S. E., Burgeff, C., Ditta, G. S., Ribas de Pouplana, L. L., Martínez-Castilla, L. & Yanofsky, M. F. 2000a. An ancestral MADS-box gene duplication occurred prior to the divergence of plants and animals. *Proc. Natl. Acad. Sci. USA* 97:5328–33.
- Alvarez-Buylla, E. R., Liljegren, S. J., Pelaz, S., Gold, S. J., Burgeff, C. N. L., Ditta, G. S., Vergara-Silva, F. & Yanofsky, M. F. 2000b. MADS-box gene evolution beyond flowers: expression in pollen endosperm, guard cells, roots and trichomes. *Plant J.* (in press).
- Ambrose, B. A., Lerner, D. R., Ciceri, P., Padilla, C. M., Yanofsky, M. & Schmidt, R. 2000. Molecular and genetic analyses of the *SILKY1* gene reveal conservation in floral organ specification between eudicots and monocots. *Mol. Cell* 5:569–79.
- Baum, D. A. 1998. The evolution of plant development. *Curr. Opin. Plant Biol.* 1:79–86.
- Bolker, J. A. & Raff, R. A. 1996. Developmental genetics and traditional homology. *Bioessays* 18:489–94.
- Bowe, L. M., Coat, G. & DePamphilis, C. W. 2000. Phylogeny of seed plants based on all three plant genomic compartments: extant gymnosperms are monophyletic and Gnetales are derived conifers. *Proc. Natl. Acad. Sci. USA* 97:4092–7.
- Bowman, J. L. 1997. Evolutionary conservation of angiosperm flower development at the molecular and genetic levels. *J. Biosci.* 22:515–27.
- Bowman, J. L. & Eshed, I. 2000. Formation and maintenance of the shoot apical meristem. *Trends Plant Sci.* 5:110–15.
- Chaw, S. M., Zharkikh, A., Sung, H. M., Lau, T. C. & Li, W.-H. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol.* 14:56–68.
- Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl. Acad. Sci. USA* 97:4086–91.
- Coen, E. S. & Meyerowitz, E. M. 1991. The war of the whorls: genetic interactions controlling flower development. *Nature* 353:31–7.
- Doyle, J. A. 1994. Origin of the angiosperm flower: a phylogenetic perspective. *Plant Syst. Evol.* 8:7–29.
- Doyle, J. J. 1994. Evolution of a plant multigene family—towards connecting molecular systematics and molecular developmental genetics. *Syst. Biol.* 43:307–28.
- Dubois, E. & Messenguy, F. 1991. In vitro studies of the binding of the ARGR proteins to the ARG5,6 promoter. *Mol. Cell Biol.* 11:2162–8.

- Ferrándiz, C., Gu, Q., Martienssen, R. & Yanofsky, M. F. 2000. Redundant regulation of meristem identity and plant architecture by *FRUITFULL*, *APETALA1* and *CAULIFLOWER*. *Development* 127:725–34.
- Frohlich, M. W. 1999. MADS about Gnetales. *Proc. Natl. Acad. Sci. USA* 96:8811–3.
- Frohlich, M. W. & Meyerowitz, E. M. 1997. The search for flower homeotic gene homologs in basal angiosperms and Gnetales: a potential new source of data on the evolutionary origin of flowers. *Int. J. Plant Sci.* 158(Suppl):S131–42.
- Frohlich, M. W. & Parker, D. S. 2000. Evolutionary origin of flowers: two theories refuted, a new theory proposed. *Syst. Bot.* (in press).
- Gehring, W. J., Affolter, M. & Bürglin, T. 1994. Homeodomain proteins. *Annu. Rev. Biochem.* 63:487–526.
- Goremykin, V., Bobrava, V., Pahnke, J., Troitsky, A., Antonov, A. & Martin, W. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbdL* data do not support gnetalean affinities of angiosperms. *Mol. Biol. Evol.* 13: 383–96.
- Gu, Q., Ferrándiz, C., Yanofsky, M. F. & Martienssen, R. 1998. The *FRUITFULL* MADS-box gene mediates cell differentiation during *Arabidopsis* fruit development. *Development* 125:1509–17.
- Hasebe, M., Kofuji, R., Ito, M., Kato, M., Iwatsuki, K. & Ueda, K. 1992. Phylogeny of gymnosperms inferred from *rbdL* sequences. *Bot. Mag. Tokyo* 105:673–9.
- Hasebe, M., Wen, C.-K., Kato, M. & Banks, J. A. 1998. Characterization of MADS homeotic genes in the fern *Ceratopteris richardii*. *Proc. Natl. Acad. Sci. USA* 95:6222–7.
- Irish, V. F. & Yamamoto, Y. T. 1995. Conservation of floral homeotic gene function between *Arabidopsis* and *Antirrhinum*. *Plant Cell* 7:1635–44.
- Jarvis, E. E., Clark, K. L. & Sprague, G. F. 1989. The yeast transcription activator PRTF, a homolog of the mammalian serum response factor, is encoded by the *MCM1* gene. *Genes Dev.* 3:936–45.
- Lawton-Rauh, A. L., Buckler IV, E. S. & Purugganan, M. D. 1999. Patterns of molecular evolution among paralogous floral homeotic genes. *Mol. Biol. Evol.* 16:1037–45.
- Lawton-Rauh, A. L., Alvarez-Buylla, E. R. & Purugganan, M. D. 2000. Molecular evolution of flower development. *Trends Ecol. Evol.* 15:144–9.
- Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L. & Yanofsky, M. F. 2000. SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* 404:766–70.
- Lilly, B., Galewsky, S., Firulli, A. B., Schulz, R. A. & Olson, E. N. 1994. D-MEF2: a MADS-box transcription factor expressed in differentiating mesoderm and muscle cell lineages during *Drosophila embryogenesis*. *Proc. Natl. Acad. Sci. USA* 91:5662–6.
- Martin, J. F., Schwarz, J. J. & Olson, E. N. 1993. Myocyte enhancer factor (MEF) 2C: a tissue-restricted member of the MEF-2 family of transcription factors. *Proc. Natl. Acad. Sci. USA* 90:5282–6.
- Martínez, E. R. & Ramos, C. H. 1989. Lacandoniaceae (Triuridales): una nueva familia de México. *Ann. Miss. Bot. Gard.* 76:128–35.
- Mendoza, L. & Alvarez-Buylla, E. R. 1998. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *J. Theor. Biol.* 193:307–19.
- Mendoza, L., Thieffry, D. & Alvarez-Buylla, E. R. 1999. Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis. *Bioinformatics* 15:593–606.
- Michaels, S. D. & Amasino, R. M. 1999. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* 11:949–56.
- Mouradov, A., Glassick, T. V., Hamdorf, B. A., Murphy, L. C., Marla, S. S., Yang, Y. & Teasdale, R. D. 1998a. Family of MADS-box genes expressed early in a male and female reproductive structures of Monterey pine. *Plant Physiol.* 117:55–61.
- Mouradov, A., Hamdorf, B., Teasdale, R. D., Kim, J. T., Winter, K.-U. & Theissen, G. 1998b. A *DEF/GLO*-like MADS-box gene from a gymnosperm: *Pinus radiata* contains an ortholog of angiosperm B class floral homeotic genes. *Dev. Genet.* 25:245–52.
- Münster, T., Pahnke, J., DiRosa, A., Kim, J. T., Martin, W., Saedler, H. & Theissen, G. 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* 94:2415–20.
- Muse, S. V. & Gaut, B. S. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics* 146:393–9.
- Mushegian, A. R. & Koonin, E. V. 1996. Sequence analyses of eukaryotic developmental proteins: ancient and novel domains. *Genetics* 144:817–28.
- Nei, M. & Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3:418–26.
- Norman, C., Runswick, M., Pollock, R. & Treisman, R. 1988. Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the *c-fos* serum response element. *Cell* 55:989–1003.
- Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E. & Yanofsky, M. F. 2000. B and C floral organ identity functions require *SEPALLATA* MADS-box genes. *Nature* 405:200–3.
- 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.* 45:392–6.
- Purugganan, M. D., Rounsley, S. D., Schmidt, R. J. & Yanofsky, M. F. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. *Genetics* 140:345–56.
- Riechmann, J. L. & Meyerowitz, E. M. 1997. MADS domain proteins in plant development. *Biol. Chem.* 378:1079–101.
- Rieppel, O. 1993. The conceptual relationship of ontogeny, phylogeny and classification: the taxic approach. *Evol. Biol.* 27:1–32.
- Rutledge, R., Regan, S., Nicolas, O., Fobert, P., Coté, C., Bosnich, W., Kauffeldt, C., Sunohara, G., Séguin, A. & Stewart, D. 1998. Characterization of an *AGAMOUS* homologue from the conifer black spruce (*Picea mariana*) that produces floral homeotic conversions when expressed in *Arabidopsis*. *Plant J.* 15:625–34.
- Samach, A., Onouchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F. & Coupland, G. 2000. Distinct roles of *CONSTANS* target genes in reproductive development of *Arabidopsis*. *Science* 288:1613–6.
- Shindo, S., Ito, M., Ueda, K., Kato, M. & Hasebe, M. 1999. Characterization of MADS genes in the gymnosperm *Gnetum parvifolium* and its implication on the evolution of reproductive organs in seed plants. *Evol. Dev.* 1:180–90.
- Shore, P. & Sharrocks, A. D. 1995. The MADS-box family of transcription factors. *Eur. J. Biochem.* 229:1–13.
- Soltis, P. S., Soltis, D. E., Wolf, P. G., Nickrent, D. L., Chaw, S. M. & Chapman, R. L. 1999. The phylogeny of land plants inferred from 18S rDNA sequences: pushing the limits of rDNA signal? *Mol. Biol. Evol.* 16:1774–84.
- Sommer, H., Beltrán, J.-P., Huijser, P., Pape, H., Lönnig, W.-E., Saedler, H. & Schwarz-Sommer, Z. 1990. *DEFICIENS*, a homeotic gene involved in the control of flower morphogenesis in *Antirrhinum majus*: the protein shows homology to transcription factors. *EMBO J.* 9:605–13.
- Sundström, J., Carlsbecker, A., Svensson, M. E., Svenson, M., Johanson, U., Theissen, G. & Engström, P. 1999. MADS-box genes active in developing pollen cones of Norway spruce (*Picea abies*) are homologous to the B-class floral homeotic genes in angiosperms. *Dev. Genet.* 25:253–66.
- Svensson, M. E., Johannesson, H. & Engström, P. 2000. The *LAMB1* gene from the clubmoss, *Lycopodium annotinum*, is a divergent MADS-box gene, expressed specifically in sporogenic structures. *Gene* 253:31–43.
- Tanabe, Y., Hasebe, M., Nozaki, H. & Ito, M. 1999. Analysis of MADS-box gene from *Chara* (*Chara braunii*) which is one of green algae closely related to land plants. XVI Int. Bot. Cong. (Abstract):297.
- Tandre, K., Albert, V. A., Sundas, A. & Engström, P. 1995. Conifer homologues to genes that control floral development in angiosperms. *Plant Mol. Biol.* 27:69–78.
- Tandre, K., Svenson, M., Svensson, M. E. & Engström, P. 1998. Conservation of gene structure and activity in the regulation of reproductive organ development of conifers and angiosperms. *Plant J.* 15:615–23.
- Theissen, G., Kim, J. T. & Saedler, H. 1996. Classification and phylogeny of the MADS-box gene multigene family suggests defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. *J. Mol. Evol.* 43:484–516.

- Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J. T., Münster, T., Winter, K.-W. & Saedler, H. 2000. A short history of MADS-box genes in plants. *Plant Mol. Biol.* 42:115–49.
- Vergara, F., Ferrandiz, C., Meyerowitz, E. & Alvarez-Buylla, E. R. 1999. Molecular basis and evolution of the inside-out flower of *Lacandonia schismatica*. XVI Int. Bot. Cong. Presentation 15.13.2.
- Winter, K.-U., Becker, A., Münster, T., Kim, J. T., Saedler, H. & Theissen, G. 1999. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. *Proc. Natl. Acad. Sci. USA* 96:7342–7.
- Witter, M. S. & Carr, G. D. 1988. Adaptive radiation and genetic differentiation in the Hawaiian silverword alliance (Compositae: Madiinae). *Evolution* 42:1278–87.
- Yanofsky, M. F., Ma, H., Bowman, J. L., Drews, G. N., Feldmann, K. A. & Meyerowitz, E. M. 1990. The protein encoded by the *Arabidopsis* homeotic gene *AGAMOUS* resembles transcription factors. *Nature* 346:35–40.
- Zhang, H. & Forde, B. G. 1998. An *Arabidopsis* MADS-box gene that controls nutrient-induced changes in root architecture. *Science* 279:407–9.

(2) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Alvarez-Buylla, E. R., Pelaz, S., Liljegren, S. J., Gold, S. E., Burgeff, C., Ditta, G. S., Ribas de Pouplana, Ll., Martínez-Castilla, L. y Yanofsky, M. F. 2000. *Proceedings of the National Academy of Sciences, USA*. 97: 5328-5333.

An ancestral MADS-box gene duplication occurred before the divergence of plants and animals

Elena R. Alvarez-Buylla^{*†‡}, Soraya Pelaz^{*}, Sarah J. Liljegren^{*}, Scott E. Gold^{*§}, Caroline Burgeff[†], Gary S. Ditta^{*}, Lluís Ribas de Pouplana[¶], León Martínez-Castilla[†], and Martin F. Yanofsky^{*‡}

^{*}Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116; [†]Instituto de Ecología, Universidad Nacional Autónoma de México, AP-Postal 70-275, México D.F. 04510, México; and [‡]Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92117

Communicated by Elliot M. Meyerowitz, California Institute of Technology, Pasadena, CA, March 13, 2000 (received for review September 8, 1999)

Changes in genes encoding transcriptional regulators can alter development and are important components of the molecular mechanisms of morphological evolution. MADS-box genes encode transcriptional regulators of diverse and important biological functions. In plants, MADS-box genes regulate flower, fruit, leaf, and root development. Recent sequencing efforts in *Arabidopsis* have allowed a nearly complete sampling of the MADS-box gene family from a single plant, something that was lacking in previous phylogenetic studies. To test the long-suspected parallel between the evolution of the MADS-box gene family and the evolution of plant form, a polarized gene phylogeny is necessary. Here we suggest that a gene duplication ancestral to the divergence of plants and animals gave rise to two main lineages of MADS-box genes: TypeI and TypeII. We locate the root of the eukaryotic MADS-box gene family between these two lineages. A novel monophyletic group of plant MADS domains (AGL34 like) seems to be more closely related to previously identified animal SRF-like MADS domains to form TypeI lineage. Most other plant sequences form a clear monophyletic group with animal MEF2-like domains to form TypeII lineage. Only plant TypeII members have a K domain that is downstream of the MADS domain in most plant members previously identified. This suggests that the K domain evolved after the duplication that gave rise to the two lineages. Finally, a group of intermediate plant sequences could be the result of recombination events. These analyses may guide the search for MADS-box sequences in basal eukaryotes and the phylogenetic placement of new genes from other plant species.

MEF2 | SRF | homeotic genes | MADS | development

Changes in genes encoding transcriptional regulators may represent the most important determinants of morphological evolution in plants and animals (1), and phylogenetic analyses provide a historical framework to identify such changes. The MADS-box genes encode a eukaryotic family of transcriptional regulators involved in diverse and important biological functions, ranging from cardiac muscle development in animals to pheromone response in yeast (2). In plants, MADS-box genes encode the three floral homeotic functions predicted by the genetic ABC model of flower organ identity (3, 4). In addition, plant MADS-box genes regulate the timing of flower initiation and flower meristem identity, as well as various aspects of ovule, fruit, leaf, and root development (4, 5).

Previously identified plant MADS-box genes encode proteins that share a stereotypical MIKC structure (Fig. 1), with the highly conserved DNA-binding MADS domain at the amino terminus. The moderately conserved K domain in the central portion of these proteins has been shown to be important for protein-protein interactions and likely forms a coiled-coil structure. The MADS and K domains are linked to one another by a weakly conserved I domain, whereas a poorly conserved carboxyl-terminal (C) region may function as a trans-activation domain (4). In animals and fungi, two distinct types of MADS-box genes have been identified, the SRF-like and MEF2-like classes (ref. 2; see Fig. 1).

This paper provides a hypothesis on the evolutionary history of the eukaryotic MADS-box gene family. Previous studies of eukaryotic MADS-box gene evolution, which included plant and animal sequences, provided unrooted trees useful to infer the phylogenetic relationships of the MADS-box lineages (6). These previous studies suggested that at least one MADS-box gene was present in the common ancestor of plants, animals, and fungi, and that probably the duplication that gave rise to the animal MEF2- and SRF-like genes occurred after animals diverged from plants but before fungi diverged from animals (6). However, previous plant and eukaryotic studies were based on a relatively small sampling of plant MADS-box sequences for a particular species (6–9). To test whether all *Arabidopsis* MADS-box sequences group in a monophyletic clade distinct from all animal and fungal MADS-box sequences, we performed phylogenetic analyses. We used 45 *Arabidopsis* MADS domain sequences, including 26 new ones, 9 sequences representative of the MEF2-like class from animals, and 8 sequences from the animal SRF-like group.

We present a rooted phylogenetic tree of the eukaryotic MADS domain lineages and postulate new hypotheses on the evolutionary history of this gene family. Our results suggest that a duplication ancestral to the divergence of plants and animals gave rise to two lineages (herein called TypeI and TypeII MADS), and that the protein motifs that define each group were fixed in the common ancestors of plants, animals, and fungi. Our analyses also identify new monophyletic clades of plant MADS-box sequences. Most plant MADS-box genes including all of the ones that have been characterized functionally in previous studies, group with the animal MEF2-like sequences in what we have named the TypeII MADS-box lineage. But we have identified a group of *Arabidopsis* MADS-box sequences that seems to be more closely related to the animal SRF-like genes forming the group that we herein call TypeI MADS. This finding suggests that both lineages are present in plants, animals, and fungi. Finally, we show that the K domain, typical of plant MADS-domain proteins, is found only in the TypeII MADS domain sequences of plants, suggesting that this domain evolved after this lineage diverged from the TypeI MADS. These results have enabled us to put forward a model for the evolution of this important family of regulatory genes in eukaryotes (see Fig. 4).

Materials and Methods

Sequence Sources and/or Accession Numbers. Sequence sources or GenBank accession numbers are as follows: *AGAMOUS* (10),

Abbreviations: MP, maximum parsimony; NJ, neighbor joining; QP, quartet puzzling; USP, Universal Stress Protein.

[†]To whom reprint requests should be addressed. E-mail: abuylla@servidor.unam.mx or marty@biomail.ucsd.edu.

[§]Present address: Department of Plant Pathology, University of Georgia, Athens, GA 30602-7274.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. 51734 solely to indicate this fact.

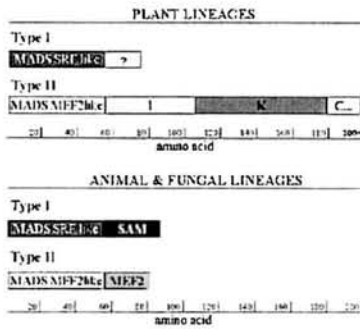


Fig. 1. Schematic representation of the protein domains of plant, animal, and fungal Type I (SRF-like) and Type II (MEF2-like) MADS-domain proteins. The scale indicates the number of amino acids along the protein. Plant Type II-like proteins have carboxyl-terminal domains that go beyond 200 amino acids. In plant Type I-like proteins the "?" indicates carboxyl-terminal domains not well defined yet and of variable lengths.

APETALA3 (11), *PISTILLATA* (12), *AGL1-6* (13), *APETALA1* (14), *AGL8* (15), *AGL9* (16), *CAULIFLOWER* (17), *AGL11*, *AGL12*, *AGL13*, *AGL14*, *AGL15* and *AGL17* (18), *AGL16* (AL137080, S.L. and M.Y., unpublished data), *AGL18* (AL137080, S.G. and M.Y., unpublished data), *AGL19* (AL161558, S.G. and M.Y., unpublished data), *AGL20* (AC003680, S.G. and M.Y., unpublished data), *AGL21* (AT120D10), *AGL22* (AC006592), *AGL23* (AC004512), *AGL24* (AF005158), *AGL25* (AF116527), *AGL26* (AF007270), *AGL27* (AC002291/cDNA sequence, S.P. and M.Y., unpublished data), *AGL28* (Y12776), *AGL29* (AC004077), *AGL30* (AC004138), *AGL31* (T45787/cDNA sequence, S.P. and M.Y., unpublished data), *AGL32* (AB007648), *AGL33* (AC004484), *AGL34* (AF058914), *AGL35* (AF058914), *AGL36* (AF058914), *AGL37* (AC00451), *AGL38* (AC004512), *AGL39* (AF007271), *AGL40* (Z99708), *ANR1* (19). *AGL23*, *AGL26*, and *AGL28-38* were recently identified by the *Arabidopsis* Genome Sequencing project. Although we lack cDNA clones for these genes, their predicted MADS-box domain sequences, on which our analyses are based, are unequivocal, because no introns have ever been found in this region.

GenBank accession numbers for the animal and fungal sequences are as follows. The MEF2-like genes used are: *Homo sapiens* *MEF2C* (L08895), *Caenorhabditis elegans* *CEMEF2* (U36198), *H. sapiens* *MEF2A* (S25831), *H. sapiens* *MEF2D* (Q14814), *Halocynthia roretzi* *ASMEF2* (D49970), *H. sapiens* *MEF2B* (X68502), *Drosophila melanogaster* *DMEF2* (U03292), *Saccharomyces cerevisiae* *SMP1* (P38128), and *S. cerevisiae* *RLM1* (D63340). The SRF-like genes used are: *H. sapiens* *SRF* (J03161), *Xenopus laevis* *SRF* (S15018), *D. melanogaster* *DSRF* (X77532), *S. cerevisiae* *MCM1* (P11746), *S. cerevisiae* *ARG80* (X05327), and *Schizosaccharomyces pombe* *PLN* (D78483). The bacterial Universal Stress Protein (*USP*) family sequences that served as outgroup for some of the analyses are: *Escherichia coli* *EcuspA* (X67639), *E. coli* *Ecyiit* (P32132), *Coxiella burnetii* *Coxymf* (P45680), and *Bacillus subtilis* *Bsxyic* (P42297).

Alignment and Phylogenetic Analyses. We used 65 amino acid sequences for the analyses. These cover the 57–60 amino acids that different authors (2, 6) have defined as the MADS domain plus a few additional conserved amino acids. These sequences were aligned by using CLUSTAL X; the alignment generated was unambiguous (complete alignment available from authors on request, and see Fig. 2). Phylogenetic analyses were conducted with unweighted maximum parsimony (MP), neighbor joining (NJ), and quartet puzzling (QP), by using the test version 4d64 of PAUP* (D. L. Swofford, Laboratory of Molecular Systematics, Smithsonian Institution, Washington, D.C.). For MP analyses,

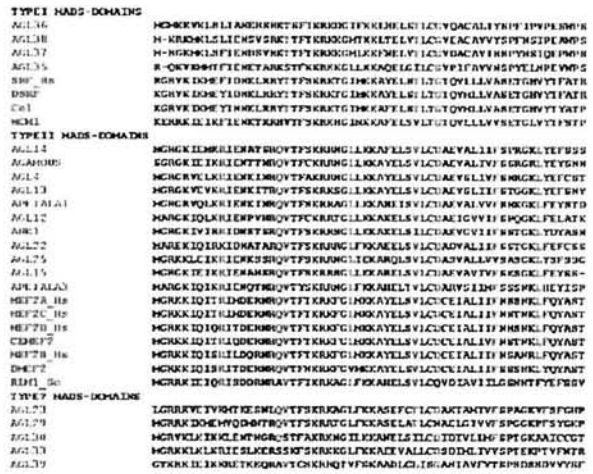


Fig. 2. Amino acid alignment of the MADS-domain (amino acids 1 to 60) for some representative members of the plant, animal, and fungal Type I (SRF-like) and Type II (MEF2-like) lineages. We also show representative sequences of the genes that are not clearly assigned to either one (MADS-domains Type?). One gene from each monophyletic clade identified in MP and NJ was selected. Conserved amino acids within each group and not found in any (or in no more than two) of the MADS domains of the other group are in red. Green names indicate plant sequences and red names, animal or fungal ones (see *Materials and Methods*).

100 replicates of random addition sequences keeping all optimal trees in each replicate, TBR branch swapping, and no maxtrees limit were used. Gaps were treated as missing data. NJ analyses were done by using the default factory settings and the p-distance (proportion of different amino acids between two sequences) as a distance estimator. This is the recommended distance measure when comparing distantly related sequences, because it has a smaller variance than other estimates (20).

Nonparametric bootstrap (100 pseudoreplicates) was used to assess the reliability of individual branches. Bootstrap proportions are considered here as an index of support for a particular clade and not a statement about probability or confidence limit in the statistical sense (21). QP trees were based on 1,000 replicates by using the factory default settings. The phylogenetic relationships inferred from the trees presented here do not depend on specific sequences used to estimate phylogeny; by using subsamples of protein sequences, the same relationships were inferred (data not shown). Trees were examined with TREEVIEW (22).

To study the branching order of MADS-box gene lineages and the timing of duplications relative to the divergence of the main groups of eukaryotes (plants, animals, and fungi), we need a rooted tree. An unambiguous root location depends on using an outgroup MADS-box domain sequence. We have attempted this rooting by using four bacterial sequences that belong to the USP family as outgroup. These share very few conserved amino acids with known eukaryotic MADS-box sequences but have been defined as MADS-domain homologues based on these few conserved residues and other functional criteria (23). A better outgroup could come from a taxon representative of a sister clade of plants, animals, and fungi, such as *Euglena*, but this is not yet available.

As an alternative way to objectively root the MADS-box tree, we used a parsimony-based approach from Page and Charleston (24, 25). This method reconciles the gene tree to the species tree and finds the rooted gene tree that minimizes the number of gene sorting events (which could include gene losses or insufficient sampling of genomes) and duplications. This is the MADS

domain tree that we put forward as a polarized phylogenetic hypothesis for this gene family. We used the species tree proposed by Baldauf and Palmer (26), in which animals and fungi are each other's closest relatives. We used groups of sequences that were shared by the NJ and MP trees, which were supported by high bootstrap values and were *bona fide* subfamilies, as possible outgroups to be tested. We tested seven alternative outgroups from the NJ and MP searches. The reconciled trees' method requires completely resolved trees. Therefore, one tree from each island sampled in the MP search was used. Trees from each island were very similar and differed only in some of the terminal branches. To avoid a bias because of the excessive number of possible losses and duplications found in the terminal branches where only taxa from either plants (only *Arabidopsis*) or one of the animal or fungal groups used were represented, we repeated this analysis by counting only basal duplications (i.e., those that are at the base of clades that combine sequences of plants, animals, and fungi).

Protein Structure Prediction. The predictions of coiled-coil regions within the protein sequences were performed with the programs PAIRCOIL and MULTICOIL (27, 28) and were based on the presence in the sequences of heptat-repeat signature motifs. In all cases, both programs used yielded the same result. A K domain was predicted to be present when the probability cutoff of finding coiled-coils downstream of the MADS-box domain was >0.35 . The default value of 0.5 has been determined empirically to work well. However, to avoid false negatives, we decreased the cutoff value by 20%. Additionally, we predicted possible protein secondary structures using discrete state-space probability models, as implemented by the program PSA (<http://bmerc-www.bu.edu/psa>; ref. 29). These predictions identified α -helices for the same sequences and were used to confirm results obtained from the coiled-coil prediction programs.

Results and Discussion

Ancient Duplications of Eukaryotic MADS-Box Sequences. We present molecular evolutionary analyses of plant, animal, and fungal MADS-domain sequences, including 26 newly identified MADS-domain sequences from *Arabidopsis*, along with 19 previously analyzed members of this extensive gene family. The most striking result of our analyses is the discovery that animal and fungal MEF2-like sequences are more closely related to most plant MADS-domain sequences than to animal SRF-like sequences. Some conserved amino acids put the MEF2-like animal and most plant sequences in a clear monophyletic clade (hereafter referred to as TypeII MADS domains), suggesting that at least one gene-duplication event occurred before the divergence of plants and animals. In addition, a group of *Arabidopsis* MADS-domain sequences (AGL34-like) seem to share a more closely related ancestor with the SRF-like sequences of animals and fungi than with other plant MADS-domain sequences. The clade formed by these two related groups is referred to hereafter as the lineage of TypeI MADS domains. However, the monophyly of this group is not as well supported as that of the TypeII MADS domains, because it is supported by very few shared and unique amino acids (Fig. 2). Finally, we found a group of intermediate plant sequences that could be the result of recombination between TypeI and II MADS-box genes. These results are based on NJ, QP, and MP phylogenies, described below.

The NJ tree rooted with the putative MADS-domain sequences from bacteria is well resolved (Fig. 3a) and is similar to the one obtained by the rooting method described below (Fig. 3b). In the tree of Fig. 3a, the TypeII MADS domains that group the animal MEF2-like and most plant sequences form a well-supported monophyletic clade. However, the rest of the clades that in Fig. 3b are grouped into the TypeI lineage do not form a monophyletic group in Fig. 3a. Results in Fig. 3a suggest that

AGL39-like sequences were lost or have not been found in animals and fungi. Both of the latter possibilities are unlikely, because yeast and *C. elegans*, whose genomes are completely sequenced, have both TypeI and TypeII MADS domains and no other types. It would be highly improbable that in both organisms the same genes were lost. We also performed MP analyses using the bacterial sequences as outgroup (not shown), but the strict consensus MP tree for these sequences does not resolve any basal branching other than that of the bacterial sequences. In the rest of the analyses, we have included only the eukaryotic MADS-domain sequences.

An alternative way to root the MADS-domain protein tree objectively is to use Page and Charleston's (24) approach to find the root position that minimizes the number of duplications and sorting events in the protein tree, when this is reconciled to the species tree (see *Materials and Methods*). We show the rooted NJ tree that minimized the reconciliation cost (49 total or 3 basal duplications and 17 sorting events) as the polarized phylogenetic hypothesis for this gene family. The bootstrap NJ tree reveals two well supported ($>50\%$) clades. The first one is constituted by the TypeI MADS-domain sequences and groups the animal SRF-like genes with two newly identified plant lineages, AGL34- and AGL23-like, plus AGL30, AGL33, and AGL39. The second, TypeII MADS-domain sequences, includes the rest of the plant sequences and the animal MEF2-like sequences.

Using MP analyses, we obtained a total of 647 most parsimonious trees (consistency index = 0.544, retention index = 0.695, rescaled consistency index = 0.378) of a length of 700 steps. The strict consensus-rooted MP tree resolves the monophyletic clade that includes animal SRF-like and plant AGL34-like sequences plus AGL30, AGL33, and AGL39, but with a low bootstrap support ($<50\%$). In contrast to the NJ tree, the strict consensus MP tree identifies the AGL-23 plant MADS-domain clade as a sister branch of the animal MEF2-like sequences, but with a very low bootstrap support ($<20\%$). The MP tree also resolves the AGL25 clade as sister to the monophyletic group formed by the rest of the plant TypeII and the animal MEF2-like sequences, also with a very low bootstrap support ($<20\%$). MP groups the animal and fungal MEF2-like sequences with the plant MADS-domain sequences in a monophyletic clade and places the animal and fungal SRF-like sequences as sister group with a good bootstrap support ($>50\%$).

When reconciled to the species tree, the least costly MP gene tree still requires a greater number of basal gene duplications and losses (49 total or 8 basal duplications and 22 sorting events) than the NJ tree shown (Fig. 3b). This MP tree also defined TypeI and TypeII groups as sister to each other. These results confirm that the most parsimonious root location among all trees tested is between the TypeI and TypeII lineages that we have identified. We compared the length of the Bootstrap NJ topology with the MP strict consensus tree using MACCLADE (Ver. 3.0) and found that they are of equal length. Therefore, based on the data at hand, we propose the tree shown in Fig. 3b as the most parsimonious hypothesis on the polarized evolutionary history of the eukaryotic MADS-box gene family. Finally, the QP tree also resolved the same TypeI and TypeII clades formed by the same family members as in the NJ tree shown (frequency value equal to 40%).

The inconsistent placement of the AGL23 clade between the NJ/QP and MP topologies, as well as the low bootstrap value for the TypeI clade in the MP strict consensus tree, suggests that some plant sequences cannot be unambiguously associated to either the TypeI or TypeII lineages. In fact, if AGL30, AGL33, AGL39, and the AGL23-like genes are removed, NJ, MP, and QP analyses yield resolved and well supported trees (bootstrap values of $>90\%$ and 50% for both lineages in NJ and MP analyses, respectively; see Fig. 3b; and 89% frequency in QP). These problematic sequences could be the result of recombina-

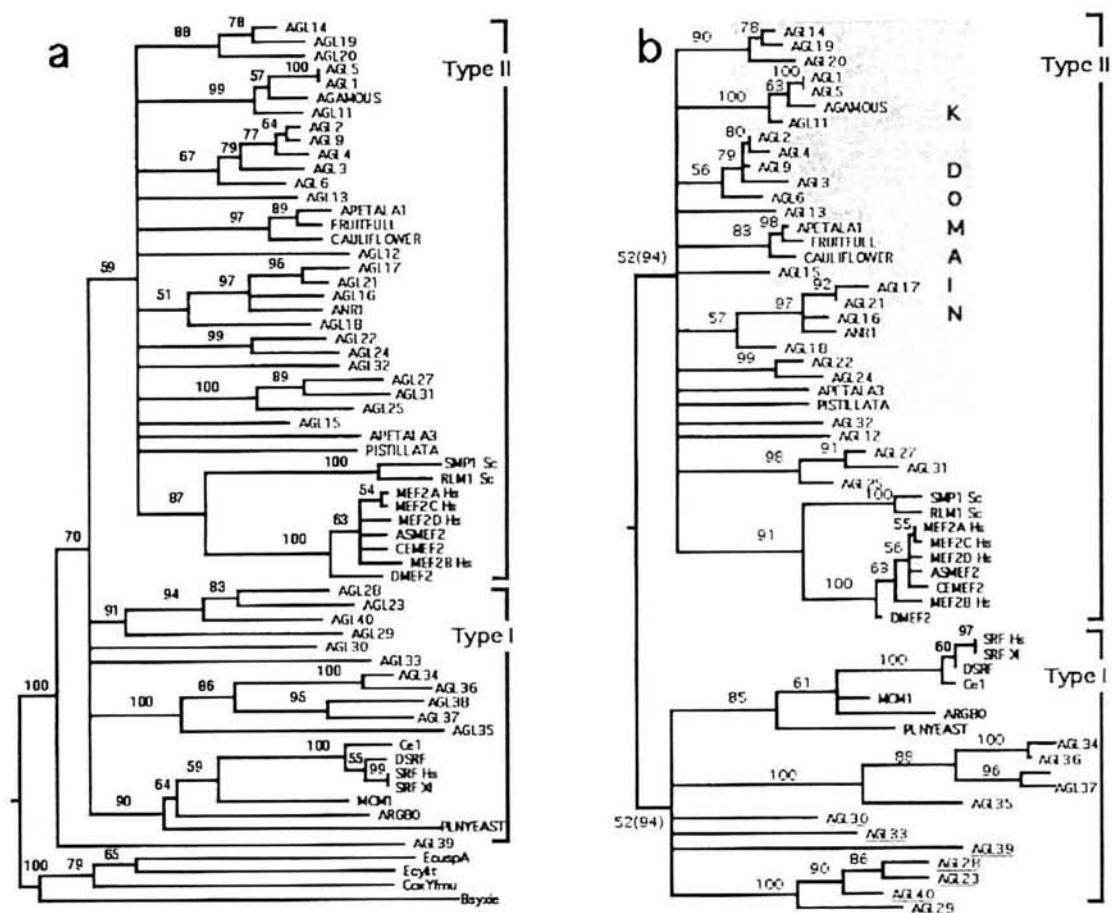


Fig. 3. Phylogeny of the eukaryotic MADS-box gene family. Animal and fungal sequences (*H. sapiens*: MEF2A_Hs, MEF2C_Hs, MEF2D_Hs, MEF2B_Hs, SRF_Hs; *X. laevis*: SRF_Xl; *C. elegans*: CEMEF2, *H. roretzi*: ASMEF2; *D. melanogaster*: DMEF2; DSRF; *S. cerevisiae*: SMP1_Sc, RLM1_Sc, MCM1, ARG80; *S. pombe*: PLNYEAST) are red; plant sequences (all from *A. thaliana*) are green; bacterial USP family sequences (*E. coli*: EcuspA, Ecyiit; *Coxiella burnetii*: CoxYfmu; *B. subtilis*: Bsyxie) (23) are blue. Type I (SRF-like) and Type II (MEF2-like) lineages are indicated by blue and pink brackets, respectively. (a) The NJ tree rooted with the bacterial USP family (see ref. 23) is shown in a, and the NJ tree rooted by minimizing the reconciliation cost (see *Materials and Methods*) is shown in b. Branch lengths are proportional to the number of amino acid substitutions. Bootstrap values shown on branches; in b, values in parentheses correspond to analyses done without the underlined sequences. Branches with bootstrap values <50% are collapsed. Sequences within purple square are those for which a coiled-coil structure downstream of the MADS-domain (K domain) was predicted.

tion between Type I and Type II sequences. This possibility is suggested because they share some of the synapomorphies that define each of the two lineages (see Fig. 2). The fact that these sequences group in a clearly monophyletic clade suggests an ancient recombination event that would have been followed by several duplications. To unambiguously resolve the origin and phylogenetic position of these genes, more information is required.

In an effort to explore further the monophyly of the Type I groups that we propose, we did MP and NJ phylogenetic analyses of this clade by using only one sequence of the MEF2-like sequences as outgroup (not shown). In these analyses, the plant AGL34-like, plus AGL30 and AGL33, plus the animal SRF-like sequences, form a well supported (bootstrap = 63%) monophyletic group, and AGL23-like and AGL39 sequences group in a clade sister to that formed by the former sequences. Both of these clades form a monophyletic lineage with 76% of bootstrap support.

The results presented here imply that features shared by proteins within the MEF2-like and SRF-like clades were present in the ancestral eukaryotes and have remained practically unchanged during the evolution of animal, fungal, and plant

lineages. The Type II MADS-domain sequences share some conserved amino acids that are found in none of the Type I MADS domains (synapomorphies; see Fig. 2). In contrast, the Type I MADS have only one synapomorphy that defines this clade and some that are shared by all but one or a few sequences. This suggests that there has been a stronger functional constraint within the Type II than the Type I MADS-domain lineages. Type I MADS domains are conserved within animals and within plants, but they differ between these two species' lineages. MADS domains from yeast from both Type I and Type II lineages are the most divergent ones.

It will be interesting to determine whether the plant Type I MADS-box sequences represent expressed genes or are instead pseudogenes. But the fact that at least one of these sequences, AGL39, is represented as an EST clone (GenBank accession no. C99890), as well as the high conservation among AGL34-like sequences, suggests that these members are indeed expressed. Future studies should be devoted to characterizing functionally these genes in *Arabidopsis*.

The conserved MADS-domain motifs within each lineage may serve as the basis of the common functional properties of all proteins within the Type I and Type II clades. Indeed, *in vitro*

DNA-binding assays revealed that chimeric proteins with either the SRF or MEF2A amino-terminal region of the MADS domain and the rest of the AP1, AP3, PI, and AG plant proteins, acquired the respective and distinct DNA-binding specificity of SRF or MEF2A. However, *in vivo* assays did not distinguish between chimeric and full-length wild-type proteins' functions. Both results put together suggest that DNA-binding specificity, which must underlie functional specificity of MADS-domain proteins, is determined not solely by sequences within the MADS-domain but also by sequences within other domains that may affect dimerization with protein partners (30).

Additional *in vivo* experiments show that although chimeric genes with the amino terminus of either the SRF or MEF2A MADS-domain and the rest of AP1 may rescue *apl-1* mutant plants when expressed under the wild-type *AP1* promoter, the chimera with the MEF2A MADS-domain amino terminus (i.e., within-lineage chimera) rescued mutant phenotypes more effectively than those harboring crosslineage constructs (i.e., from SRF; ref. 31). Our phylogenetic results support, as suggested by these functional analyses, that differences between TypeII and TypeI MADS domains have a role in defining function. Indeed, ectopic expression experiments of chimeric proteins suggest that the MADS and I domains define functional specificities of APETALA1 and AGAMOUS (32, 33), both TypeII (MEF2-like) plant members. However, the conservation of MADS-domain sequences within each lineage and additional functional studies (see below) also suggest that domains outside the MADS domain are important for functional specificity. The K domain, typical of previously characterized plant TypeII proteins, is one such domain.

Evolution of the Plant K Domain. The K domain is an ≈ 70 -aa domain located downstream of the DNA-binding MADS domain, typically spanning positions 110 to 180 of plant MADS proteins. It has a regular spacing of hydrophobic amino acids, and it is assumed to adopt a coiled-coil structure (see Fig. 1). This structural motif has been described for the great majority of previously identified plant MADS-domain proteins (4). To investigate the origin and evolution of the K domain, we used protein-structure programs to predict whether the AGL34 and AGL23 clade members, as well as the other plant and animal MADS-domain sequences analyzed, contain a K domain. In Fig. 3b, we boxed the sequences with a predicted coiled-coil structure downstream of the MADS domain.

Coiled-coil structures were not predicted for any of the animal sequences, any of the plant AGL34 or AGL23-like, or for AGL30, AGL33, and AGL39. These sequences also lack any significant sequence similarity to other plant MADS-domain sequences outside of the MADS domain. Interestingly, whereas protein-structure prediction programs clearly identify a coiled-coil domain for most plant members of the TypeII lineage (MEF2-like), they fail to predict such a structure for a few members of this group (the AGL25-like and AGL12) that seem to lack some of the conserved hydrophobic amino acids. This result suggests that the absent amino acids might be critical for the formation of the coiled-coil structure. Both methods used here have been reported to identify positively all of the sequences that form coiled coils in Protein Data Bank structures containing this type of helical structure (27). Thus, the coiled-coil predictions presented in this work have a high level of reliability (>95%), well above standard secondary structure prediction methods.

Animal SRF- and MEF2-like proteins contain additional conserved regions, referred to as SAM and MEF2 domains (2). These and the K domain could be the regions involved in the functional divergence among members of each MADS-domain lineage. Ectopic expression experiments of chimeric proteins suggest that functional specificities of APETALA3 and PISTILL-

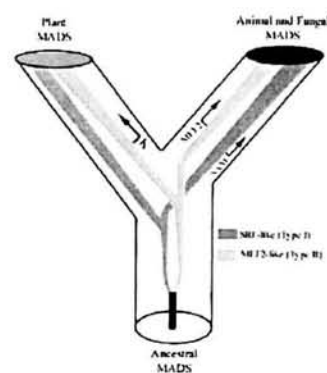


Fig. 4. Model for the evolution of the MADS-box gene family in eukaryotes. At least one duplication of the ancestral MADS-box gene is postulated to have occurred before the divergence of plants and animals. The K domain was probably added to the plant TypeII (MEF2-like) lineage. Similarly, animal MADS-domain proteins evolved specific domains (SAM and MEF2) in SRF-like and MEF2-like lineages, respectively. Pink, TypeI (SRF-like) lineage; blue, TypeII (MEF2-like) lineage.

LATA MADS-domain proteins in organ determination rely on the I and K domains of these genes (31, 32). Recent experiments for two plant MADS-domain proteins (*APETALA1* and *CAULIFLOWER*) suggest that differences between the K domains of these two recently duplicated genes explain at least part of the functional differences between these paralogous loci (E.R.A.-B. and M.F.Y., unpublished results).

Evolution of MADS-Domain Proteins in Eukaryotes: A Synthesis. The results described here suggest a hypothetical scenario for the evolution of the MADS-box gene family in eukaryotes (Fig. 4). From our analyses, it appears that at least one ancestral MADS-box gene duplicated in the common ancestor of the major eukaryotic kingdoms more than a billion years ago to give rise to the distinct TypeI (SRF-like) and TypeII (MEF2-like) lineages found in plants, fungi, and animals today. In yeast and *C. elegans* genomes, MADS-box sequences of both TypeI and TypeII have been found (several of each in yeast and one of each in *C. elegans*). These results support our proposition that eukaryotic MADS-box sequences can be assigned to either of two main lineages that are both present at least in fungi and animals. The *Arabidopsis* genome will be sequenced to completion soon, and we will then be able to test unambiguously the presence of these and additional lineages in plants. Phylogenetic analyses that include MADS domains from basal eukaryotes and TypeI sequences from other plants will help confirm the uniqueness of the ancestral duplication and the monophyly of the TypeI clade.

The evolution of additional domains beyond the MADS domain could have occurred independently along the animal and plant lineages after their divergence from each other, as suggested in our model (Fig. 4), or these could have been present in the ancestral MADS-box genes and then lost along different lineages. In plants, the K domain evolved within the TypeII (MEF2-like) lineage but not the TypeI (SRF-like) lineage. Because most of the TypeII class of plant MADS-box genes are predicted to encode a K domain, this plant-specific domain probably evolved before the extensive duplications that generated this particular lineage. Interestingly, some of the recently cloned MADS-box genes from ferns (33) are predicted to contain K domains (data not shown), indicating that this domain was present at least 395 million years ago in the common ancestors of ferns and seed plants.

We can use parsimony to argue that the K domain originated after the duplication that led to the MEF2- and SRF-like animal

MADS-box genes. However, based on the phylogeny of Fig. 3b, we cannot distinguish whether it evolved along the plant lineage after it diverged from the animal one, or whether it was present in the ancestral TypeII-like gene and then lost in animal and some plant lineages. A recent phylogenetic analysis of the M, I, and K domains of all plant protein sequences, (E.R.A.-B., S.L., S.P., S.G., C.B., G.D., and M.Y., unpublished work) suggests that AGL12 and the AGL25-like sequences are basal to the rest of the *Arabidopsis* TypeII AGLs. This result supports the hypothesis that the K domain evolved along the plant lineage after it diverged from animals and fungi (Fig. 4). Identification of MADS-box genes within the most basal extant green plant lineages (including green algae and the bryophytes) and in one of the extant common ancestors of plants and animals (e.g., *Euglena*) should provide experimental tests for the hypotheses postulated in this model of MADS-box gene family evolution. Animal SRF- and MEF2-like domains (see Figs. 1 and 4) may have evolved within animal lineages (as suggested in Fig. 4), or they could have been present also before the divergence of plants and animals and subsequently lost and replaced in plants.

MADS-box genes probably played key roles in the early evolution of flowering plants and in plant evolution in general, perhaps analogous to the roles played by homeobox genes in the evolution of animal form (34, 35). This scenario is suggested by the fact that MADS-box gene mutations, as those of homeobox genes in animals, also produce homeotic conversions in flowers, suggesting that they occupy similar places in the regulatory networks that control development (36). Like homeobox genes, MADS-box genes are also highly conserved among distantly related plants, and orthologous genes form monophyletic clades (6–9). To test the long-suspected parallel between the molecular evolution of the MADS-box gene family and the evolution of

plant form, a polarized gene phylogeny is necessary. We have proposed here a hypothesis for the evolutionary history of the MADS-domain protein family, including the nearly complete *Arabidopsis* MADS-box sequence complement, which suggests that eukaryotic MADS-box sequences can be assigned to two main lineages and locates the root of the whole family between them. These analyses may be used to guide the search for MADS-box sequences in basal eukaryotes and the assignment of newly cloned genes from other plant species to one of the clades proposed in this study. Further phylogenetic and population genetic studies (e.g., ref. 37) as well as functional analyses of the MADS-box family and other important transcriptional regulators should lead to a better understanding of the molecular evolution of developmental mechanisms. These mechanisms underlie the morphological evolution of plants and animals, the understanding of which is still elusive to evolutionary biologists.

We thank C. Ferrándiz, W. Crosby, C. Gustafson-Brown, and S. Rounsley for providing unpublished sequences. Special thanks to A. Chaos and F. Vergara for illuminating phylogenetic conversations and for help during figure preparation. Two anonymous reviewers put important effort into helping us improve this paper. Many thanks to A. Cortés for help in various tasks and to R. Salas for help in Fig. 4. C. Ferrándiz, S. Kempin, M. Ng, and A. Sessions provided useful suggestions. This work was supported by a grant from the National Science Foundation to M.F.Y. and a CONACYT (Consejo Nacional de Ciencia y Tecnología, México) grant to E.R.A.-B. Also, E.R.A.-B. was a Pew Foundation Fellow during the completion of this work. S.P. had a long-term postdoctoral fellowship from the Human Frontiers Science Program Organization, S.L. had a Lucille P. Markey predoctoral fellowship, and C.B. and L.M.C. had a Ph.D. scholarship from CONACYT and Universidad Nacional Autónoma de México.

- Doebley, J. & Lukens, L. (1998) *Plant Cell* **10**, 1075–1082.
- Shore, P. & Sharrocks, A. D. (1995) *Eur. J. Biochem.* **229**, 1–13.
- Coen, E. S. & Meyerowitz, E. M. (1991) *Nature (London)* **353**, 31–37.
- Riechmann, J. L. & Meyerowitz, E. M. (1997) *J. Biol. Chem.* **378**, 1079–1101.
- Liljegren, S. J., Ferrándiz, C., Alvarez-Buylla, E. R., Pelaz, S. & Yanofsky, M. F. (1998) *Flowering Newsletter* **25**, 9–19.
- Theissen, G., Kim, J. T. & Saedler, H. (1996) *J. Mol. Evol.* **43**, 484–516.
- Doyle, J. J. (1994) *Syst. Biol.* **43**, 307–328.
- Purugganan, M. D., Rounsley, S. D., Schmidt, R. J. & Yanofsky, M. F. (1995) *Genetics* **140**, 345–356.
- Purugganan, M. D. (1997) *J. Mol. Evol.* **45**, 392–396.
- Yanofsky, M. F., Ma, H., Bowman, J. L., Drews, G. N., Feldmann, K. A. & Meyerowitz, E. M. (1990) *Nature (London)* **346**, 35–39.
- Jack, T., Brockman, L. L. & Meyerowitz, E. M. (1992) *Cell* **68**, 683–697.
- Goto, K. & Meyerowitz, E. M. (1994) *Genes Dev.* **8**, 1548–1560.
- Ma, H., Yanofsky, M. F. & Meyerowitz, E. M. (1991) *Genes Dev.* **5**, 484–495.
- Mandel, M. A., Gustafson-Brown, C., Savidge, B. & Yanofsky, M. F. (1992) *Nature (London)* **360**, 273–277.
- Mandel, M. A. & Yanofsky, M. F. (1995) *Plant Cell* **7**, 1763–1771.
- Mandel, M. A. & Yanofsky, M. F. (1998) *Sexual Plant Reprod.* **11**, 22–28.
- Kempin, S. A., Savidge, B. & Yanofsky, M. F. (1995) *Science* **267**, 522–525.
- Rounsley, S. D., Ditta, G. S. & Yanofsky, M. F. (1995) *Plant Cell* **7**, 1259–1269.
- Zhang, H. & Forde, B. G. (1998) *Science* **279**, 407–409.
- Burke, W. D., Eickbush, D. G., Xiong, Y., Jacubczak, J. & Eickbush, T. H. (1993) *Mol. Biol. Evol.* **10**, 163–185.
- Hillis, D. M. & Bull, J. J. (1993) *Syst. Biol.* **42**, 182–192.
- Page, R. D. M. (1996) *CABIOS* **12**, 357–358.
- Mushegian, A. R. & Koonin, E. V. (1996) *Genetics* **144**, 817–828.
- Page, R. D. M. & Charleston, M. A. (1997) *Mol. Phyl. Evol.* **7**, 231–240.
- Page, R. D. M. & Holmes, E. C. (1998) *Molecular Evolution. A Phylogenetic Approach* (Blackwell Scientific, Oxford).
- Baldauf, S. L. & Palmer, J. D. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 11558–11562.
- Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M. & Kim, P. S. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 8259–8263.
- Wolf, E., Kim, P. S. & Berger, B. (1997) *Protein Sci.* **6**, 1179–1189.
- Stultz, C. M., White, J. V. & Smith, T. F. (1993) *Protein Sci.* **2**, 305–314.
- Riechmann, J. L. & Meyerowitz, E. M. (1997) *Mol. Biol. Cell* **8**, 1243–1259.
- Krizek, B. A., Riechmann, J. L. & Meyerowitz, E. M. (1999) *Sex Plant Reprod.* **12**, 14–26.
- Krizek, B. A. & Meyerowitz, E. M. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 4063–4070.
- Münster, T., Pahnke, J., DiRosa, A., Kim, J., Martin, W., Saedler, H. & Theissen, G. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 2415–2420.
- Raff, R. A. (1996) *The Shape of Life: Genes, Development, and the Evolution of Animal Form* (Univ. of Chicago Press, Chicago).
- Carroll, S. B. (1995) *Nature (London)* **376**, 479–485.
- Mendoza, L. & Alvarez-Buylla, E. R. (1998) *J. Theor. Biol.* **193**, 307–319.
- Purugganan, M. D. & Suddith, J. I. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 8130–8134.

(3) Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny. Martínez-Castilla, L. P. y Alvarez-Buylla, E. R. 2003. *Proceedings of the National Academy of Sciences, USA*. 100: 13407-13412.

NOTA: Las secciones de material suplementario que se mencionan en el artículo se encuentran al final de este apéndice.

Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny

León Patricio Martínez-Castilla and Elena R. Alvarez-Buylla*

Laboratorio de Genética Molecular, Desarrollo y Evolución de Plantas, Instituto de Ecología, National Autonomous University of Mexico, Ap Postal 70-275, Mexico D.F., 04510, Mexico

Communicated by José Sarukhán, National Autonomous University of Mexico, Mexico D.F., Mexico, September 11, 2003 (received for review May 31, 2003)

Gene duplication is a substrate of evolution. However, the relative importance of positive selection versus relaxation of constraints in the functional divergence of gene copies is still under debate. Plant MADS-box genes encode transcriptional regulators key in various aspects of development and have undergone extensive duplications to form a large family. We recovered 104 MADS sequences from the *Arabidopsis* genome. Bayesian phylogenetic trees recover type II lineage as a monophyletic group and resolve a branching sequence of monophyletic groups within this lineage. The type I lineage is comprised of several divergent groups. However, contrasting gene structure and patterns of chromosomal distribution between type I and II sequences suggest that they had different evolutionary histories and support the placement of the root of the gene family between these two groups. Site-specific and site-branch analyses of positive Darwinian selection (PDS) suggest that different selection regimes could have affected the evolution of these lineages. We found evidence for PDS along the branch leading to flowering time genes that have a direct impact on plant fitness. Sites with high probabilities of having been under PDS were found in the MADS and K domains, suggesting that these played important roles in the acquisition of novel functions during MADS-box diversification. Detected sites are targets for further experimental analyses. We argue that adaptive changes in MADS-domain protein sequences have been important for their functional divergence, suggesting that changes within coding regions of transcriptional regulators have influenced phenotypic evolution of plants.

positive Darwinian selection | duplication | functional divergence | *Arabidopsis thaliana* | development

Gene duplication provides a substrate for evolution, and understanding the fate of duplicates is fundamental to clarifying mechanisms of genetic redundancy and the link between gene family diversification and phenotypic evolution (1). Several empirical studies have evaluated the roles of duplication in adaptation and diversification (2), but the evolutionary forces at play during functional divergence of duplicates are still under debate (3, 4). Genomic studies are revealing that eukaryotes harbor large families of genes that have arisen during evolution through duplication and have persisted for longer periods of time than expected by classical models (5). Models that incorporate positive selection (6, 7) provide alternative explanations for the persistence of duplicates.

Empirical studies to test models on the fate of duplicates and the evolutionary forces driving their functional divergence will need complete and resolved gene family phylogenies. Several studies suggest that positive Darwinian selection (PDS) might have been important in protein evolution. However, most previous studies have involved few members of a gene family from various species (8, 9). In this article, we annotate, align, and analyze the complete MADS-box gene family of the plant model system *Arabidopsis thaliana* and provide resolved phylogenies as a basis to infer the role of PDS in protein evolution in this gene family.

The detection of an excess in the ratio of the rate of nonsynonymous (dN) over the synonymous (dS) substitutions (that is $dN/dS > 1$; dN/dS is also denoted ω) is a nonambiguous indicator of PDS at the coding sequence level. Early studies estimated this ratio as an average over all codon sites within complete or partial sequence stretches and over the entire evolutionary time that separates the sequences compared. This method appears to be conservative because many sites might be under purifying selection because of functional constraint (10). However, in adaptive evolution in developmental regulatory loci, such as the MADS (11) PDS most likely occurs along particular lineages and at specific sites. In such cases, average dN/dS ratios over time and sites might not be significantly greater than 1, even if PDS has occurred.

MADS-box genes are present in plants, animals, and fungi, and previous studies suggested the existence of two main monophyletic lineages (type I and II) among all eukaryotes that probably derived from at least one duplication event before the divergence of plants and animals (14). The trees presented here recover *Arabidopsis* type II genes as a strongly supported monophyletic lineage, and the type I genes seem to be monophyletic but comprise several divergent sublineages.

MADS-box genes encode transcriptional regulators with diverse functions that could have been key during important events of plant diversification (12, 13). Hence, phylogenetic analyses of MADS-box genes are useful guides for studying their roles in plant evolution. Plant MADS-box gene phylogeny resolution, especially at its basal nodes, has been hindered by incomplete data and by limitations of inference methods (14, 15). Here we show resolved gene phylogenies of the *Arabidopsis* MADS-box genes.

More than half of the *Arabidopsis* MADS-box sequences are type I and only share with type II the MADS-box (14). All but one (16) functionally characterized plant MADS-box genes are type II and encode the three floral homeotic functions of the flower development ABC model (17–19). They also encode regulators of flower initiation, flower meristem identity, and various aspects of ovule, fruit, leaf, and root development (11, 20–23). All characterized plant type II MADS-box genes encode proteins that share a stereotypical MIKC structure, with highly conserved MADS and K domains that are putative DNA-binding and protein–protein interaction domains, respectively, and less conserved I and COOH regions.

We show here that sequences of type I and II have contrasting gene structure and chromosome distribution, supporting the idea that these two lineages had different evolutionary histories with a contrasting role of PDS. These contrasting histories also support placing the root of the family tree between the two lineages. Indeed,

Abbreviations: PDS, positive Darwinian selection; LRT, likelihood ratio test; AGL, agamous-like.

*To whom correspondence should be addressed. E-mail: ealvarez@miranda.ecologia.unam.mx.

© 2003 by The National Academy of Sciences of the USA

we found a significant role of PDS at fixing specific residues within the MADS-domain after different duplications in the type I lineage, but in the type II lineage we found evidence of PDS only with branch-site models along specific lineages. We addressed whether PDS played a significant role during the evolution of genes that regulate the transition to flowering, a trait that is clearly linked to plant fitness. Our findings identify target proteins and residues for future functional analyses and suggest that changes in coding sequences of transcriptional regulators, and not only in their regulatory regions, played important roles during phenotypic evolution.

Materials and Methods

Sequences and Alignment. To detect putative MADS-box genes in the *Arabidopsis* genome, two TBLASTN searches were performed on the complete *Arabidopsis* database (Table 3, which is published as supporting information on the PNAS web site). Sequences were assigned to either type I or II based on exon number and on careful comparisons of their MADS boxes (14, 24). Type I and II were then aligned separately with CLUSTALW (25) launched from BIOEDIT (26) and hand-corrected by using published alignments as guides (refs. 27 and 28; details can be found in *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site).

Phylogenetic Reconstruction. Bayesian phylogenetic analyses were performed on MRBAYES 2.01 (29). All searches were started from a random tree, on four different Markov chains for 2,500,000 generations and saving every 100th tree. At convergence ($\approx 10,000$ generations), the first 15,000 trees were discarded and a consensus was built. Posterior Bayesian probabilities were used to evaluate branch support. According to the recommendations of Foster (www.bioinf.org/molys/data/like.pdf), we used the GTR model with a substitution rate that varies in an intracodon position-specific manner (GTR + SS).

Statistical Tests for Positive Selection. We applied the approach of Yang and coworkers (9, 30) to test for positive selection. First, we ran a test for the existence of sites with dN/dS ratios > 1 by using a likelihood ratio test (LRT) to compare a model that does not allow for sites with dN/dS > 1 to a model that does. If the LRT was statistically significant, then we identified the sites that were under positive selection. We calculated the posterior probability (PP) that a site was drawn from a given dN/dS class. Sites with PP > 0.5 are reported but we focus on those with a PP > 0.95 . The program HYPHY 0.901b (S. L. Kosakovsky-Pond and S. V. Muse, www.hyphy.org) was used. Models tested were M3, M2, and M8 vs. null models M0, M1, and M7, respectively. We used the codon substitution model of Goldman and Yang (31) and 10 classes in the gamma distribution of M7 and M8. To avoid false positives, sites detected by models M3, M2, and M8 were considered as bona fide results only if the same sites were detected with at least two of these models and in both cases the LRT result was significant (32). To further minimize false positives, we performed all analyses on unambiguous and compact alignments (available on request). Thus, our conclusions on the role of PDS are based on very conservative analyses.

Additionally, we performed the branch-site analyses of Yang and Nielsen by using model MB (PAML 3.13, http://abacus.gene.ucl.ac.uk/software/paml.html; refs. 33 and 34) at the basal branch of two clades of flowering-time type II genes [*FLC*- and *SVP*-like genes (35–37)]. We hypothesized that functional change would be important precisely at the origin of these clades that evolved a distinct function (details of procedures can be found in refs. 31, 33, and 34). We also performed site-branch analyses for the branch that leads to type I *PHERES1* gene.

Gene conversion and concerted evolution may violate the assumptions of site-specific positive selection models. We used

GENECONV 1.81 (www.math.wustl.edu/~sawyer/geneconv/index.html) and MEGA 2.1 (38) on the alignments tested for PDS.

Results

Annotation, Nomenclature, Gene Structure, and Duplications. The list of 104 MADS-box sequences found in the *Arabidopsis* genome database (www.ncbi.nlm.nih.gov/blast/Genome/ara.html), given agamous-like (AGL) number, accession number, chromosome location, intron–exon structure, and type of duplication are shown in Table 3.

Type I and II genes have contrasting gene structure and chromosome location. Type I genes have always one or two exons, whereas type II genes have more than five, and typically six to eight, suggesting that these two types of genes have different predisposition to gain or loss of introns or, alternatively, a difference in exon shuffling in the building of both types of genes, perhaps since their origin. All type II genes have a clear MIKC structure, except AGL33 with a very short transcript and AGL30-related sequences that do not have a clear IKC structure but share conserved motifs in their MADS-boxes with the rest of the type II genes. Also, a coiled-coil domain similar to the K domain is inferred for at least AGL104 (data not shown).

We used chromosome map locations to make qualitative inferences on past duplication events during MADS-box gene family evolution. The distribution of type I sequences among the five chromosomes is distinct to that of type II genes. Whereas the former are concentrated in chromosome I and V ($\chi^2 = 26.77$; $P < 0.001$ rejects uniform distribution among chromosomes correcting with chromosome size), type II are uniformly distributed ($\chi^2 = 2.62$; $P > 0.1$) among chromosomes. Also, most type I genes can be traced to intrachromosomal duplications, whereas approximately half of type II genes seem to have originated from interchromosomal duplications. Interestingly, Lynch and Conery (39) have found that recent duplications happened more frequently within than between two chromosomes, suggesting that most type I genes diverged more recently than type II genes (40). Moreover, a survey of the *A. thaliana* paralogous blocks database (http://wolfe.gen.tcd.ie/athal/dup) indicates that there are more duplicates from the type II group that seem to have persisted than those from type I. Eighteen out of 26 type II sequences that are found in a nonspurious duplicated region had a close paralog in a sister region, whereas among type I genes only 2 out of 22 did. Among type II genes found in nonspurious duplicated blocks, 14 are found in interchromosomal duplications and 12 in intrachromosomal duplications, whereas among type I genes the corresponding numbers are 15 and 7. But when we consider gene pairs of terminal clades of the trees, 13 out of 18 type I gene pairs involve intrachromosomal duplications, whereas in the type II we found this to be true of 6 out of 15 gene pairs (Fig. 1).

MADS-Box Gene Family Phylogeny. To corroborate the monophyletic origin of the two lineages that we had previously proposed (14), we obtained trees that included 103 of the MADS-box sequences found in the *A. thaliana* genome (Fig. 1). The global phylogeny recovers the two lineages (types I and II) of MADS-box sequences as two monophyletic groups with both alignments used (see supporting information) if the tree root is placed between type I and II genes. However, type I genes are more divergent among them than type II genes (Tables 4 and 5, which are published as supporting information on the PNAS web site). Nonetheless conserved motifs after the MADS suggest that type I sequences are not pseudogenes, an idea supported by the recent characterization of *PHERES1* (16).

AGL30 had been incorrectly assigned to type I. However, this and related genes seem to be divergent type II. This is supported by their affiliation to type II-like moss genes that bear K domains as well as by their exon number (27, 41), and by conserved MADS-box motifs with respect to other type II genes. In our global tree (Fig. 1), these genes are resolved in a different position to that in the type

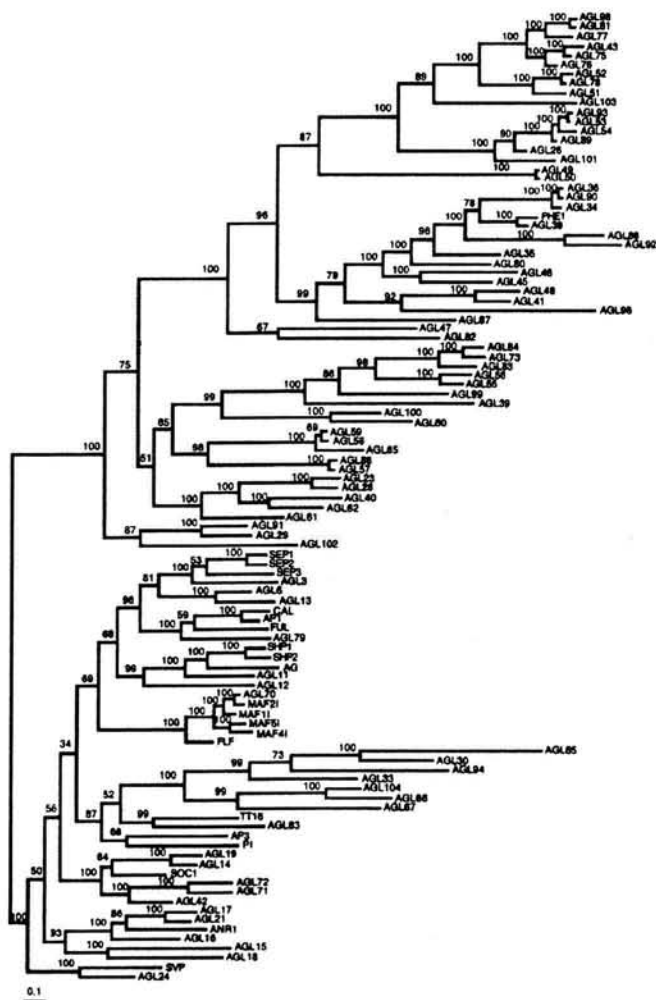


Fig. 1. *A. thaliana* MADS-box gene family Bayesian phylogeny. Numbers above or below branches represent Bayesian posterior probabilities of finding a given clade. Branch lengths are proportional to number of nucleotide substitutions. AGL105 was excluded.

II tree (Fig. 2*b*). Nonetheless, this remaining ambiguity does not affect PDS analyses. We ran PDS analyses for MIKC genes with an alternative type II topology similar to that in Fig. 1 and results recover the same sites with high PP as those obtained with the topology shown in Fig. 2*b* (data available on request).

Given the ambiguity of the alignment of sequences after the MADS-boxes of both lineages, we resolved the internal phylogenetic relationships of each lineage separately. For the type I lineage tree, we used type II sequences as outgroups and vice versa for type II lineage. In both cases, tree topologies were very similar if regions downstream of the MADS-boxes of outgroup sequences were assumed to be homologous to those of ingroup sequences or were displaced. In contrast to previously published phylogenies, the trees shown here for both lineages resolve the branching sequence of the monophyletic groups (Fig. 2).

In the type I lineage, several strongly supported monophyletic clades are resolved and these are confirmed in trees with various outgroups (data not shown). It is noteworthy that no type I sequences bear the IKC region typical of lineage II and that several of the previously identified (14) primitive amino acids within the putative MADS-domain of this lineage are found in most available sequences. Two main monophyletic groups are distinguished within the type I, suggesting at least one ancestral duplication within this lineage. DNA sequences beyond the MADS-box in the *AGL23*-like sublineage are conserved within each small monophyletic group

resolved within this sublineage but are very divergent among groups (Fig. 2*a* and Table 4). Within the *AGL26*-like genes, two groups are resolved [*AGL26*-like genes themselves and *PHERES1*-like genes that includes the first type I gene functionally characterized (16)] and within each a very high degree of conservation is found in the putative domains beyond the MADS (Fig. 2*a* and Table 4). Sisters to these groups, *AGL47* and *AGL82*, are resolved but have divergent putative domains beyond the MADS.

In the type II lineage, which includes all of the MIKC genes functionally characterized up to now, several clades are resolved and all are well supported (Fig. 2*b*). It is noteworthy that the *AGAMOUS* clade has *AGL12*, which had been reported as root-specific (42), as its sister gene. Another important finding is the strong association of *AGL79*, expressed in roots (data not shown), with *API*, *CAL*, and *FUL*, which are well characterized flower development genes (43). Therefore, this tree suggests that not all monophyletic groups resolved include genes with similar expression patterns and functions as previously thought (15, 21), but formal and robust inferences on evolution of MADS-box gene expression and function will have to await more experimental data and the inclusion of genes from additional taxa.

Positive Selection in MADS-Box Gene Evolution. We compared Models M0 and M3 to evaluate whether there had been dN/dS ratio variation among codon positions below each node of type I and II trees from Fig. 2 (Tables 1 and 2). We found rate variation at deep, intermediate, and recent duplications of both type I and II lineages (data available on request).

Secondly, we applied the LRT to compare data fit to models M1 vs. M2 and M7 vs. M8 to address whether PDS promoted divergence of MADS-box genes below nodes and whether the action of selection has been heterogeneous among protein domains codified by these genes, using the trees of Fig. 2 (Tables 1 and 2 and Fig. 3). Below many of the deep nodes of type II tree, model M3 and at least one of M2 or M8 had significant LRT results (Fig. 2*b*). However, none of the sites with high PP were detected by more than one model comparison with significant LRT results (data available on request). In contrast, a similar analysis for type I lineage reveals several nodes below which specific sites appear to have been under PDS (Tables 1 and 2). For instance, in nodes AH and AL, positions 72–74 appear to have been under PDS with high PP, with position 72 showing five different amino acids for the seven sequences involved (Fig. 2*a* and Table 4). Positions 73 and 74 are part of the otherwise highly conserved “RQVTF” motif, and in the human serum response factor, position 72 has been shown to be involved in DNA contact (44). Below nodes AH and AR, position 82 was also found to be under PDS in two of the model pairs compared, although model M3 collapsed to only two rate classes. At this position, amino acid diversity is very high. For example, 10 different residues can be found for 20 sequences analyzed below node AR. In contrast, the homologous position of type II genes (position 26 in Table 4) shows only six different amino acids for 45 sequences.

Strong evidence for positive selection in type I evolution was also found in less variable positions. This is the case for position 123 of node AG and position 58 of nodes AP and AQ. In position 123, there are only two different amino acids, although their distribution suggests that this site mutated twice during the history of descendants of node AG. In position 58, there are only five variable sites out of 14 sequences compared.

In the above analyses, PDS is detected at individual sites only if the average dN rate across lineages is higher than the average dS rate. This is observed mainly when recurrent positive selection occurs. However, PDS may change a few key residues of a protein but only at particular moments during its evolutionary history. In the latter case, detecting a significantly elevated dN would be hard if an average across-lineages estimate is considered. Thus, we applied the branch-site model of Yang and Nielsen (33) to test for PDS affecting individual sites along the branches leading to the

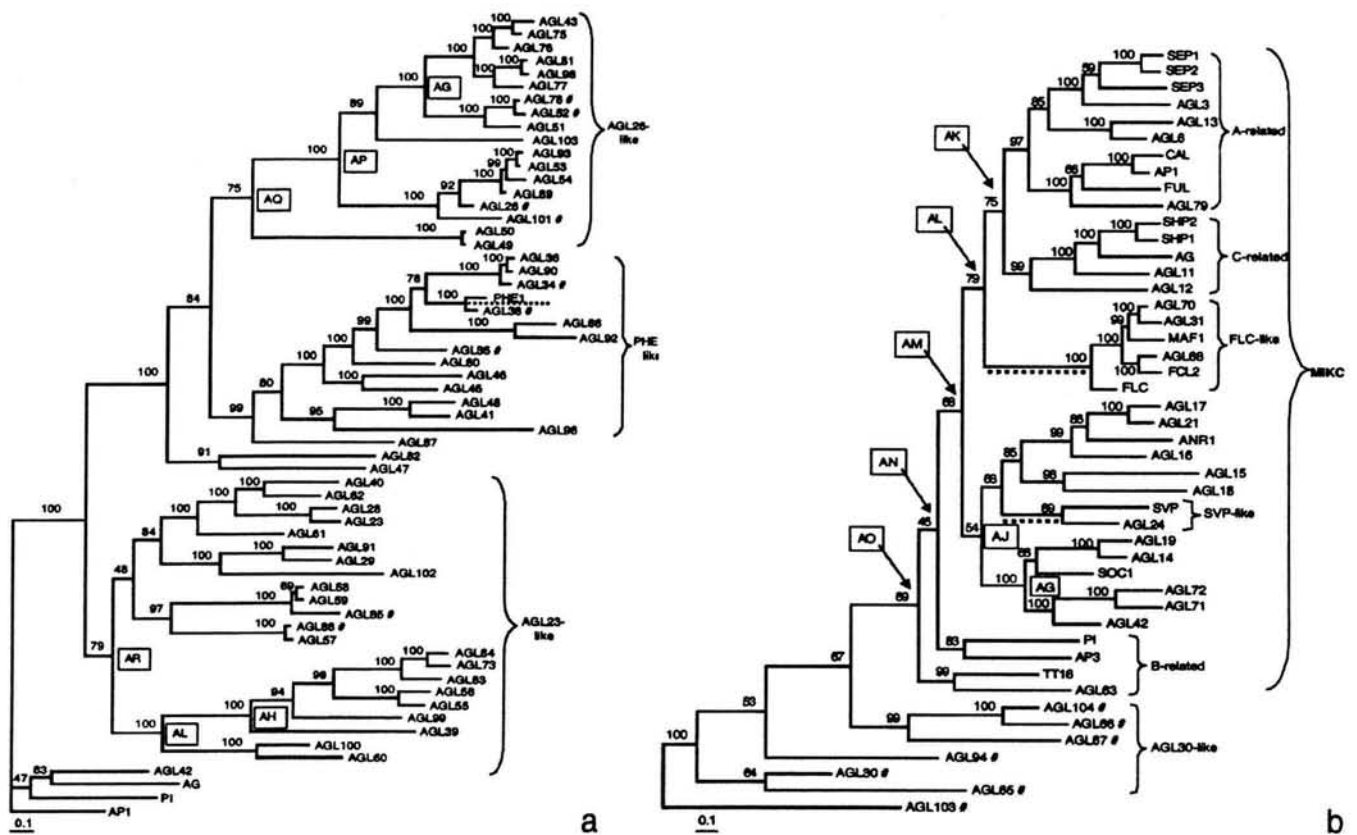


Fig. 2. Type I and type II *A. thaliana* MADS-box Bayesian phylogenies. (a) Type I tree polarized with type II sequences. (b) Type II tree polarized with type I sequences. Numbers above or below branches represent posterior probabilities. Branch lengths are proportional to the number of nucleotide substitutions. Boxed letters identify clades in which site-specific tests of positive selection yielded statistically significant LRT results for at least two model comparisons and in which at least one of the models detected sites under PDS with $PP > 0.90$. Branches underlined with a broken line identify cases in which the branch-site analyses yielded significant PDS results. #, Excluded from the site-specific analyses.

flowering-time *FLC*- and *SVP*-like genes in type II lineage and along the branch leading to the only functionally characterized type I protein (*PHERES1*).

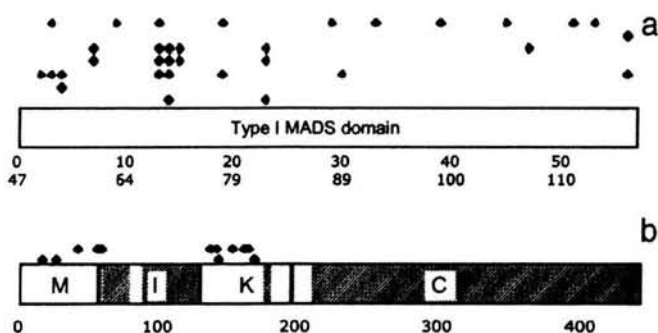


Fig. 3. Schematic representation of the distribution of sites under PDS in type I (a) and type II (b) sequences. MADS, I, K, and COOH domains are indicated. In a, the upper row corresponds to branch-site analysis along the branch leading to *PHERES1*, and the rest of the rows correspond, from top to bottom, to the site analyses performed below the nodes marked with the boxed letters AG, AH, AL, AP, AQ, and AR in Fig. 2a. In b, the upper row corresponds to branch-site analysis along the *FLC* branch and the lower row corresponds to branch-site analysis along the *SVP* branch. In a, the upper scale corresponds to amino acid position along the MADS domain and the lower scale corresponds to amino acid position along our alignment. In b, the scale corresponds to amino acid position along our alignment. Sites with $PP > 0.70$ are included. Shaded regions in b were excluded from PDS to avoid false positives. All sites are listed in Tables 1 and 2.

Parameter estimates suggest that in the basal branch of the *FLC*-like lineage, four residues were fixed by PDS with $PP > 0.95$ (Tables 1 and 2 and Fig. 3). Two of the residues were found within the MADS and correspond to amino acids that in other MADS-domain proteins participate in interactions between subunits (44–46). For example, site 42 is homologous to a site that in the myocyte enhancer factor-2 has been shown to intervene in subunit folding (45). Another site with high PP is 154. This site is within the K-box and has been shown to intervene in AP3/PI dimerization and determines functional specificity in AP1 and AG (47).

PDS seems to have been important also along the branch leading to the *SVP*-like genes, and we found two sites that appear to have been fixed by PDS ($PP > 0.95$). Site 16 is within the MADS-domain, and its homologous position in *MEF2A* plays a role in DNA-protein complex stabilization (46). Position 144 is found within the K-domain. *SVP* and *AGL24* are the only type II proteins that have a lysine at that position.

Branch-site models also detected strong ($PP > 0.95$; see Tables 1 and 2) support for PDS along the branch leading to *PHE1* at sites 92 and 72, which have been reported (44) to be important for α - β folding, and for position 105, which is involved in both dimerization and α - β folding.

Our analyses suggest that gene conversion or concerted evolution has not been prevalent during MADS-box genes evolution and does not bias PDS inferences. The overall modified Nei-Gojobori (48) means of synonymous–nonsynonymous differences were relatively high: 0.18 and 0.09 (transition/transversion ratio of 1.72 and 1.7, respectively) for types I and II, respectively (49). Gene conversion does not seem to have played significant roles in MADS-box

Table 1. Sites under PDS in the *A. thaliana* MADS-box gene family: "Site-specific analyses"

Site-specific analyses (type I)	<i>n</i>	dN/dS (ω) under M0	$2\Delta\ell$ M3 vs. M0 (df LRT 3)	$2\Delta\ell$ M2 vs. M1 (df 2)	$2\Delta\ell$ M8 vs. M7 (df 2)	Parameter estimates (β and ω) under M8 $\beta(p,q)$	Positively selected sites under M3	Positively selected sites under M2	Positively selected sites under M8
Node AG	7	0.22	28.76***	10.10*	10.57*	$P_1 = 0.98, \omega = 8.44$ $\beta(0.87, 3.82)$	123	No rate classes with dN/dS > 1	123
Node AH	7	0.26	59.09***	11.75**	7.36*	$P_1 = 0.81, \omega = 1.96$ $\beta(0.87, 3.48)$	73 74 61 72 82 107 63	73 74 72 61 82 63 107	73 72 74 61 82 107 63 48 58 81 110
Node AL	9	0.25	90.35***	18.39***	4.55	$P_1 = 0.93, \omega = 1.24$ $\beta(0.80, 2.28)$	61 72 73 82 107 110 48 74 86 58 64 66 63 65 109	72 73 82 74 61 107 48 110	72 73 82 48 61 74 107 110 58
Node AP	12	0.25	130.47***	21.26***	7.59*	$P_1 = 0.75, \omega = 1.18$ $\beta(0.56, 3.46)$	58	58 48 57 78 89 72 123 73 81	48 57 58 72 73 78 89 123 59 65 81 47 60 66 92 74 100
Node AQ	14	0.22	116.71***	14.97***	7.73*	$P_1 = 0.97, \omega = 4.67$ $\beta(0.47, 1.52)$	58 89	58	58
Node AR	20	0.15	102.65***	30.18***	8.22*	$P_1 = 0.96, \omega = 1.12$ $\beta(1.64, 8.94)$	No rate classes with dN/dS > 1.	73 82 72	82 73

Each comparison has *n* sequences, dN/dS is average ratio over sites under a codon model with one ω . Proportion of the component of positively selected sites (P_1) and parameters *p* and *q* of the beta distribution $\beta(p,q)$ are given under M8. *, $P < 0.5$; **, $P < 0.005$; ***, $P < 0.001$; bold underlined, $PP \geq 0.99$ of being under positive selection; bold, $0.99 > PP \geq 0.95$; italics, $0.95 > PP \geq 0.90$; underlined, $0.90 > PP \geq 0.70$; normal, $0.70 > PP \geq 0.50$.

evolution because only three conversion events were detected among type II genes, but none of these included genes for which we infer PDS and no conversion events were detected among type I sequences.

Discussion

We presented an annotated list of 104 MADS-box sequences from the complete *A. thaliana* genome database. Our phylogenetic analyses provide a resolved evolutionary hypothesis for the *A. thaliana* MADS-box gene family. This will be a useful reference for establishing orthology relationships, postulating functional hypotheses for uncharacterized MADS-box genes, and evaluating the role of MADS-box genes in plant morphological evolution.

The monophyly of the type II lineage is strongly supported in the present analyses, and type I comprises several sublineages with divergent putative domains after the MADS. However, previous analyses (14), as well as contrasting exon-intron structure and chromosomal distribution between type I and II sequences, still support the placement of the root between the type I and II genes in the tree of the complete gene family (Fig. 1). This tree hence resolves type I and II sequences in two monophyletic lineages. Nonetheless, genes from other plant, animal, and fungal species should be included in future analyses to trace MADS-box gene

duplications with respect to taxa divergence and to specifically reevaluate the number of MADS-box gene duplications that occurred before the divergence of plants and animals (14). Such analyses will provide further evidence to reevaluate the monophyly of type I and II lineages.

Gene family structure has to be understood in the context of extensive gene duplications that have occurred in the evolutionary history of *A. thaliana* (40). Duplications leading to the chromosome stretches identified in the *Arabidopsis* Genome Initiative occurred 65 million years ago or before (refs. 50 and 51, but see ref. 52). Most retained groups within these stretches belong to type II genes, and duplications among type II seem to have been more ancient than those among type I, as suggested by their differential distribution among chromosomes. Retention due to a balance between genetic drift and mutation (5, 53) would depend on population characteristics (mainly effective population size) and hence would affect sequences of type I and II equally. But contrasting roles of selection between these two lineages could underlie the contrasting retention rates observed between them. The different evolutionary histories could have been determined by the fact that genes from these two lineages were recruited for different functions.

Interestingly, although duplications of type I seem to have occurred more recently than those leading to the type II lineage,

Table 2. Sites under PDS in the *A. thaliana* MADS-box gene family: "Branch-site analyses"

Branch-site analyses	<i>n</i>	$2\Delta\ell$ M3 (K = 2) vs. MB (df 2)	Parameter estimates under MB	Positively selected sites under MB
Branch leading to the FLC-like genes	39	20.50***	Type II $P_0 = 0.45, P_1 = 0.38, (P_2 + P_3 = 0.17),$ $\omega_0 = 0.06, \omega_1 = 0.36, \omega_2 = 4.47$	42 56 59 154 138 142 163 165 4 15 134 147 158 176
Branch leading to the SVP-like genes	39	13.13**	$P_0 = 0.44, P_1 = 0.37, (P_2 + P_3 = 0.19),$ $\omega_0 = 0.06, \omega_1 = 0.37, \omega_2 = 2.01$	16 144 26 170 2 4 7 55 84 129 132 133 184 186 188 197
Branch leading to PHERESI (AGL37)	48	10.10**	Type I $P_0 = 0.35, P_1 = 0.41, (P_2 + P_3 = 0.24),$ $\omega_0 = 0.10, \omega_1 = 0.32, \omega_2 = 6.52$	105 92 72 120 57 99 63 118 78 88 66 89 47

Each comparison has *n* sequences. Proportions of the component site classes 0 (P_0), 1 (P_1), and 2 + 3 ($P_2 + P_3$), as well as the values for the background ratios ω_0 and ω_1 and the foreground ratio ω_2 , are given under MB. **, $P < 0.005$; ***, $P < 0.001$; bold underlined, $PP \geq 0.99$ of being under positive selection; bold, $0.99 > PP \geq 0.95$; italics, $0.95 > PP \geq 0.90$; underlined, $0.90 > PP \geq 0.70$; normal, $0.70 > PP \geq 0.50$.

type I sequences are more divergent among them in comparison to type II. This finding would suggest that whereas type II genes have been affected by sporadic PDS at the origin of new functions, followed by strong functional constraint, type I genes have been subject to recurrent events of PDS. In turn, this could be an indication that the functional roles of type I genes are overall distinct to those of type II genes. This pattern also allows us to put forward the hypothesis that type I orthologues from other taxa are less conserved than most type II orthologues.

Indeed, PDS analyses presented here suggest that type I and II lineages have been subject to overall contrasting selection regimes. We found that recurrent positive selection could have played a role in fixing specific amino acids after several duplication events during the evolution of type I genes. In contrast, analyses of site models did not provide strong evidence for PDS selection among type II genes. We had to use site-branch models to detect a role for PDS among type II genes. Indeed, we found evidence for PDS along the branches leading to the groups of genes that control flowering time that evolved a new function with respect to most other genes characterized up to now that are involved in cell- or organ-type specification. Indeed, probably by their control of life-history traits, flowering-time genes may have directly impacted plant fitness, and this could also explain the prevalence of positive selection during protein evolution among them.

Sites with high PP of having been fixed by natural selection in both lineages were found mainly in the MADS and K domains. However, our analyses are biased toward these domains because we only focused on conserved stretches that may be unambiguously aligned and excluded most variable domains. Future studies focusing on particular closely related genes for several species will be useful to address the role of PDS within COOH and other divergent domains. Indeed, a recent study showed that regions within the C-terminal domain determine functional specificity in AP3 and PI and may be relevant for floral organ evolution (54). The localization of the sites with high PP identified here suggest a role for PDS in MADS-domain protein diversification through interactions with protein partners and changes in affinity to binding motifs (46, 47, 55). Sites and genes identified here to have been under PDS become interesting targets for functional evaluations.

Evaluations of assumptions and predictions made by models of gene duplication and persistence will require phylogenetically driven analyses of functional and population level data for related genes in different species. The MADS-box gene family might become a good "model family" for such a purpose. For example, our site model analyses did not find that PDS played a role in the divergence of redundant AP1, CAL, and FUL (56). However, population-level data do suggest a role for positive selection in the divergence of these genes (57). More powerful analyses should be performed to rule out false negatives in our analyses due to low gene number (32). Moreover, our conclusions were based on conservative analyses and unambiguously aligned sequences to avoid false positives. Additional tests (data available on request; Fig. 2b) suggest that the role of PDS in MADS-box gene evolution might be more widespread. Other approaches (58) and the inclusion of sequences for additional taxa should be considered when further investigating the role of PDS in MADS evolution.

Our results suggest a role for positive selection during MADS-box evolution in plants. Previous studies have emphasized the role of changes in cis-regulatory regions of transcriptional regulators during plant evolution (59). Fewer recent studies, however, have also demonstrated that the evolution of transcriptional regulators' cDNA sequences played important roles in plant evolution (54). The detection of positive selection in MADS protein sequences that are developmentally important also indicates that changes in cDNA, and not only in the regulatory regions of these genes, have played a role in the evolution of plant body plans.

We shared unpublished results and agreed on AGL numbers with Lucia Colombo and her collaborators. We thank Lorenzo Segovia and Julio Collado for computer time. Discussions with Francisco Vergara-Silva were insightful, and Alejandra Vázquez-Lobo, Rodolfo Salas, and Lev Jardón made comments and helped with the figures. Gary Ditta and Marty Yanofsky provided a preliminary list of MADS-like sequences. The comments of three anonymous reviewers improved this paper. Elizabeth Núñez helped with logistical tasks. This work was supported by Ph.D. fellowships to L.P.M.-C. from Consejo Nacional de Ciencia y Tecnología (CONACYT) and Dirección General de Estudios de Posgrado (DGEP) at the National Autonomous University of Mexico, and by grants from CONACYT, Programa de Apoyo para Proyectos de Investigación e Innovación Tecnológica, the Human Frontiers Science Program, and the University of California-Mexico (to E.R.A.-B.).

- Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, Heidelberg, Germany).
- Zhang, J., Zhang, Y.-P. & Rosenberg, H. F. (2002) *Nat. Genet.* **30**, 411–415.
- Ohta, T. (2000) *Gene* **259**, 45–52.
- Hughes, A. L. (2002) *Trends Genet.* **18**, 433–434.
- Lynch, M., O'Hely, M., Walsh, B. & Force, A. (2001) *Genetics* **159**, 1789–1804.
- Clark, A. G. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 2950–2954.
- Wagner, A. (1999) *J. Evol. Biol.* **12**, 1–16.
- Messier, W. & Stewart, C.-B. (1997) *Nature* **385**, 151–154.
- Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 2509–2514.
- Yang, Z. & Bielawski, J. P. (2000) *Trends Ecol. Evol.* **15**, 496–503.
- Jack, T., Brockmann, L. L. & Meyerowitz, E. M. (1992) *Cell* **68**, 683–697.
- Doyle, J. (1994) *Syst. Biol.* **43**, 307–328.
- Purugganan, M. D. (1998) *BioEssays* **20**, 700–711.
- Alvarez-Buylla, E. R., Pelaz, S., Liljegren, S. J., Gold, S. E., Burgeff, C., Ditta, G. S., Ribas de Pouplana, L., Martínez-Castilla, L. & Yanofsky, M. F. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 5328–5333.
- Purugganan, M. D. (1997) *J. Mol. Evol.* **45**, 392–396.
- Köhler, C., Hennig, L., Spillane, C., Pien, S., Gruijssem, W. & Grossniklaus, U. (2003) *Genes Dev.* **17**, 1540–1553.
- Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H. & Sommer, H. (1990) *Science* **250**, 931–936.
- Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. (1991) *Development (Cambridge, U.K.)* **112**, 1–20.
- Carpenter, R. & Coen, E. S. (1990) *Genes Dev.* **4**, 1483–1493.
- Goto, K. & Meyerowitz, E. M. (1994) *Genes Dev.* **8**, 1548–1560.
- Alvarez-Buylla, E. R., Liljegren, S. J., Pelaz, S., Gold, S. E., Burgeff, C., Ditta, G. S., Vergara-Silva, F. & Yanofsky, M. F. (2000) *Plant J.* **24**, 457–466.
- Liljegren, S. J., Ferrándiz, C., Alvarez-Buylla, E. R., Pelaz, S. & Yanofsky, M. F. (1998) *Flowering Newslett.* **25**, 9–19.
- Zhang, H. & Forde, B. G. (1998) *Science* **279**, 407–409.
- Johansen, B., Pedersen, L. B., Skipper, M. & Frederiksen, S. (2002) *Mol. Phylogenet. Evol.* **23**, 458–480.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680.
- Hall, T. A. (1999) *Nucleic Acids Symp. Ser.* **41**, 95–98.
- De Bodd, S., Raes, J., Florquin, K., Rombauts, S., Rouze, P., Theissen, G. & Van De Peer, Y. (2003) *J. Mol. Evol.* **56**, 573–586.
- Kramer, E. M., Dorit, R. L. & Irish, V. F. (1998) *Genetics* **149**, 765–783.
- Huelsbeck, J. P. & Ronquist, F. (2001) *Bioinformatics* **17**, 754–755.
- Yang, Z., Nielsen, R., Goldman, N. & Krabbe-Pedersen, A.-M. (2000) *Genetics* **155**, 431–449.
- Goldman, N. & Yang, Z. (1994) *Mol. Biol. Evol.* **11**, 725–736.
- Anisimova, M., Bielawski, J. P. & Yang, Z. (2001) *Mol. Biol. Evol.* **18**, 1585–1592.
- Yang, Z. & Nielsen, R. (2002) *Mol. Biol. Evol.* **19**, 908–917.
- Yang, Z. (1997) *Comput. Appl. Biosci.* **13**, 555–556.
- Michaels, S. D., Ditta, G., Gustafson-Brown, C., Pelaz, S., Yanofsky, M. & Amasino, R. M. (2003) *Plant J.* **33**, 867–874.
- Hartmann, U., Hohmann, S., Nettesheim, K., Wisman, E., Saedler, H. & Huijser, P. (2000) *Plant J.* **21**, 351–360.
- Ratcliffe, O. J., Kumimoto, R. W., Wong, B. J. & Riechmann, J. L. (2003) *Plant Cell* **15**, 1159–1169.
- Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17**, 1244–1245.
- Lynch, M. & Conery, J. S. (2000) *Science* **290**, 1151–1155.
- The Arabidopsis Initiative (2000) *Nature* **408**, 796–815.
- Henschel, K., Kofuji, R., Hasebe, M., Saedler, H., Munster, T. & Theissen, G. (2002) *Mol. Biol. Evol.* **19**, 801–814.
- Rounsley, S. D., Ditta, G. S. & Yanofsky, M. F. (1995) *Plant Cell* **7**, 1259–1269.
- Riechmann, J. L. & Meyerowitz, E. M. (1997) *J. Biol. Chem.* **272**, 1079–1101.
- Pellegrini, L., Tan, S. & Richmond, T. J. (1995) *Nature* **376**, 490–498.
- Santelli, E. & Richmond, T. J. (2000) *J. Mol. Biol.* **297**, 437–449.
- Huang, K., Louis, J. M., Donaldson, L., Lim, F.-L., Sharrocks, A. D. & Clore, G. M. (2000) *EMBO J.* **19**, 2615–2628.
- Krizek, B. A. & Meyerowitz, E. M. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 4063–4070.
- Zhang, J., Rosenberg, H. F. & Nei, M. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 3708–3713.
- Nei, M., Rogozin, I. B. & Piontkivska, H. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 10866–10871.
- Vision, T., Brown, D. G. & Tansley, S. D. (2000) *Science* **290**, 2114–2117.
- Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. & Van de Peer, Y. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 13627–13632.
- Blanc, G., Hokamp, K. & Wolfe, K. H. (2003) *Genome Res.* **13**, 137–144.
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-I. & Postlethwait, J. (1999) *Genetics* **151**, 1531–1545.
- Lamb, R. S. & Irish, V. F. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 6558–6563.
- Riechmann, J., Wang, M. & Meyerowitz, E. (1996) *Nucleic Acids Res.* **24**, 3134–3141.
- Ferrándiz, C., Gu, Q., Martienssen, R. & Yanofsky, M. F. (2000) *Development (Cambridge, U.K.)* **127**, 725–734.
- Purugganan, M. D. & Suddith, J. I. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 8130–8134.
- Suzuki, Y. & Nei, M. (2002) *Mol. Biol. Evol.* **19**, 1865–1869.
- Doebley, J. & Lukens, L. (1998) *Plant Cell* **10**, 1075–1082.

Supporting Materials and Methods

Sequences and Alignment. The query sequences for the BLAST searches were composite sequences made from the MADS- and K-domain of the following *Arabidopsis* proteins: *AG*, *AP3*, *PI*, *SHPI*, *SEP1*, *AGL3*, *SEP2*, *SHPI2*, *AGL6*, *API*, *FUL*, *SEP3*, *CAL*, *AGL11*, *AGL12*, *AGL14*, *AGL15*, and *AGL17*. The threshold for accepting a retrieved sequence as MADS-box was an *E* value of 10⁻³. The searches yielded a total of 112 and 29 distinct sequences, respectively, including the sequences used to make the composites. All of the sequences detected with the K-domain composite were also represented in the MADS-domain search. Sequences with <40% positives and those annotated as pseudogenes were excluded, except *AGL88*, which has a stop codon in position 23, but because *AGL88* is highly similar to *AGL57*, we believe that the stop may be spurious. Subsequent recurrent TBLASTN searches made with the newly found sequences as query detected six additional sequences with positives stretching at least half of the MADS-domain. Data from expression studies were used to help annotation, revealing that several predicted ORFs from the flowering-time *FLC*-like clade were part of a group of alternative splicings (1). In this study, we used the most conservative splicing form of those genes (Table 3).

The COOH region of type II genes was aligned by eye to assign the same homologies in B-function, B-sister, and AGL30-AGL66 clades postulated by De Bodt *et al.* (2) and Kramer *et al.* (3) and to match regions with periodic hydrophobic residues from AGL66-like genes with the corresponding regions of the K-box from other type II genes. To test the homogeneity of the phylogenetic signal, tree searches for type I and II genes were also performed on an alignment that did not include the less conservative sites discarded by Johannsen and coworkers (4). Topologies derived from complete and partial alignments (conserved regions) were compared to refine the alignment of the ambiguous stretches (Tables 4 and 5). AGL105 is excluded because it was very hard to align unambiguously.

In alignments that included both type I and type II sequences, the alignment of the MADS-boxes was unambiguous, but for the regions downstream of the MADS-box, two different strategies were followed: in one, no attempt was made to assign homology between type I and II regions downstream of the MADS-box and thus the MADS-box was implicitly given more weight. In the other, the aligned regions downstream of the MADS-box were treated as homologous blocks and put immediately after the MADS-box without further refining of the alignment between type I and II genes. Presumably, this strategy gives a similar weight to all regions in resolving branching order.

Annotation, Gene Structure, and Duplications. In addition to the functionally characterized genes, 10 sequences show high levels of sequence and intron-exon structure conservation with respect to one or more cloned genes belonging to the same clade, indicating that the former are probably well predicted. Finally, another 57 sequences group in monophyletic groups with high conservation among themselves, also suggesting that they are generally well predicted. For AGL65 and AGL66, which clearly group with AGL30-like sequences, we decided to exclude certain positions (amino acids 113-149 and 198-200 for AGL65, and 145-155 and 188-201 for AGL66) because these appeared to be indels with respect to the rest of the sequences in this group that could be otherwise easily aligned. These indels could be due to erroneous annotation or alternative splicing. Nonetheless, until all these latter cDNAs are cloned, their gene structures should be considered preliminary.

Reliability of Type II Alignment. To resolve the main clades of type II lineage we performed preliminary searches that included only the unambiguously aligned regions previously identified (4). These searches yielded the same main clades of type II genes as in shown Fig. 2b but with a relatively high statistical support at the base of each clade but with a very low support in the resolution of the relationships among clades (data not shown). In those analyses and the complete-alignment analyses, branching order varied depending on the type I gene used to polarize the phylogeny. These analyses confirmed that the main source of ambiguity in the alignment of the type II sequences derives from the highly divergent AGL30-like genes because these, together with the B-function genes, are the only genes that had one of two alternative positions depending on outgroup (the topology of the type II genes that is alternative to the one shown in Fig. 2b can be seen in Fig. 1).

1. Ratcliffe, O. J., Kumimoto, R. W., Wong, B. J. & Riechmann, J. L. (2003) *Plant Cell* **15**, 1159–1169.
2. De Bodt, S., Raes, J., Florquin, K., Rombauts, S., Rouze, P., Theissen, G. & Van De Peer, Y. (2003) *J. Mol. Evol.* **56**, 573–586.
3. Kramer, E. M., Dorit, R. L. & Irish, V. F. (1998) *Genetics* **149**, 765–783.
4. Johansen, B., Pedersen, L. B., Skipper, M. & Frederiksen, S. (2002) *Mol. Phylogenet. Evol.* **23**, 458–480.

Table 3. MADS-box sequences detected in the *Arabidopsis thaliana* genome

Gene name, symbol, and synonyms	Accession no.	Gene ID	BAC no. identifier (ORF number)	Chromosome	Predicted number of exons	Type of sister paralogon duplication *	Comments and references for functional studies
AGAMOUS / AG	AL021711 / AL161549	At4g18960	F13C5.130	4	8	Interchromosomal †	1
APETALA3 / AP3	AL132971 / AY070397	At3g54340	T12E18_30	3	7	Not duplicated in sister region	2
PISTILLATA / PI	AB035137	At5g20240	F5O24.130	5	6	Not duplicated in sister region	3
AGL1 / SHATTERPROOF1 / SHP1	AL353032	At3g58780	T20N10.130	3	7	Interchromosomal	4
AGL2 / SEPALLATA1 / SEP1	M55551	At5g15800	F14F8.180	5	7	Interchromosomal	4, 5
AGL3	AC006836	At2g03710	F19B11.16	2	8	Intrachromosomal	4
AGL4 / SEPALLATA2	AC009755	At3g02310	F14P3.4	3	7	Interchromosomal	4, 5
AGL5 / SHATTERPROOF2 / SHP2	AC006931	At2g42830	F7D19.17	2	7	Interchromosomal	4, 6
AGL6	AC003680	At2g45650	F17K2.18	2	8	Interchromosomal	4
AGL7 / APETALA1 / AP1	AC008262	At1g69120	F4N2.9	1	8	Intrachromosomal	7
AGL8 / FRUITFULL / FUL	AB008269	At5g60910	MSL3.3	5	8	Interchromosomal †	8
AGL9 / SEPALLATA3 / SEP3	AC002396	At1g24260	F3I6.19	1	8	Intrachromosomal †	5
AGL10 / CAULIFLOWER / CAL	P5K112	At1g26310	F28B23.25	1	8	Intrachromosomal	9
AGL11	AL049481 / AL161516	At4g09960	T5L19.90	4	7	Interchromosomal	10
AGL12	AC012654 / AL016163	At1g71692	F14O23.5	1	7	Interchromosomal	10, 11
AGL13	AL137898	At3g61120	T20K12.20	3	8	Interchromosomal	10
AGL14	AL078606 / AL161531 / AL161533	At4g11880	T26M18.90	4	7	Intrachromosomal	10

Gene name, symbol, and synonyms	Accession no.	Gene ID	BAC no. identifier (ORF number)	Chromosome	Predicted number of exons	Type of sister paralogon duplication *	Comments and references for functional studies
AGL15	AB005230	At5g13790	MXE10.8	5	8	Not duplicated is sister region	10
AGL16	AL137080	At3g57230	F28O9.80	3	6	Not duplicated is sister region	12
AGL17	AC006340	At2g22630	T9I22.7	2	7	Interchromosomal	11
AGL18	AL137080 / AF312663	At3g57390	F28O9.240	3	8	Intrachromosomal †	12, ‡
AGL19	AL031018 / AL161558	At4g22950	F7H19.130	4	7	Intrachromosomal	12
AGL20/ SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1 / SOC1	AC003680	At2g45660	F17K2.19	2	7	Not duplicated is sister region	13, 14
AGL21	AL035538 / AL161592	At4g37940	F20D10.60	4	7	Interchromosomal	11
AGL22/ SHORT VEGETATIVE PHASE / SVP	AC006592	At2g22540	F14M13.6	2	6	Interchromosomal †	15
AGL23	AC004512	At1g65360	T8F5.14	1	1	Not duplicated is sister region	
AGL24	AL035356 / AL161561	At4g24540	F22K18.260	4	8	Interchromosomal †	16, 17
AGL25/ FLOWERING LOCUS C / FLOWERING LOCUS F / FLC / FLM	AF116528 / AL356332	At5g10140	T31P16_130	5	7	Intrachromosomal	18, 19
AGL26	AF007270	At5g26870	F2P16.19	5	1	Not duplicated is sister region	
AGL27 / FLOWERING LOCUS M / FLM /FK1 / MAF1	AC002291	At1g77080	F22K20.15	1	6	Not duplicated is sister region	12, 20, §
AGL28	Y12776 / AC061957	At1g01530	F22L4.7	1	2	Not duplicated is sister region	
AGL29	AC004077	At2g34440	T31E10.22 (=F13P17.1)	2	2	Interchromosomal †	¶

Gene name, symbol, and synonyms	Accession no.	Gene ID	BAC no. identifier (ORF number)	Chromosome	Predicted number of exons	Type of sister paralogon duplication *	Comments and references for functional studies
AGL30	AC004138	At2g03060	T17M13.23	2	6	Not duplicated is sister region	
AGL31 / MAF2-I	AB019236 /	At5g65050	MXK3.30	5	6	Not duplicated is sister region	12, 21,
AGL32 / TRANSPARENT TESTA 16 / TT16	AB007648	At5g23260	MKD15.12	5	5	Not duplicated is sister region.	22
AGL33	AC004484	At2g26320	T1D16.4	2	2	Not duplicated is sister region.	**
AGL34	AF058914	At5g26575	F21E10.14	5	2	Not duplicated is sister region	
AGL35	AF058914	At5g26625	F21E10.9	5	1	Intrachromosomal †	
AGL36	AF058914	At5g26645	F21E10.10	5	1	Intrachromosomal †	¶
AGL37	AC004512	At1g65330	T8F5.11	1	1	Intrachromosomal †	
AGL38	AC004512	At1g65300	T8F5.8	1	1	Not duplicated is sister region	
AGL39	AF007271	At5g27130	A_TM021B04.16/ T21B4.40	5	1	Not duplicated is sister region	
AGL40	Z99708 / AL161589	At4g36590	C7A10.770	4	2	Not duplicated is sister region	
AGL41	AC005168	At2g26880	F12C20.8	2	2	Intrachromosomal †	
AGL42	AB016880	At5g62165	MTG10.20	5	7	No match found in database	††, ‡‡
AGL43	AB010699	At5g40220	MSN9.13	5	1	Interchromosomal †	
AGL44/ ABNORMAL NITRATE RESPONSE / ANR1	AC007210	At2g14210	F15N24.5	2	8	Not duplicated is sister region	24
AGL45	AC012393	At3g05860	F10A16.16	3	2	Not duplicated is sister region	
AGL46	AC007184	At2g28700	T11P11.1	2	1	Not duplicated is sister region	
AGL47	AB009050	At5g55690	MDF20.13	5	1	Not duplicated is sister region	
AGL48	AC018721 / AF085279	At2g40210	T7M7.9	2	1	Interchromosomal	
AGL49	AC005966 / AC007258	At1g60040	T2K10.9	1	1	Not duplicated is sister region	
AGL50	AC007258 / AC005966	At1g59810	F23H11.13	1	1	Not duplicated is sister region	

Gene name, symbol, and synonyms	Accession no.	Gene ID	BAC no. identifier (ORF number)	Chromosome	Predicted number of exons	Type of sister paralogon duplication *	Comments and references for functional studies
AGL51	AF075597 / AL161494	At4g02240	T2H3.15	4	2	Not duplicated is sister region	
AGL52	AL161531	At4g11250	F8L21.40	4	1	Not duplicated is sister region	
AGL53	AF160760	At5g27070	F15P11.40	5	1	Not duplicated is sister region	§§
AGL54	AF160760	At5g27090	F15P11.1	5	1	Not duplicated is sister region	
AGL55	AC018908	At1g60920	T7P1.6	1	1	Not duplicated is sister region	
AGL56	AC018908	At1g60880	T7P1.3	1	1	Not duplicated is sister region	
AGL57	AC016829	At3g04100	T6K12.28	3	1	Not duplicated is sister region	
AGL58	AC010155	At1g28450	F3M18.11	1	1	Not duplicated is sister region	
AGL59	AC010155	At1g28460	F3M18.10	1	1	Not duplicated is sister region	
AGL60	AC016529	At1g72350	T10D10.18	1	1	Intrachromosomal	
AGL61	AC006585	At2g24840	F27C12.24	2	1	Not duplicated is sister region	
AGL62	AB011483	At5g60440	MUF9.24	5	2	Not duplicated is sister region	
AGL63	AC004793	At1g31140	F28K20.7	1	5	Not duplicated is sister region	
AGL64						Not used in this study	¶¶
AGL65	AC011809	At1g18750	F6A14.14	1	7	Not duplicated is sister region	
AGL66	AC009243	At1g77980	F28K19.20	1	9	Intrachromosomal	
AGL67	AC009243	At1g77950	F28K19.16	1	5	Not duplicated is sister region	
AGL68/ FLOWERING C LOCUS1 / FCL1 / MAF5-I	AB02663 (partially) AB013395 (partially) AF214485	At5g65080	F15O5.4	5	6	Not duplicated is sister region	21,

Gene name, symbol, and synonyms	Accession no.	Gene ID	BAC no. identifier (ORF number)	Chromosome	Predicted number of exons	Type of sister paralogon duplication *	Comments and references for functional studies
AGL69 / FLOWERING C LOCUS 2 / FCL2 /MAF4	AB026633	At5g65070	F15O5.3	5	8	Intrachromosomal	21
AGL70 / MAF3-1	AB026633	At5g65060	F15O5.2 F15O5(N3)	5	7	Not duplicated is sister region	21, ***
AGL71	AB025623	At5g51870	MJM18.2	5	7	Not duplicated is sister region	
AGL72	AB010074/ AB025623	At5g51860	MIO24.20	5	7	Not duplicated is sister region	† † †
AGL73	AB005231	At5g38620	MBB18.17	5	1	Intrachromosomal †	
AGL74						Not used in this study	¶¶
AGL75	AB010072	At5g41200	MEE6.27	5	1	Intrachromosomal †	
AGL76	AB010699	At5g40120	MSN9.2; MUD12.4	5	1	Not duplicated is sister region	
AGL77	AB011478	At5g38740	MKD10.6	5	1	Not duplicated is sister region	
AGL78	AB011479	At5g65330	MNA5.6	5	1	Not duplicated is sister region	
AGL79	AP001314	At3g3026/At3g30270	T6J22.1	3	8	Not duplicated is sister region	‡ ‡ ‡
AGL80	AB015468	At5g48670	K15N18.16	5	1	Interchromosomal †	
AGL81	AB016876	At5g39750	MKM21.6	5	1	Interchromosomal †	
AGL82	AB016885	At5g58890	K19M22.9	5	1	Not duplicated is sister region	
AGL83	AB023033	At5g49490	K6M13.3	5	1	Not duplicated is sister region	
AGL84	AB023034	At5g49420	K7J8.9	5	1	Intrachromosomal †	
AGL85	AC005388	At1g54760	T22H22.17	1	1	Not duplicated is sister region	
AGL86	AC074360.1	At1g31630	F27M3_17	1	1	Not duplicated is sister region	
AGL87	AC006551	At1g22590	F12K8.7	1	3	Not duplicated is sister region	
AGL88	AC007045	At2g11990	F23M2.15	2	1	No match	§§§
AGL89	AC007478	At5g27580	F15A18.40	5	1	Intrachromosomal †	
AGL90	AC007627	At5g27960	F15F15.30	5	1	Not duplicated is sister region	¶¶¶

Gene name, symbol, and synonyms	Accession no.	Gene ID	BAC no. identifier (ORF number)	Chromosome	Predicted number of exons	Type of sister paralogon duplication *	Comments and references for functional studies
AGL91	AC036106	At3g66656	T8E24.5	3	1	Interchromosomal †	
AGL92	AC074360.2	At1g31640	F27M3_16	1	2	Intrachromosomal †	
AGL93		At5g26950	F2P16.17	5	1	Intrachromosomal †	
AGL94	NM_105623 / AC073178	At1g69540	F10D13_25	1	8	Not duplicated is sister region	
AGL95						Not used in this study	¶¶
AGL96	AP002543	At5g06500	F15M7.3 (also MHF15.29)	5	1	Not duplicated is sister region	
AGL97						Not used in this study	¶¶
AGL98		At5g39810	MKM21.13	5	2	Not duplicated is sister region	¶¶
AGL99	AL162875	At5g04640	T32M21_240 T1E3.5	5	1	Not duplicated is sister region	
AGL100	AC026479	At1g17310	T13M22.2	1	1	Intrachromosomal	
AGL101	AF160760	At5g27050	F15P11.2	5	1 (varies in different accessions)	Not duplicated is sister region	****
AGL102	AC012463	At1g47760	T2E6.17	1	2	Not duplicated is sister region	
AGL103	AB026654	At3g18650	MVE11.1	3	1	Interchromosomal †	
AGL104	AC069252	At1g22130	F2E2.20	1	10	Intrachromosomal	
AGL105	AP000607	At5g37420	T25O11.7	5	2	Not duplicated is sister region	

Consecutive AGL names for previously unnamed genes were assigned following agreement with Lucia Colombo and collaborators (personal communication).

*From the paralogons in the *A. thaliana* Database (<http://wolfe.gen.tcd.ie/athal/dup>).

†Paralogon contains fewer than seven genes and therefore could be spurious.

‡Reported as AGL15 in some GenBank entries.

§The MAF1-I splicing form (21) was used for this study.

¶Sequence At5g26650 (= AGL36) has been wrongly identified with AGL29.

||The MAF2-I splicing form (21) was used for this study.

**Reported as having eight exons in ref. 23.

††T. Nawy, J. E. Malamy, S. Thongrod, J. Jung, and P. N. Benfey,

www.arabidopsis2002.com/abstractspublic/abstract_expediente.asp?cdabstract=1147.

‡‡There is another MTG10.20 that is not a MADS protein.

§§This entry appears to have two MADS-box genes within the same ORF; we have called these AGL53 and AGL101, the former being identical to ORF F15P11.40 (see AGL101).

¶¶Personal communication by Lucia Colombo and Lucie Parenicová (not included in this study).

|||The MAF5-I splicing form (21) was used for this study

***The MAF3-I splicing form (21) was used for this study.

†††Also called MJM18.1 (but there is another MJM18.1 that is not a MADS-box gene).

‡‡‡In some reports it is wrongly identified as AGL8.

§§§Reported as a pseudogene.

¶¶¶Reported as synonymous with AGL29 (it is on a different chromosome).

||||Almost identical to F15P11.40 and to half of AGL53.

****Reported as a separate gene but could be part of AGL53 because AGL53 looks like two MADS proteins in the same ORF.

1. Yanofsky, M. F., Ma, H., Bowman, J. L., Drews, G. N., Feldman, K. A. & Meyerowitz, E. M. (1990) *Nature* **346**, 35–39.
2. Jack, T., Brockmann, L. L. & Meyerowitz, E. M. (1992) *Cell* **68**, 683–697.
3. Goto, K. & Meyerowitz, E. M. (1994) *Genes Dev.* **8**, 1548–1560.
4. Ma, H., Yanofsky, M. F. & Meyerowitz, E. M. (1991) *Genes Dev.* **5**, 484–495.
5. Pelaz, S., Ditta, G. S., Bauman, E., Wisman, E. & Yanofsky, M. F. (2000) *Nature* **405**, 200–203.
6. Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L. & Yanofsky, M. F. (2000) *Nature* **404**, 766–770.
7. Mandel, M. A., Gustafson-Brown, C., Savidge, B. & Yanofsky, M. F. (1992) *Nature* **360**, 273–277.
8. Gu, Q., Ferrandiz, C., Yanofsky, M. F., Bowman, J. L. & Martienssen, R. (1998) *Development (Cambridge, U.K.)* **125**, 1509–1517.

9. Kempin, S. A., Savidge, B. & Yanofsky, M. F. (1995) *Science* **267**, 522–525.
10. Rounsley, S. D., Ditta, G. S. & Yanofsky, M. F. (1995) *Plant Cell* **7**, 1259–1269.
11. Burgeff, C., Liljegren, S. J., Tapia-López, R., Yanofsky, M. F. & Alvarez-Buylla, E. R. (2002) *Planta* **214**, 365–372.
12. Alvarez-Buylla, E. R., Liljegren, S. J., Pelaz, S., Gold, S. E., Burgeff, C., Ditta, S. G., Vergara-Silva, F. & Yanofsky, M. F. (2000) *Plant J.* **24**, 457–466.
13. Onouchi, H., Igeno, M. I., Perilleux, C., Graves, K. & Coupland, G. (2000) *Plant Cell* **12**, 885–900.
14. Samach, A., Onouchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F. & Coupland, G. (2000) *Science* **288**, 1613–1616.
15. Hartmann, U., Hohmann, S., Nettekheim, K., Wisman, E., Saedler, H. & Huijser, P. (2000) *Plant J.* **21**, 351–360.
16. Michaels, S. D., Ditta, G., Gustafson-Brown, C., Pelaz, S., Yanofsky, M. & Amasino, R. M. (2003) *Plant J.* **33**, 867–874.
17. Yu, H., Xu, Y., Tan, E. L. & Kumar, P. P. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 16336–16341.
18. Sheldon, C. C., Burn, J. E., Perez, P. P., Metzger, J., Edwards, J. A., Peacock, W. J. & Dennis, E. S. (1999) *Plant Cell* **11**, 445–458.
19. Michaels, S. D. & Amasino, R. H. (2001) *Plant Cell* **13**, 935–941.
20. Scortecci, K. C., Michaels, S. D. & Amasino, R. D. (2001) *Plant J.* **26**, 229–236.
21. Ratcliffe, O. J., Kumimoto, R. W., Wong, B. J. & Riechmann, J. L. (2003) *Plant Cell* **15**, 1159–1169.
22. Nesi, N., Debaujon, I., Jond, C., Stewart, A. J., Jenkins, G., Caboche, M. & Lepinier, L. (2002) *Plant Cell* **14**, 2463–2479.
23. De Bodt, S., Raes, J., Florquin, K., Rombauts, S., Rouze, P., Theissen, G. & Van De Peer, Y. (2003) *J. Mol. Evol.* **56**, 573–586.
24. Zhang, H. & Forde, B. G. (1998) *Science* **279**, 407–409.

Table 4. Alignment of *Arabidopsis thaliana* type I MADS domain proteins

	← MADS domain					

	5	15	25	35	45	55
AGL43	-----	-----	-----	-----	---MTMRS-	-----SLPS
AGL75	-----	-----	-----	-----	---MTMRS-	-----SSPS
AGL76	-----	-----	-----	-----	---MTMR-	-----SLPF
AGL81	-----	-----	-----	-----	---MAIRSL	PSSSRCSSSS
AGL98	-----	-----	-----	-----	---MAIRSL	PSSSGCSNSS
AGL77	-----	-----	-----	-----	---MTTIRSS	PSSSRCSSNS
AGL78	-----	-----	-----	-----	---MK-	-----QASS
AGL52	-----	-----	-----	-----	---MK-	-----QASS
AGL51	-----	-----	-----	-----	---MK-	-----QSSF
AGL103	-----	-----	-----	-----	---MASSSSSSL	SFSTSKKNKT
AGL93	-----	-----	---MDSS	MSTKKKTKLS	VRNQTCFK--	-----KSSL
AGL53	-----	-----	-----	-----	---NQTCFK--	-----KSSL
AGL54	-----	-----	-----	-----	---NQTCFK--	-----KSSL
AGL89	-----	-----	---MDSS	MSTKKKTKLS	VRNQTCFK--	-----KSSL
AGL26	-----	-----	-----	-----	-----	-----
AGL101	-----	-----	-----	-----	-----	-----
AGL50	-----	-----	-----	-----	---APRQ-	-----KKPN
AGL49	-----	-----	-----	-----	---APRQ-	-----KKPN
AGL36	-----	-----	-----	-----	---MGMKK-	-----VKLS
AGL90	-----	-----	-----	-----	---MGMKK-	-----VKLS
AGL34	-----	-----	-----	-----	---MGMKK-	-----VKLS
AGL37	-----	-----	-----	-----	---MRGK-	-----MKLS
AGL38	-----	-----	-----	-----	-----	-----MKLS
AGL86	-----	-----	-----	-----	---MRS-	-----KIKLS
AGL92	-----	-----	-----	-----	---MRT-	-----KTKLV
AGL35	-----	-----	-----	-----	-----	-----MT
AGL80	-----	-----	-----	-----	---MTRK-	-----KVKLA
AGL46	-----	-----	-----	-----	---MARK-	-----KLNLT
AGL45	-----	-----	-----	-----	---MTRK-	-----KLNLS
AGL48	-----	-----	-----	-----	---MTRK-	-----KVKLV
AGL41	-----	-----	-----	-----	---MTRK-	-----KVKLA
AGL96	-----	-----	-----	-----	---MARK-	-----KVRAA
AGL87	-----	-----	-----	-----	---MGRR-	-----KVTHQ
AGL82	-----	-----	-----	-----	---MVPK-	-----VVDLQ
AGL47	-----	-----	-----	-----	-----MG-	-----RKMVKMT
AGL40	-----	-----	-----	-----	---MVRSTKG-	-----RQKIEMK
AGL62	-----	-----	-----	-----	---MVKSKSG-	-----RQKIEMV
AGL23	-----	-----	-----	-----	---MVKKTLG-	-----RRKVEIV
AGL28	-----	-----	-----	-----	---MARKNLG-	-----RRKIELV
AGL61	-----	-----	-----	-----	---IG-	-----RQKIPMV
AGL91	-----	-----	-----	-----	---MG-	-----RRKIKME
AGL29	-----	-----	-----	-----	---MG-	-----RRKIKME
AGL102	-----	-----	-----	-----	---MG-	-----RRKIEIK
AGL58	-----	-----	-----	-----	---KG-	-----KQKINIK
AGL59	-----	-----	-----	-----	---KG-	-----KQKINIK
AGL85	-----	-----	-----	-----	---MKTDW-	-----SHYLSVE
AGL88	-----	-----	-----	-----	-----	-----
AGL57	-----	-----	-----	-----	---KGRKTKG-	-----KQKIEMK
AGL84	-----	-----	-----	-----	---MVKKGGT-	-----KRKIAIE
AGL73	-----	-----	-----	-----	---GT-	-----KRKIAIE
AGL83	---MRFVPLYL	YEIERLWLSL	VNYLSPRKNK	NRRCGEIDKI	RMVKKGGT-	---KRKIAIE
AGL56	-----	-----	-----	-----	---MGGK-	---KTKIEIK
AGL55	-----	-----	-----	-----	---GT-	---KRKIEIK
AGL99	-----	-----	-----	-----	---MGGV-	---KRKISIE
AGL39	-----	-----	-----	-----	---GT-	---KRKIEIK
AGL60	---MEDGEA	STITFLPTE	PKPLQNPPLL	AK-PKKETKQ	KKPKTTKG-	---RQKIEIK
AGL100	MKDLFMEGER	ETSSMTCLTP	KDSVQSPNML	VRQPKKETTT	QTPKTTTRG-	---RQKIEIK
AGL33	-----	-----	-----	MKRTIKNKNK	QIVKENMG-	---RKKLKLK

	MADS domain					

	65	75	85	95	105	115
AGL43	SSSAYS----	-LASTSLSNR	LETIFKKASE	LCTLC-DIEA	CVIYYGPDGE	-----LKT
AGL75	SSSSYS----	-LAFTSLSNR	LETIFKKASE	LCTLC-DIEA	CVIYYGPDGE	-----LKT
AGL76	SSSSYS----	-LASTSLSNR	LETIFKKASE	LCTLC-DIEA	CVIYYGPDGE	-----LKT
AGL81	SSSSYS----	-LASTSLSNR	LETIFKKASE	LCTLC-DIEA	CVIYYGPDGE	-----LKT
AGL98	SSSSYS----	-LASTSLSNR	LETIFKKASE	LCTLC-DIEA	CVIYYGPDGE	-----LKT
AGL77	SSSSYS----	-LASTSLSNR	LETIFKKASE	LCTLC-DIEA	CVIYYGPDGE	-----LKT
AGL78	SSSS-----	-RNSTSLTNR	LKTIFKKAEE	LSILC-AIEV	CVIYYGPDGE	-----LRT

AGL52	SSS-----	-CNPTSLTNR	LKTIFKKAEE	LSILC-AIDV	CVIYYGPDGD	-----LRT
AGL51	SSSSSS----	-RNSTSLTNR	LKTIFKKAEE	LSILC-AIDV	CVIYYGPDGE	-----LRT
AGL103	FFKKPNSAFS	SSRATSLIKR	QQTVPFKAKE	LSILC-DIDV	CVICYGNGE	-----LKT
AGL93	SSSSTA----	-KKTNNLSMR	EQTMPFKKALE	LSTLC-NIDV	CVIYYGRDGK	-----LIKT
AGL53	SSSSTA----	-KKTNNLSMR	EQTMPFKKALE	LSTLC-NIDV	CVIYYGRDGK	-----LIKT
AGL54	SSS-NA----	-KKTNNLSMR	EQTMPFKKALE	LSTLC-DIEV	CVIYYGRDGK	-----LIKT
AGL89	SSSSTA----	-KKTNNLSMR	EETMPFKKALE	LSTLC-DIEV	CVIYYGRDGE	-----LIKT
AGL26	-----	-----MR	EDTMPFKKALE	LSTLC-DIEV	CVILYSRDGE	-----LIKT
AGL101	-----	-----	---MPFKKALE	LSTLC-NIEV	CVIYYGRDGE	-----LFKT
AGL50	KSDDDD-DLR	RKKQSFFKQR	FPGFKKKASE	LSVLC-GNSV	GFICYGSDS-	-----DLHV
AGL49	KSDDDDGDLH	RKKQSFFKQR	FPGFKKKASE	LSVLC-GNSV	GFICYGPDN-	-----DLHV
AGL36	LIANER----	-SRKTSFIKR	KDGIFKKLHE	LSTLC-GVQA	CALIYSPFIP	-----VPES
AGL90	LIANER----	-SRKTSFMKR	KNGIFKKLHE	LSTLC-GVQA	CALIYSPFIP	-----VPES
AGL34	LIANEI----	-SRETSFMKR	KNGIMKKLYE	LSTLC-GVQA	CTLIYSPFIP	-----VPE-
AGL37	FIENDES----	-VRKTTFTKR	KKGMLKKFNE	LVTLC-GVDA	CAVIRSPYNS	-----IQEP
AGL38	LIENSV----	-SRKTTFTKR	KKGMLKKLTE	LVTLC-GVEA	CAVVYSPFNS	-----IPEA
AGL86	LIANKT----	-SRRTTFRKR	KGGITNKLHE	LTTLC-GVKA	CAVISSPYEN	-----PVV
AGL92	LIPDRH----	-FRRATFRKR	NAGIFKKLHE	LTTLIC-DIKA	CAVIYSPFEN	-----PTV
AGL35	FIENET----	-ARKSTFKKR	KKGLKKAQE	LGILC-GVPI	FAVVNSPYEL	-----NPEV
AGL80	YISNDS----	-SRKATFKKR	KKGLMKKVHE	LSTLC-GITA	CAIYSPYDT	-----NPEV
AGL46	YIFNDR----	-MRKRSFKQR	REGFLKKLND	LKVLC-DVNA	CAVVYSPFNS	-----NPDV
AGL45	YITNES----	-MRKATFNKR	KKGLVKKIHE	LSVLC-GIEA	CAVIYSPFNS	-----NPEV
AGL48	WIENDK----	-SRATSLQKM	RVGLLKKVKE	LTILC-AVRA	IVIIFSPDKV	-----GPLV
AGL41	WIENDN----	-TRAIASLKR	RVGLVKKVRE	LSILC-DIKA	CTIVFSPNEA	-----ELMV
AGL96	WIRDDR----	-MRRASLKR	LTGLIKKVVNE	LSILC-DMRA	SVVVFNREEE	-----QLTA
AGL87	LISDNA----	-TRRVTFRKR	KDGLLKKIYE	LTVLC-GLPA	CAIYSEYKD	-----GPEL
AGL82	RIANDK----	-TRITTYKKR	KASLYKKAQE	FSTLC-GVET	CLIVYGPTKA	TDVVISEPEI
AGL47	RIANEK----	-TRITTYKKR	KACLYKKAQE	FSTLC-GVDT	CVIYGPSRA	GDEMVMPEL
AGL40	KMENES----	-NLQVTFSKR	RFGLFKKASE	LCTLS-GAEI	LLIVFSPGGK	-----VFS
AGL62	KMKNES----	-NLQVTFSKR	RSGLFKKASE	LCTLC-GAEV	AIVVFSPPGRK	-----VFS
AGL23	KMTKES----	-NLQVTFSKR	KAGLFKKASE	FCTLC-DAKI	AMIVFSPAGK	-----VFS
AGL28	KMTNES----	-NLQVTFSKR	RSGLFKKASE	LCTLC-DAEI	AIVVFSPPSGK	-----AYS
AGL61	KIKKES----	-HRQVTFSKR	RAGLFKKASE	LCTLC-GAEI	GIIVFSPAKK	-----PFS
AGL91	KVQDTN----	-TKQVTFSKR	RLGLFKKASE	LATLC-NAEV	GIVVFSPPGNK	-----PYS
AGL29	MVQDMN----	-TRQVTFSKR	RTGLFKKASE	LATLC-NAEL	GIVVFSPPGGK	-----PFS
AGL102	FIEDSI----	-ERKATFSRR	RNGIFKKADE	LAKLC-NVEI	AVLVISPTNI	-----PYT
AGL58	KIEKDE----	-DRSVTLSCR	LNAIYTMIE	LSILC-GVEV	AFIYSCSGK	-----PYT
AGL59	KIEKDE----	-GRSVTFSCR	LNGIYTKISE	LSILC-GVEV	AFIYSCSGK	-----PYT
AGL85	MES-----	-----TISN-	-----E	LSILC-GAEV	AFLGYSKSGK	-----PYT
AGL88	-----	-----	-----MNE	LVAMC-DVEV	AFLIFSPPKK	-----PYT
AGL57	KVENYG----	-DRMITFSKR	KTGIFKKMNE	LVAMC-DVEV	AFLIFSPPKK	-----PYT
AGL84	TIQKRD----	-SLRVCTCKR	REGLYSKASQ	LCLLS-DAQI	AILATPPSSE	S---NVSFYS
AGL73	TIQKRD----	-SLRVCTCKR	REGLYSKASQ	LCLLS-DAQI	AILATPPSSE	S---DVSFYS
AGL83	TIQKSD----	-YLRVCTCKR	REGLYSKASQ	LCLLS-DAQI	AILATPPSSE	S---NISFYS
AGL56	KIINKP----	-AKTVAFTRK	REGLFRKASQ	LCLLSPATQI	AILAAPMTSK	S---HASFYS
AGL55	RIEDKN----	-VRAVAFTRK	KSGLFHKASE	LCLLSPGTQI	AILATPLSSH	S---HASFYS
AGL99	LIEKKD----	-SRAVAFTRK	SRGLYSKASD	LCLLS-DAQI	AILATPVSSK	S---NVSFYT
AGL39	KRETKE----	-QRAVTCSSK	RQTVFSKAAD	LCLIS-GANI	AVFVTSPPSDS	S---DVSFYS
AGL60	EIMLET----	-RRQVTFSKR	RSGLFKKAAE	LSVLC-GAQI	GIITFSRCDR	-----IYS
AGL100	KIEEET----	-KRQVTFSKR	RRGLFKKSAE	LSVLT-GAKI	AVITFSKCDR	-----IYR
AGL33	RIESLK----	-ERSKFSKR	KKGLFKKAAE	VALLC-DSDI	MLIVVSPTEK	-----PTV

MADS domain →

	125	135	145	155	165	175
AGL43	WPPER----	E KVRDIALRYS	Q-LNEALRRK	KSVNLHGFLN	KKKK-----N	KGLKNTDK-K
AGL75	WPKEK----	E KVRDIALRYS	L-LNEALRRK	KSVNLHGFLN	KKK-----N	KGLKNPNK-K
AGL76	WPKEK----	E KVRDIALRYS	Q-LNEALRSK	KSVNLHGFLN	KKKKK---KK	KGLKNPNK-K
AGL81	WPPER----	E KVEDIALRYS	Q-LNEALRRK	KSVTLYDFLN	KKKDKTNLEK	KAKITDND-D
AGL98	WPPER----	E KVEDIALRYS	Q-LNEALRRK	KSVTLYDFLN	KKKNKTNLEK	KAKIKDND-L
AGL77	WPKER----	E KVRDIALRFN	Q-LNEALRHK	KSVNLHGFLN	KKK-----KN	KGLKNPNK-K
AGL78	WPKER----	E TVKDMALRYK	---EARKRK	KSRNLHEFLE	KE-----	---KDKD--K
AGL52	WPKDR----	E TVKDMALRYK	---EDRKRK	KCLNLHEFLE	KE-----	---VKDKDKYK
AGL51	WPKER----	N TVKDMASRYK	---EATKRK	K-----	-----	-----
AGL103	WPEER----	E KVKAIAIRYK	E-LSETKRRK	GSVDLHEFLE	KMNKD-----	DPEKEEKKKI
AGL93	WPDDQ----	S KVRDMAERFS	R-LHERERCK	KRTNLSLFLR	KK-----	-----
AGL53	WPEDQ----	S KVRDMAERFS	R-LHERERCK	KRTNLSLFLR	KK-----	-----
AGL54	WPEDQ----	S KVRDMAERFS	R-LHERERCK	KRTNLSLFLR	KQ-----	-----
AGL89	WPEDQ----	S KVRDMAERFS	K-LHERERRK	KRTNLSLFLR	KK-----	-----
AGL26	WPEDQ----	S KVRDMAERFS	K-LHERERRK	KRTNLSLFLR	KK-----	-----
AGL101	WPEDE----	S KVRDMAERFT	K-LNERERRK	KRTNLSLFLR	KK-----	-----
AGL50	WPQSQDHPN	Q ALHEIVAKFN	A-LSDERRKN	HACDLNDFP-	-----	-----

AGL49	WPQSQDHNPO	ALHEIVAKFN	A-LSDERRKN	HACDLNDFP-	-----	-----
AGL36	WPSR-----E	GAKKVASRFL	E-MPPTARTK	KMMDQETYLM	ER-----	-----
AGL90	WPSR-----E	GAKKVASRFL	E-MPRTARTR	KMMDQETHLM	ER-----	-----
AGL34	-----FL	E-MSPTARTR	KMMNQETYLM	ER-----	-----	-----
AGL37	WPSR-----E	GVEEVMSKFM	E-FSVLDRTK	KMVDQETFLR	QR-----	-----
AGL38	WPSR-----E	GVEDVVSFKM	E-LSVLDRTK	KMVDQETFLS	QR-----	-----
AGL86	WPST-----E	GVQEAVSMFM	E-RPATEQSK	LMMSHETYLQ	DK-----	-----
AGL92	WPST-----E	GVQEVISEFM	E-KPATERSK	TMMSHETFLR	DQ-----	-----
AGL35	WPSR-----E	AANQVVSQWK	T-MSVMDKTK	KMVNQETFLQ	QR-----	-----
AGL80	WPSN-----S	GVQRVVSEFR	T-LPEMDQHK	KMVDQEGFLK	QR-----	-----
AGL46	WPSK-----S	EVNNIKKFE	M-LPETQKKV	KSVNHEEFLN	-----	-----
AGL45	WPSN-----S	EVKNVMENFE	M-LTKLEQEK	KMVSHEGFIR	QN-----	-----
AGL48	WPSP-----Q	ATHGLLDEFF	A-LPKSVQKK	KESNVESYLK	EK-----	-----
AGL41	WPS-----	-----VE	RLMDIELFLN	EK-----	-----	-----
AGL96	WPSN-----E	AANSLIDNFI	S-LTDHERTM	KAVDPES---	-----	-----
AGL87	WPNL-----N	EVRSILNRLS	E-LPVEKQTK	YMMDQKDLNM	KM-----	-----
AGL82	WPKDE---T	KVRAIIRKYK	DTVSTSCRKE	TNVETFVNDV	GK-----	-----
AGL47	WPKDG---S	KVREILTKYR	DTASSSCTKT	YTVQECLEKN	-----	-----
AGL40	FGHP-----	SVQELIHRFS	N-PNHNSAIV	HHQNNN----	-----	-----
AGL62	FGHP-----	NVDSVIDRFI	N---NNPLP	PHQHNN----	-----	-----
AGL23	FGHP-----	NVDVLLDHRF	G----CVVG	HNN-----	-----	-----
AGL28	FGHP-----	NVNKLLDHSL	G----RVIR	HNN-----	-----	-----
AGL61	FGHP-----	SVESVLDRYV	S-RN-NMSLA	QSQ-----	-----	-----
AGL91	FGKP-----	NFDVIAERFK	N-EFEEEEEG	DSCET-----	-----	-----
AGL29	YGKP-----	NLDSVAERFM	R-EYDDSDSG	D-----	-----	-----
AGL102	YGYP-----	CFNDVVERIQ	N---P-----	-----	-----	-----
AGL58	FGSP-----	SFQAVVERFL	N-G--EASSS	SSSSLQ----	-----	-----
AGL59	FGSP-----	SFQAVVERFL	N-G--DASSS	SSS-----	-----	-----
AGL85	FGSP-----	SFQAVVERFL	N-R--EASSS	LQR-----	-----	-----
AGL88	FAHP-----	SMKEVADRLK	N---PSRQEP	LEKDN-----	-----	-----
AGL57	FAHP-----	SMKKVADRLK	N---PSRQEP	LERDD-----	-----	-----
AGL84	FGHS-----	SVDAVVSAFL	S---GQR--P	VPKDNKETRE	DV-----	-----
AGL73	FGHS-----	SVDAVVSAFL	S---GKRVPV	APKDNKETRE	DV-----	-----
AGL83	FGHS-----	SVDAVVSSFL	S---GQRVCP	LQEDTKEMRE	DV-----	-----
AGL56	FGHS-----	SVDNVVSSLL	Y---DHPPLT	ANQDN-----	-----	-----
AGL55	FGHS-----	SVDHVVSSLL	H-N-QHPSLP	TNQDN-----	-----	-----
AGL99	FGHS-----	SVDNVVAAFL	T-N---QRP	RE-----	-----	-----
AGL39	FSGYS-----	SAYEIADCYL	NRKPPPKIVN	PA-----	-----	-----
AGL60	FGN-----	-VNSLIDKYL	R-KAPVMLRS	HPGGN-----	-----	-----
AGL100	FGH-----	-VDALIDKYL	R-KSPVKLEG	YSGDN-----	-----	-----
AGL33	FNTRSRSFHT	I-----	-----LER	FCMLSLQERE	ERCDLSYFYI	IIT-----

	185	195	205	215	225	235
AGL43	RKTSCLK-VN	VLKYPLADHY	PPDQVSP--	-IQSLELHVS	KFHE-RLEFL	ESR-KQNETQ
AGL75	MKTSCLK-VN	ILKYPLADHY	PPDQVSP--	-IQSLELHVS	KFQE-RLRFL	ESQ-KQNQTK
AGL76	RKTCLKNVN	VLKYPLADHY	PPDQVSQL--	-TQSLKLHVS	KFQE-RLRFL	ESQ-KQ--TK
AGL81	LKTCLKN-VN	ILKYPLADHY	SPDQVSQL--	-IQSLEPHVS	KVRE-RIRFV	ESQ-KHKETK
AGL98	KRL-----	-----	-----	---SLEPHVS	KVRE-RIRFV	ESQ-KHKETK
AGL77	KKTSCLK-VN	VLKYPLADHY	SPDQVSQL--	-TQSLELNVS	KFQE-RLRFL	ESQ-KQNETK
AGL78	GKTNLKK---	-NWYPNFDHY	SPQQLSQL--	-IQSLERTLS	TLQE-RLRIV	EAQ-KLQNTN
AGL52	GKTNYVK--N	PNWYPNFDHY	SPQQLSQL--	-IQSLERTLS	TLQK-RLRIV	ESQ-KKQNTN
AGL51	-----	---KR	-----	-----TLS	TLQE-RLRIV	ESQ-KQQNKN
AGL103	KVRRVPKVY	PVWDPFRDNY	SVEQLMGL--	-VQSLERNLT	RIQH-RTCAV	VEA-QGQRRV
AGL93	-----	---ILDDTK	LSEKVLEM--	-EDSLESGLR	VLQD-KLLLL	QPE-KNQTEF
AGL53	-----	---ILDDTK	LSEKVLEM--	-EDSLESGLR	VLQD-KLLLL	QPE-KNQTEF
AGL54	-----	---ILHDKK	LSEKVLEM--	-EDSLESGLR	VLQD-KLLLL	QPE-KNQTEL
AGL89	-----	---ILDDNK	LSEKVLEM--	-KDSLESGLR	VLQD-KLLLL	QPE--KNQTEL
AGL26	-----	---ILDNSK	LSEKVLEM--	-KDSLESGLR	VLQD-KLLLL	QPE-KNQTEL
AGL101	-----	---ILDDNK	LSGKVLEM--	-KDSLERGLR	VLQD-KLLLL	QPE--NQTKS
AGL50	-----	---HHL	KGLSREELR-	-KHLHLDSQ	LLGV-REQKI	EIL-KKTLTG
AGL49	-----	---HHL	KGLSREELR-	-KHLHLDSQ	LLGV-REQKI	EIL-KKTLTG
AGL36	-----	---ITK	AKEQLKNLA-	-AENRELQVR	RFMF-DCVEG	KMS-QYHYDA
AGL90	-----	---ITK	AKEQLKNLA-	-AENRELQVR	RFMF-DCVEG	KMS-QYRYDA
AGL34	-----	---ITK	AKEQLQNLV-	-GANQELQYR	-----	-----YDA
AGL37	-----	---IAK	ETERLQKLR-	-DENRNSQIR	DLMF-GCLKG	EVD-VSHLHG
AGL38	-----	---IAK	EKEQLQKLR-	-DENHNSQIR	ELMF-GCLKG	ETN-VYNLDG
AGL86	-----	---ITK	ETKKLES LR-	-RENRESQLR	QFMF-DCVEG	KMS-EHQYGA
AGL92	-----	---ITK	EQNKLES LR-	-RENRETQLK	HFMF-DCVGG	KMS-EQQYGA
AGL35	-----	---ITK	ATESWKKLR-	-KENKELEMK	NIMF-DCLSG	CTL-VSSIEK
AGL80	-----	---IAK	ATETLRRQR-	-KDSRELEMT	EVMF-QCLIG	NME-MPHLNI
AGL46	-----	---LYISK	VEKQSKKLI-	-VENKETCLK	EVMF-KCLGG	NMG-DFVMND

AGL45	-----	-----	ISK	TMESNNKKM-	-IDNAERTMK	EAMF-QLLSG	KGE-KLNLTD
AGL48	-----	-----	THK	FQEQLKKS-	-KKNKEHVID	ELMM-QLQSG	R-E-IADLNQ
AGL41	-----	-----	TNK	VNEKLIKSC-	-KKNKEYVSN	ELMM-QLQRG	R-R-IHDLNL
AGL96	-----	-----	YVQT	VIEKIEKKR-	-ADTRKVITE	FEMDELMFQV	QNGRELADLS
AGL87	-----	-----	IQD	AEKKLEK---	-----	EKMHT	RAMK-LGLMA
AGL82	-----	-----	GNEV	VTKKRVKREN	KYSSWEEKLD	KCSREQLHGI	FCAVDSKLINE
AGL47	-----	-----	NTK	VEKTIAT---	-----	KYPTWDK	KLDQCSLNDL
AGL40	-----	-----	LQ	LVETRPDR--	---	NIQYLNN	ILTE-VLANQ
AGL62	-----	-----	MQ	LRETRRNS--	---	IVQDLNN	HLTQ-VLSQL
AGL23	-----	-----	TN	LDESUYTKL--	---	HVQMLNK	SYTE-VKAEV
AGL28	-----	-----	TN	FAESRTKL--	---	RIQMLNE	SLTE-VMAEK
AGL61	-----	-----	Q	LQGS--PA--	---	ASCELNM	QLTH-ILSEV
AGL91	-----	-----	SG	YSRGNRAR--	---	QEKKICK	RLNS-ITEEA
AGL29	-----	-----	EE	KS-GNYRP--	---	KLKRLSE	RLDL-LNQEY
AGL102	-----	-----				SASS-KLRSL	MKELEQIKEF
AGL58	-----	-----	RS	VKNAHQQA--	---	KIQELCK	RYNR-LVEEL
AGL59	-----	-----	SL	VMNAHQQA--	---	KIQELCK	KYNR-LVEEL
AGL85	-----	-----	S	VMNAHQQA--	---	KIQELCK	VYNR-MVEEA
AGL88	-----	-----	TRP	LVEAYKKQ--	---	RFHDLIK	KMEA-LEEEL
AGL57	-----	-----	TRP	LVEAYKKR--	---	RLHDLVK	KMEA-LEEEL
AGL84	-----	-----	GIC	LTRNNLGL--	---	GFWWNDE	SLAR-SENPO
AGL73	-----	-----	GIC	LTRKNLGL--	---	GFWWNDE	SLVR-SENPO
AGL83	-----	-----	AIC	LSRTNLGL--	---	GFWWNNE	SLNK-SENPO
AGL56	-----	-----	RS--GL		---	GFWWEDK	RFDV-SENVE
AGL55	-----	-----	RS--GL		---	GFWWEDQ	AFDR-LENVD
AGL99	-----	-----	GL--GL		---	DYWWEDE	RLSK-SEdle
AGL39	-----	-----	GS--KL		---	GFWWEDP	DLYHSCDDLS
AGL60	-----	-----	VA	NGEENDNG--	---	LMWWEWA	VESV-PE--E
AGL100	-----	-----	AA	DEESRRP---	---	WWERP	VESV-PE--E
AGL33	-----	-----					ELE-EYMAAL

	245	255	265	275	285	295
AGL43	PDHHS----	LASSSLNHQT	QSLNPSQFSL	FMYNHGD-NT	LSQIPVS---	-----
AGL75	PDHQS----	LTPSSLNHYT	QSLNPSQFSL	FMYNHGD-NT	LSQIPVS---	-----
AGL76	PDHQS----	LTPSSLNHQT	QSLNPRQFSL	FMYNHGD-NT	LSQISVS---	-----
AGL81	PDHQS----	LASSSLNHQT	QSLNPSQFSL	FMYNHGD-NT	LSQIPVS---	-----
AGL98	PDHQS----	LASSSLNNQT	QSLNPSQFSL	FMYNHGD-NI	LSQIPVS---	-----
AGL77	PDHQS----	LTS--ISSLN	QSLNPSQFSL	FMYNHGY-NT	LSQIPVS---	-----
AGL78	LVBQS----	LTP-SYLNQT	QHLNPSKFSL	FMYNHGD-AT	LSQLPLS---	-----
AGL52	LVBQS----	LTP-SYLNQT	QHLDPKSFSL	YMYNHGD-AT	LSQLPLS---	-----
AGL51	LVBQS----	LTP-SYLNQI	QHLNPSNFSP	YMYNHGDAAT	LSQLPLS---	-----
AGL103	QYTNMANQEL	MMA-NTMNQL	QQHSN-QVSM	YLWNHGNGA-	FSQIPVS---	-----
AGL93	GQTRA----	VSS--TTN	PLSPPPS---	LIEDHRHQQR	TEPLMSG---	-----
AGL53	GQTRA----	VSS--TTN	PLSPPPS---	LIEDHRHQQW	TEPLMSG---	-----
AGL54	GQSCA----	VYS--TTY	PLSSPSP---	IEDHQHQQW	TEPLSNT---	-----
AGL89	GQSRA----	VSS--TTN	PLSSP---	--EDHHHQW	TEPLVTG---	-----
AGL26	GQIPV----	INN--GQN	HW-----			-----
AGL101	-LTRS----	VSS--LDY	LRTGNMKKKR	CILSSQIKLI	VENEFCS---	-----
AGL50	SSEKD----	GARVS-ENSA	ISDHKLKIEP	HLKDILSEDH	LIRVSDK---	-----
AGL49	SSEKD----	GARVS-ENSA	ISDHKLKIEP	NLTDILSEDH	LIRVSDK---	-----
AGL36	KDLQD----	LQSCINLYLD	QLNGRIESIK	ENGESLL-SS	VSPFPTRIGV	DEIGDESFS
AGL90	KDLQD----	LLSCINLYLD	QLNGRIESIK	ENGESLL-SS	VSPFPTRIGV	DEIGDESFS
AGL34	KDLQD----	LLSCINLYLD	QLNGRIEILK	EHGDSLP--S	VSPFPTRIGV	EETGDESSD
AGL37	RDLLD----	LNVFLNKYLN	GVIRRVEILK	ENGESSSSVP	PPIGVA---	-----
AGL38	RDLQD----	LSLYIDKYLN	GLTRRIEILI	ENGESSSSLP		-----
AGL86	RDLQD----	LSLYIDHYIN	QLNSSVMLLT	NNGASSSSFP	PPLHT-----	S
AGL92	RDLQD----	LSLFTDQYLN	QLNARKKFLT	EYGESSSSVP	PLFDVAGANP	PVVADQAAVT
AGL35	TELRD----	FGYVIEQQLK	DVNRIEILK	RNNEPSSALV	--PVAAP---	-----
AGL80	VDLND----	LGYMIEQYLK	DVNRIEILR	NSGTEI-GES	SSVAVAASEG	NIPM-----
AGL46	NDRLD----	LCKFIDHYLR	NLYHHKNVTL	NNPNFEIGES	SSLMDMAP--	-----
AGL45	RNRD----	LCKYIDQYLK	ELYHHKNKTI	NQSHIE----	--PGESS---	-----
AGL48	SEMYA----	LARSFRDITL	LCRKKLAFMQ	FPPLRDP--P	VFPFEIQVEE	FKTTTNDGFV
AGL41	SEIYT----	LLSYSRETIM	SFRKKFDFMQ	HSPLRDP--P	VLPFEVQVEQ	FKSTTKDAFL
AGL96	PTEAD----	KLIPYADKKL	MWLSKRMGST	GVDALRASNV	ASGSGGN---	-----
AGL87	DCSEE----	LARAADVVDK	KLKAIRERIK	AVEAGA----	--PIIKR---	-----
AGL82	AVTRQ----	ERS--MFRVN	HQAMDTFPFQ	NLMDQQFMPQ	YFHEQPQ---	-----
AGL47	IQEAT----	NRN-QTFPDT	SCWNSDQLGL	CGYNRQCFEQ	YQLFPLP---	-----
AGL40	LDLLK----	ES--REQVG	NWYEKDV---	KDLDMN----	--ETN-----	-----
AGL62	LKKIR----	EK--TKALG	NWVEDPV---	EELALS----	--QLE-----	-----
AGL23	RAQNE----	RE--NENAE	EWWSKSP---	LELNLN----	--QST-----	-----

AGL28	IVQNE----	RE---NKDAE	KWRNNSP---	TELNLA----	--QST-----	-----
AGL61	MEEMR-----	KES--VRRSMI	NWWEKPV---	EEMNMV-----	--QLQ-----	-----
AGL91	LHKWL-----	ES----AEQ	DKFNKPI---	EELTLE-----	--ELK-----	-----
AGL29	SQEKL-----	ES----AGD	ERFKESI---	ETLTLT-----	--ELN-----	-----
AGL102	QEDLR-----	KKQQRN-LEK	SNMK-ENVDL	KLEDLV-----	-----	-----
AGL58	AAALA-----	ET--RAVNKD	AWWKADPN--	DVKDHE-----	--KAK-----	-----
AGL59	AAALA-----	ET--RVVNKD	VWVKVDPN--	DVKDHE-----	--KAK-----	-----
AGL85	AAALA-----	ET--MPVDED	AWWKVDPK--	EVEDHE-----	--EAK-----	-----
AGL88	KESRN-----	EKK----LDK	MWWNFPSEGL	SVKELQ-----	--ERH-----	-----
AGL57	KESRN-----	EKK----LDK	MWWNFPSEGL	SAKELQ-----	-----	-----
AGL84	RTLLR-----	NLK-ELRADE	ALACNQAFVN	DREDLK-----	--NNDKC-----	-----
AGL73	WTLLS-----	NLK-ELRADE	ACVN-----	DHKDLK-----	--KNEKS-----	-----
AGL83	LTLLS-----	NLK-ELSGEE	ALVN-----	DHKDLK-----	--KNERS-----	-----
AGL56	SRMLN-----	NVR--CRLND	AVK-S-----	TQRDGGLEIL	HHQE-----	----EEVLQT
AGL55	SRMLN-----	NVR--LRLDD	AVK-S-----	NQRDGSLS-VI	HQED-----	----EEVLQL
AGL99	SKMLK-----	DLK-DLQNR	DCEEDVKKKG	VLHGTHQKQT	FNPE-----	----SCSVNF
AGL39	QRMKK-----	HVMACLEKEE	KSQLV-----	SSFDQNPNST	CSLDVEDCDG	SSYSQIASTF
AGL60	SVLRE-----	NLL--TRIQY	MSGD-----	RTVEN-----	-----	-----
AGL100	SMLRE-----	NIG--KKIVA	MGND-----	RTVDM-----	-----	-----
AGL33	-----	-----	-----	-----	-----	-----

	305	315	325	335	345	355
AGL43	-----	-----ASNF	NQDYFSALLE	QSELKN----	---QLMKQEI	CGNDQONNMW
AGL75	-----	-----ASNF	NQDYFSALLE	ESELKN----	---QLMKPEI	CGYDQONQMS
AGL76	-----	-----ASNF	NQNYFSALLE	QSELKN----	---QLMKQD-	-GYDQONQMR
AGL81	-----	-----ASNF	NQDYFSALLE	QSELKS----	---QIMKQDL	CGYEQNMCMs
AGL98	-----	-----ASNF	NQDYFSALLE	QSELKS----	---QIMKQEV	CGYEQNMCMs
AGL77	-----	-----ASNF	NQDYFSALLE	QSELKS----	---QIMKQEV	CGYEQNMCMs
AGL78	-----	-----APHS	NQLINYQNHL	MQHGF-----	---QNMCSDN	ITNNN-FEHP
AGL52	-----	-----ASQS	NQLINYQ---	MQHGF-----	---QNMCLDN	ITNNNFOHP
AGL51	-----	-----ASLS	NQLQLPES--	-----	-----	-----
AGL103	-----	-----ALASNQ	TQSLAPIPPE	LMIYPN-----	---SDAGNYS	GSLGVQGTGI
AGL93	-----	-----VSNT	EQDLSTSSLS	QNQSKF-----	---SVFLYN-	-----
AGL53	-----	-----VSNT	EQDLSTSSLS	QNQSRI-----	---SVFLYN-	-----
AGL54	-----	-----E--	-----	-----	-----	-----
AGL89	-----	-----VSNT	EQDLSTSPLS	NHQSKY-----	---SVFVYN-	-----
AGL26	-----	-----	-----	-----	-----	-----
AGL101	-----	-----KQNL	TKKSTKIFE	KQNFLK-----	---CFFIYG-	-----
AGL50	-----	-----KLG	SCDVFDELAY	VVRGS-----	---RNLNENV	SKYE-----
AGL49	-----	-----KLG	SCDVFDELAY	VVRGS-----	---RNLNENV	SNYE-----
AGL36	SPIHATTGVV	DTLNATNPVH	LTGDMTPFLD	ADATAVTASS	RFFDHIPYEN	MNMSQNLHEP
AGL90	SPIHSTTRVV	DTPNATNPVH	LAGDMTPFLD	ADANAVTAPS	RFSDHIQYEN	MNMSQNLHEP
AGL34	SPILATTTGVV	DTPNATNPRV	LVADTTHFLD	ANATAVTAPF	GFSNHIQYKN	MNMSQDLHRP
AGL37	-----	-----PTV	VDAASVPIGFD	G-----	---RMIQDQN	QNQ-----
AGL38	-----	-----LPIV	ANAAAPVGF	G-----	---PMFYHN	QNQ-----
AGL86	VAGAGAGAGA	APLVVAGAGA	APLAVAGAGA	S-----	-----	-----
AGL92	VPPLFAVAGA	NLPVVADQAA	VTVPPLFAVA	G--ANLPVVA	DQAAVNVPTG	FHNMNVNQNO
AGL35	-----	-----	TTSSVMPVVE	MGSSS-----	---FAAVANF	VNP-----
AGL80	-----	-----PNLVA	TTAPTPTTIE	VGSSSS-----	---FAAVANF	VNP-----
AGL46	-----	-----TAT	TGNMATTVVD	EG-----	-----	-----
AGL45	-----	-----GAT	NAMTPTSVE	NPNFN-----	-----	-----
AGL48	GGGQD-----	-----NKRAG	RTDEATRFIN	TDIFK-----	---QSKSYFF	FDEWVFPSP
AGL41	GGDLL-----	-----VERAR	NTNEATRIIN	IDSLR-----	---ENKSYLL	IDQWFPTEP
AGL96	-----	-----GL	NMMETGRSFY	YVDK-----	-----	-----
AGL87	-----	-----D-	-----	-----	-----	-----
AGL82	-----	-----FQGF	PNNFNMGFS	LISPHDG---	-----	-----
AGL47	-----	-----TMDY	NGLSFFPFNN	QMTS-----	-----	-----
AGL40	-----	-----	QLISALQDVK	KK-----	-----	-----
AGL62	-----	-----	GFKGNLENLK	KV-----	-----	-----
AGL23	-----	-----	CMIRVLKDLK	KI-----	-----	-----
AGL28	-----	-----	SMKCDLEALK	KE-----	-----	-----
AGL61	-----	-----	EMKYALEELR	KT-----	-----	-----
AGL91	-----	-----	EFEAKIKKIS	CG-----	-----	-----
AGL29	-----	-----	EYKDRLQTVH	GR-----	-----	-----
AGL102	-----	-----	AFKAKLEAYQ	AG-----	-----	-----
AGL58	-----	-----	KMMEKYQELK	EK-----	-----	-----
AGL59	-----	-----	KMMEKYQELY	DK-----	-----	-----
AGL85	-----	-----	KIMEKCEGLY	EK-----	-----	-----
AGL88	-----	-----	---QAMLELR	DN-----	-----	-----
AGL57	-----	-----	QRYQAMLELR	DN-----	-----	-----

AGL84	-----	-----DFVS	DHETHDQTLI	LQSASPICCI	PENLNEITQE	PNQTLNIQSS
AGL73	-----	-----D---	VHGTQDQTLI	FQSASAVCCI	PENLNDITQE	PNQTLDIQSS
AGL83	-----	-----DVVL	QHGTQYETLN	-----	-----	PN-----SN
AGL56	RN-----	-----	-----	-----	-----	-----
AGL55	GY-----	-----	-----	-----	-----	-----
AGL99	DG-----	FN KNTTEFDLDE	IFDYVSTAEA	LSMNLDMDDV	SVVTTN-----	-----
AGL39	TP-----	NSVN EYCSDQTFSS	FH---GDQNP	NLSSPSFDQD	CY--SS-----	-----
AGL60	-----	-----	LPAPPNEMAM	AD-----	-----	-----
AGL100	-----	-----	VPAPWINVMG	WK-----	-----	-----
AGL33	-----	-----	-----	-----	-----	-----

	365	375	385	395	405	415
AGL43	-MG-----	---NITNNN	FQLPCVS---	-VQESVNNFG	-----	-----
AGL75	-MG-----	---DITNNK	FQDPCVSNKE	AVQESVNNFG	L-----	-----
AGL76	-MG-----	---DITNNN	FQLPYFSKKE	AVQESVNYFG	M-----	-----
AGL81	NHG-----	---DATLSQ	IPLSASNLNQ	DFSALLQDES	-----	-----
AGL98	NNG-----	---DATLSQ	IPLSASNFNQ	EFALLQDES	-----	-----
AGL77	NHG-----	---DATLSQ	IPFSASNFNQ	DFSANNFNQH	SFVSNTQDYY	SVQKSVNNNY
AGL78	-GV-----	---SNTQDY	SPLLSVQASA	VNNYGLNNHL	M-----	-----
AGL52	-GV-----	---SNTQDY	SPLLS-----	ANNYGLNNHL	M-----	-----
AGL51	-----	---LDA	AWFWS-----	--EHVFGQHH	Q-----	-----
AGL103	-NG-----	---LQNMNM	LTYNININSVN	DFSKQFDQ-N	S-----	-----
AGL93	-----	---HDNCS	FYQVP-----	---DSVSSFD	S-----	-----
AGL53	-----	---HDNRS	FYQVP-----	---DSVSSFD	-----	-----
AGL54	-----	-----	-----	-----	-----	-----
AGL89	-----	---HDSGS	FYQVP-----	---DSICF--	-----	-----
AGL26	-----	-----	-----	-----	-----	-----
AGL101	-----	---VIFSY	EIKPT-----	---AQFVSLG	-----	-----
AGL50	-----	---SKD	ADNTG-----	---LDHLVTLG	G-----	-----
AGL49	-----	---SKD	AAYTG-----	---MDHLGTFG	G-----	-----
AGL36	FQHLVPTNVC	DFQONQNMNQ	VQYQAP----	---NNLFNQIQ	R-----	-----
AGL90	FQHLVPTNVC	DFYQONQNMNQ	VQYQAP----	---NNLFNQIQ	R-----	-----
AGL34	FQHLVPTNFC	DFQONQNMNQ	VQYQAPP----	---NDMFNQIQ	R-----	-----
AGL37	-----	---QEPVQ	FQYQAL----	---YDFYDQIP	K-----	-----
AGL38	-----	---QKPVQ	FQYQAL----	---YDFYDQIP	K-----	-----
AGL86	-----	-----	-----	---PLAVAG	VGA-----	-----
AGL92	YEPVQPYVPT	GFSDHQYQN	MNFNQNQEP	VHYQALAVAG	AGLPMTQN--	-----
AGL35	-----	-----	-----	---VGFYDKVR	-----	-----
AGL80	-----	---IDLQQ	FRHPAAQH--	---VGLNEQPQ	N-----	-----
AGL46	-----	---MTPLL	IAEGSSSS-F	LNSPLFNSPQ	-----	-----
AGL45	-----	---HLSHNQ	YQYQQQFG--	---YPILVQDG	-----	-----
AGL48	PKY-----	---EIPQQ	MENGNPN---	---PKSYRLYQ	G-----	-----
AGL41	PKP-----	---VTYQQ	IGYETSN---	---RRGYNPYQ	G-----	-----
AGL96	-----	---WVFD	PQVQNP----	---CDVETHLP	T-----	-----
AGL87	-----	-----	-----	-----	-----	-----
AGL82	-----	---QIQMD	PNLMEKWTDL	ALTQSLMMSK	G-----	-----
AGL47	-----	---NTAEV	SSFSNVT---	---EPMIANGQ	S-----	-----
AGL40	-----	---LVREM	SQYSQVN---	---VSQNYFGQ	SS-----	-----
AGL62	-----	---VTVEA	SRFFQAN---	---VPNFYVGS	SSNNAAF---	-----
AGL23	-----	---VDEKA	IQLIHQT---	---NPNFYVGS	SSNAAAP---	-----
AGL28	-----	---VDEKV	AQLHHR---	---NLNFYVGS	SSNVAAP---	-----
AGL61	-----	---VVTNM	ASFNEAK---	---DDVFGFLD	NK-----	-----
AGL91	-----	---IQSNI	SHMQASS---	---SLMFLSND	N-----	-----
AGL29	-----	---IEGQV	NHLQASS---	---CLMLLSRK	-----	-----
AGL102	-----	---LKRKH	VEMEDLS---	---SPSILSK-	-----	-----
AGL58	-----	---LREEV	ALRIKRG---	---HDENNNK-	-----	-----
AGL59	-----	---LCEQA	ASRIKRG---	---HDENNNK-	-----	-----
AGL85	-----	---LCNEA	AARIORG---	---DAENNNK-	-----	-----
AGL88	-----	---LS-----	-----	-----	-----	-----
AGL57	-----	---LCDNM	AHLRLGK---	---DCGSSSS-	-----	-----
AGL84	TSA-----	---ICCV	DNSPENF---	---NEITEEQD	QIRSICETFC	VMDNN-----
AGL73	SSA-----	---ICCV	DKSPEIF---	---NEITEEQD	QILSICETFC	VTDNNNNNNN
AGL83	TTT-----	---ICCV	DELPANS---	---NEIVGISP	-----	---N-----
AGL56	-----	---D-----	---ETKT---	---NQTHEFEG	G-----	-----
AGL55	-----	-----	---KDT---	---NQITKLEG	-----	-----
AGL99	-----	---QNPVS	ASETVED---	---RELVVHKN	MD-----	-----
AGL39	-----	---LYQIC	GESSSQV---	---ASFQDNPS	SE-----	-----
AGL60	-----	-----	---WKL---	---NENLMARN	-----	-----

AGL100
AGL33

-----PTMD-----MQKLENLT-----

	425	435	445	455	465	475
AGL43	----LMHKE	FGCDHNMSV	GNINSNSCEH	-----	PCVSSTQHYS	AVEESVNNPW
AGL75	---NQLMYKE	FGCDQNMMS	GNINSNSFQN	-----	PCVSNTQHYS	AVEESVKNPW
AGL76	---NQLMLKE	LYGCDQNMCM	GNINSNSFQH	-----	PCVSKAQHYS	AVEGSVNNQR
AGL81	----GLMQQE	LCGYDQNMFM	NNN-NFQHSF	-----	VSNTQDHSAP	VVQESVNNN-
AGL98	----GLMQQE	LCNYDQNMFM	NNN-NFQHSF	-----	VSNTQDHSAP	AVQESVNNN-
AGL77	GLKNQLMKHD	LCGYEHNMC	SNHGDATAFSQ	-----	IPLSASNFNQ	DFSVSIQEE-
AGL78	----QQQDQ	LHGFDQNMCM	VSE----I--	-----	INNNNGLQHP	NLSNTVPHEF
AGL52	----QQQDQ	LHGFDQNLCM	MSE----I--	-----	INNNNGLQHP	NLSNTVPHEF
AGL51	----QQQ--	-----	-----	-----	LSTSWRVKHT	RIL-----
AGL103	----RAESY	SSLLGVHEDG	NNEFE----	-----	NPNMSSRRNF	NVQDCAGLLG
AGL93	----LTSTG	LLGEQGSGLG	SSFDLP-MVF	-----	PPQMOTQTP	VFFDQFAPWN
AGL53	----QSA	LLGEQGSGLG	SNFDLPPMVF	-----	PPQMOTQTP	VFFDQFAAWN
AGL54	-----	-----	-----	-----	-----	-----
AGL89	-----	-----	-----	-----	-----	-----
AGL26	-----	-----	-----	-----	-----	-----
AGL101	-----	LVKSISN-II	WKYKNFTILY	-----	NPEPCKKTKL	SVR-----
AGL50	-----D	YLQEAAAELY	QTYNLGNFCD	-----	-DHVWDLFEFA	SRLPPLHTFS
AGL49	-----N	YLQEAAAELY	QTYNLGNFCD	-----	-DHVWDLFEFA	SRLPPLHTFS
AGL36	-----E	FYNINLNLNL	NLNSNQYLNQ	QQSF----MN	PMVEQHMHV	GGRESIPFVD
AGL90	-----E	FYNINLNLNL	NLNSNQYLNQ	QQSF----MN	PMVEQHMHV	GGRESIPFVD
AGL34	-----E	FYNINLNQ--	--KSNQYMNQ	QQPF----MN	PMVEQHMSHV	GGRESIPFMD
AGL37	-----K	LHDFNMKMN-	-IDPNQ----	-----	-SMNLDLNDG	ED-EGIPCMD
AGL38	-----K	IHGFMNMN-	-KDSNQSMVL	-----	-DLNQNLNDG	ED-EGIPCMD
AGL86	-----	---APLAVAG	AGPPMAQNQY	EPIQPYIPTA	FSDNIQYQAP	VDFNHQIQHG
AGL92	-----QYEP	VHYQSLAVAG	GGLPMSQLQY	EPVQPYIPTV	FSDNVQYQHM	NLYQNQQEPV
AGL35	-----	-----	-----	-----	-----	-----
AGL80	-----L	NLNLNQNYN-	--QNQE-WFM	EMM-----	-NHPEQMRYQ	TEQMGYQFMD
AGL46	-----	LTN-----	-----	-----E	LQLIVSQNHR	LENSLASNLF
AGL45	-----	IYN-----	-----	-----	PSQIQNQHEE	WLDDHMMNHS
AGL48	-----S	SSNGNPHLEM	DPFRLQMMTS	QGLAGSVSQP	LQHHSMINNP	TMAMNQPSQD
AGL41	-----S	SSNGNPNFEM	MSVSPK----	-----	-----	-----
AGL96	-----	-----	-----	-----	MVSGLDLME	PSDEDLGTYS
AGL87	-----	-----	-----	-----	-----	-----
AGL82	-----ND	GTQFMQRQEQ	PYYNREQVVS	-----	RSAGFNVNPF	MGYQVPFNIP
AGL47	-----L	FGYSCSDGPY	GPMVQR----	-----	TAYMEPIHWG	LGNSMFNNVK
AGL40	-----GV	IGG--GNVGI	DLDFQR----	-----R	NAFNYPNMV	FPNHTPPMFG
AGL62	-----GI	DDGSHINPDM	DLFSQRRMMD	-----I	NAFNYNQNI	HPNHALPPFG
AGL23	-----AT	VSG-----	-----	-----	--GNISTNQG	FFDQNGMTTN
AGL28	-----AA	VSG-----	-----	-----	--GNISTNHG	FFDQNGNSTS
AGL61	-----	-----	-----	-----	---VT----	VPPYVNMPSG
AGL91	-----	-----	-----	-----	-----	-----
AGL29	-----	-----	-----	-----	-----	-----
AGL102	-----	-----	-----	-----	--NTKNKMMR	TEYSSGQSKG
AGL58	-----	-----	-----	-----	-----	-----
AGL59	-----	-----	-----	-----	-----	-----
AGL85	-----	-----	-----	-----	-----	-----
AGL88	-----	-----	-----	-----	-----	-----
AGL57	-----	-----	-----	-----	--VRVGRRVS	GGVRLFDREA
AGL84	-----AA	LPEMNLDY--	---DQDIGF	-----D	TPFESALNDW	FSDNTTHQEI
AGL73	N-----AA	LPEVNLYYNQ	DMAIDQLIDF	-----N	TPFESSIDDW	FSDNTTHQET
AGL83	-----	-PLIMLEK--	-----KKS	-----Q	IEEK-FEKEW	QVSVTRIENE
AGL56	-----	-----	-----	-----	--ETSGSASW	LENEDDILHF
AGL55	-----	-----	-----	-----	--ETSASASL	LKNVVDNLHI
AGL99	-----EDN	IHVSDMDDKD	TMLMISDKNN	-----VLP	ENLDE-FDQE	LDLDQLLDFE
AGL39	-----	-----	---IQG	-----	FETEEEINQI	NLLLQETQTE
AGL60	-----	-----	-----	-----	-----DRG	YGGNGDLEF
AGL100	-----	-----	-----	-----	-----DGV	NRCRVG----
AGL33	-----	-----	-----	-----	-----	-----

	485	495	505	515	525	535
AGL43	L-----	-----N	QLMQNELYGY	GRKVLDAASF	TSFESSWIIG	GLLIT-----
AGL75	L-----	-----N	QLMQNELYGY	G-----YAGF	C-----	-----
AGL76	Q-----	-----S	ELMQQELCGY	EQNMCFNNN	FQVSNKEAVQ	ESVTN-----
AGL81	-----	-----Y	GLMPHVPCGY	DQN--LFTSD	ITNNNLLINN	SMFL-----
AGL98	-----	-----Y	GLMPHVPCGY	DQN--LFTSD	ITNNNLLIDN	SMFL-----
AGL77	-----	-----S	GLMQQELCGY	DQNQNMSMGD	ITNNNFQVTC	ASVLESVNNF
AGL78	-----	-----S	SDFNQNPYGN	A-----VG-N	ISFSQDMFSS	YDASS-----
AGL52	-----	-----P---	-----	-----YG-N	TSFSQDMFSS	YDGSS-----
AGL51	-----	-----	-----	-----T	VSGTICSE	-----
AGL103	-----	-----	MQGAGTNGLO	-----SM-N	MHDYSNNNSI	NSNGLSHQYV
AGL93	-----	-----	-----	-----QA	PSFADPMMFP	YN-----
AGL53	-----	-----	-----	-----QA	PSFADPMMFP	YN-----
AGL54	-----	-----	-----	-----	-----	-----
AGL89	-----	-----	-----	-----	-----	-----
AGL26	-----	-----	-----	-----	-----	-----
AGL101	-----	-----	-----	-----	-----	-----
AGL50	-----	-----	-----	-----DPL	MTTNTCQTMS	TDMISI-----
AGL49	-----	-----	-----	-----DPL	MTTNTCQTMS	SDMISI-----
AGL36	-----	-----	GNCYNYHQLP	SNQLPAVDHA	STSYMPSTTG	VYDPY-----
AGL90	-----	-----	RNYYNYNQLP	-----AVDLA	STSYMPSTTD	VYDPY-----
AGL34	-----	-----	GNYYNYNQLP	-----VVDHG	STSYMPSTTG	VYDPY-----
AGL37	-----	-----	-----NNNYHP	-----EIDCL	ATVTT-APTD	VCAPN-----
AGL38	-----	-----	-----NNNYHP	-----EIDCL	ATVTT-APTD	VCAPN-----
AGL86	IYDNLSDPN	-----	-----HQY	PFQDDPFMEM	LMEYPYEQVG	YAAEHAHIPF
AGL92	HYQALGVAGA	GLPMNQNYE	PVQPYVPTGF	SDHFQFENMN	LNQOQEPVQ	YQAPVDFNHQ
AGL35	-----	-----	-----DQIQI	-----	TLNMKQTTND	LDLNK-----
AGL80	-----	-----	DNHHNHIIHQ	-----PQEHQ	HQIHDESSNA	LDAANSS---
AGL46	FS-----	-----	-----	-----EG	QDICIPDMNQ	VCML-----
AGL45	-----	-----	-----	-----KE	ISHPLMDDNN	FYYQQP----
AGL48	PFDYMRSELG	INEGININNS	QFYMSNNTIT	ANDGVRQEPY	PNVTTAGENN	GDATTS----
AGL41	-----	-----	-----	-----	-----	-----
AGL96	-----	-----	-----	-----GE	SSMAGGAEDD	AE-----
AGL87	-----	-----	-----	-----	-----	-----
AGL82	N-----	-----	-----	-----WRLSG	NQVENWELSG	KKTI-----
AGL47	-----	-----	-----	-----QFQD	YPRFAQVND	LEDSSKLSM-
AGL40	-----	-----	-----	-----YNDGVLPVI	SNNMYSSYN	FNQS-----
AGL62	NNA-----	-----	-----	-----YGINEGFVPE	YVNVFRPEYN	PNQNQIQNQN
AGL23	P-----	-----	-----	-----TQT	LLFGFDIMNR	TPGV-----
AGL28	-----	-----	-----	-----APT	LPFGFNVNMR	TPAGYNSYQL
AGL61	-----	-----	-----	-----PS	NIYNFANGNG	CF-----
AGL91	-----	-----	-----	-----	-----	-----
AGL29	-----	-----	-----	-----	-----	-----
AGL102	MYEFRAFPGG	FLGTI-----	-----	-----	-----	-----
AGL58	-----	-----	-----	-----	-----	-----
AGL59	-----	-----	-----	-----	-----	-----
AGL85	-----	-----	-----	-----	-----	-----
AGL88	-----	-----	-----	-----	-----	-----
AGL57	-----	-----	-----	-----	-----	-----
AGL84	-----	-----	-----	-----SAS	ILN--AVVDD	QVSVDLTPFS
AGL73	-----	-----	-----	-----TSA	SILNDVGVD	QVSVDTPPFS
AGL83	-----	-----	-----	-----ATS	SYAK-----R	RRSI-----
AGL56	-----	-----	-----	-----DDD	FYTG-----ID	PLF-----
AGL55	-----	-----	-----	-----DDR	YY-----	-----
AGL99	TNYESLLKSC	EMEDYASM--	-----	-----VTT	KQNLCSNPEA	VEDG-----
AGL39	ANVNLDDEIC	FWNDSND--	-----	-----DVF	GLN-----	-----S-----
AGL60	-----	-----	-----	-----AFM	PQNG-----R	Q-----
AGL100	-----	-----	-----	-----AFM	-QNG-----D	-----
AGL33	-----	-----	-----	-----	-----	-----

	545	555	565	575	585	595
AGL43	--SQYLLVSV	WYILQSQVME	IYPPEITVVF	FYNLCGMLIS	PPVCRKRLDL	LAA-----
AGL75	-----	-----	-----	-----	-----	-----
AGL76	--FGLMQHEL	YGCDQNMSG	NIINNSFQQR	LKHRTRICE-	-----	-----
AGL81	-----	-----	-----	-----	-----	-----
AGL98	-----	-----	-----	-----	-----	-----
AGL77	GLNQLMHKEF	YGCHQNMSG	NINNSFQHP	WVSNADHTRR	YKNL-----	-----
AGL78	-----	--LLQTSSLP	PLHNIPSSYC	FPGNSRLL--	-----	-----

Table 5. Type II MADS domain sequences

	← MADS Domain →					

	5	15	25	35	45	55
SEP1	MGRGRVELKR	IENKINRQVT	FAKRRNGLLK	KAYELSVLCD	AEVALIIFSN	RGKLYEFCSS
SEP2	MGRGRVELKR	IENKINRQVT	FAKRRNGLLK	KAYELSVLCD	AEVSLIVFSN	RGKLYEFCST
SEP3	MGRGRVELKR	IENKINRQVT	FAKRRNGLLK	KAYELSVLCD	AEVALIIFSN	RGKLYEFCSS
AGL3	MGRGKVELKR	IENKINRQVT	FAKRRNGLLK	KAYELSVLCD	AEIALLIFSN	RGKLYEFCSS
AGL13	MGRGKVEVKR	IENKINTRQVT	FSKRRSGLLK	KAYELSVLCD	AEVSLIIFST	GGKLYEFSNV
AGL6	MGRGRVEMKR	IENKINRQVT	FSKRRNGLLK	KAYELSVLCD	AEVALIIFSS	RGKLYEFGSV
CAL	MGRGRVELKR	IENKINRQVT	FSKRRTGLLK	KAQEISVLCD	AEVSLIVFSH	KGKLFEYSSE
AP1	MGRGRVQLKR	IENKINRQVT	FSKRRAGLLK	KAHEISVLCD	AEVALVVFSS	SGKLFYEYSTD
FUL	MGRGRVQLKR	IENKINRQVT	FSKRRSGLLK	KAHEISVLCD	AEVALIVFSS	KGKLFYEYSTD
AGL79	MGRGRVQLKR	IENKIRRQVT	FSKRRTGLVK	KAQEISVLCD	AEVALIVFSP	KGKLFEYSAG
FLF	MGRKKLEIKR	IENKSSRQVT	FSKRRNGLIE	KARQSVLCD	ASVALLVSSA	SGKLYFSFGS
SHP1	LGRGKIEIKR	IENTTNRQVT	FSKRRNGLLK	KAYELSVLCD	AEVALVIFST	RGRLYEYANN
SHP2	IGRGKIEIKR	IENTTNRQVT	FSKRRNGLLK	KAYELSVLCD	AEVALVIFST	RGRLYEYANN
AG	SGRGKIEIKR	IENTTNRQVT	FSKRRNGLLK	KAYELSVLCD	AEVALVIFSS	RGRLYEYANN
AGL11	MGRGKIEIKR	IENSTNRQVT	FSKRRNGLLK	KAYELSVLCD	AEVALVIFST	RGRLYEYANN
AGL12	MARGKIQIKR	IENPVHRQVT	FSKRRTGLLK	KAYELSVLCD	AEIGVVFSS	QGKLFELATK
AGL70	MGRRKVEIKR	IENKSSRQVT	FSKRRKGLIE	KARQSLILCE	SSIAVVAVSG	SGKLYDSASG
MAF2I	MGRKKVEIKR	IENKSSRQVT	FSKRRNGLIE	KARQSLILCE	SSIAVLLVSS	SGKLYKSASG
MAF1	MGRKKIEIKR	IENKSSRQVT	FSKRRNGLID	KARQSLILCE	SSVAVVVVSA	SGKLYDSSSG
MAF5I	MGRRRVEIKR	IENKSSRQVT	FSKRRNGLME	KARQSLILCE	SSVALPIVSS	TGKLYNSSSG
MAF4I	MGRKKVEIKR	IENKSSRQVT	FSKRRNGLME	KARQSLILCE	SSVALIISA	TGRLYSFGSS
AGL17	MGRGKIVIQK	IDDSTSRQVT	FSKRRKGLIK	KAKELAILCD	AEVCLIIIFSN	TDKLYDFASS
AGL21	MGRGKIVIQK	IDDSTSRQVT	FSKRRKGLIK	KAKELAILCD	AEVGLIIFSS	TGKLYDFASS
ANR1	MGRGKIVIRR	IDNSTSRQVT	FSKRRSGLLK	KAKELAILCD	AEVGVIIIFSS	TGKLYDYASN
AGL16	MGRGKIAIKR	INNSTRQVT	FSKRRNGLLK	KAKELAILCD	AEVGVIIIFSS	TGRLYDFSSS
AGL15	MGRGKIEIKR	IENANSRQVT	FSKRRSGLLK	KARELSVLCD	AEVAVIVFSK	SGKLFEYSST
AGL18	MGRGRIEIKK	IENINSRQVT	FSKRRNGLIK	KAKELAILCD	AEVALIIFSS	TGKIYDFSSV
SVP	MAREKIQIRK	IDNATARQVT	FSKRRRGLFK	KAEELSVLCD	ADVALIIFSS	TGKLFECSS
AGL24	MAREKIRIKK	IDNITARQVT	FSKRRRGIFK	KADELSVLCD	ADVALIIFSA	TGKLFECSSS
AGL19	MVRGKTEMKR	IENATSRQVT	FSKRRNGLLK	KAFELSVLCD	AEVALVIFSP	RSKLYEFSSS
AGL14	MVRGKTEMKR	IENATSRQVT	FSKRRNGLLK	KAFELSVLCD	AEVALIIFSP	RGKLYEFSSS
SOC1	MVRGKTQMKR	IENATSRQVT	FSKRRNGLLK	KAFELSVLCD	AEVSLIIFSP	KGKLYEFASS
AGL72	MVRGKIEIKK	IENVTSRQVT	FSKRRSGLFK	KAHELTVLCD	AQVAAMIFSQ	KGRLYEFASS
AGL71	MVRGKIEIKK	IENVTSRQVT	FSKRRSGLFK	KAHELTVLCD	AQVAIVFSQ	SGRLHEYSSS
AGL42	MVRGKIEMKK	IENATSRQVT	FSKRRNGLLK	KAYELSVLCD	AQLSLIIFSQ	RGRLYEFSSS
PI	MGRGKIEIKR	IENANNRVVT	FSKRRNGLVK	KAKEITVLCD	AKVALIIFAS	NGKMIDYCCP
AP3	MARGKIQIKR	IENQTNRQVT	YSKRRNGLFK	KAHELTVLCD	ARVSIIMFSS	SNKLEHYISP
TT16	MGRGKIEIKK	IENQTNRQVT	FSKRRTGLIK	KTRELSILCD	AHIGLIVFSA	TGKLEFCSE
AGL63	MRGKRVVIKK	IEEKIKRQVT	FAKRRKSLIK	KAYELSVLCD	VHLGLIIFSH	SNRLYDFCSN
AGL104	MGRVKLEIKR	IENTTNRQVT	FSKRRNGLIK	KAYELSVLCD	IDIALIMFSP	SDRLSLFSGK
AGL66	MGRVKLEIKR	IENTTNRQVT	FSKRRNGLIK	KAYELSVLCD	IDIALLMFSP	SDRLSLFSGK
AGL67	MGRVKLEIKR	IEKSTNRQIT	FSKRRKGLIK	KAYELSTLCD	IDLALLMFSP	SDRLCLFSGQ
AGL94	MGRVKLIKIK	LQNMNGRQCT	YTKRRHGIMK	KAKELSVLCD	IDVLLMFSP	MGKASICIGK
AGL30	MGRVKLIKIK	LENTNGRQST	FAKRRKNGILK	KANELSVLCD	IDIVLLMFSP	TGKAAICCGT
AGL65	MGRVKLIKIKR	LESTSNRQVT	YTKRRKNGILK	KAKELSVLCD	IDIVLLMFSP	TGRATAFHGE

	← I Region →					

	65	75	85	95	105	115
SEP1	SN-----	-----M	LKTLDRYQK-	--CSYGSIEV	NNKPAKEL--	-----
SEP2	SN-----	-----M	LKTLERYQK-	--CSYGSIEV	NNKPAKEL--	-----
SEP3	SS-----	-----M	LRTLERYQK-	--CNYGAPEP	NV-PSREALA	VE-----
AGL3	PS-----	-G-----	ARTVDKYRK-	--HSYATMDP	NQ-SAKDL--	-----
AGL13	G-----	-----V	GRTIERYR-	--CKDNLDD-	ND-TLEDT--	-----
AGL6	G-----	-----I	ESTIERYNR-	--C-YNCSLS	NNKPEETT--	-----
CAL	SC-----	-----M	EKVLERYER-	--YSYAERQL	IA-PDSHVNA	Q-----
AP1	SC-----	-----M	EKILERYER-	--YSYAERQL	IA-PESDVN-	-----

FUL	SC-----	-----M	ERILERYDR-	--YLYSDKQL	VGR--DVSQS	-----
AGL79	SS-----	-----M	ERILDYER-	--SAYAG-QD	IPTPNLDSQ-	-----
FLF	DN-----	-----L	VKILDYRGK-	--QHADDLKA	L-----	-----
SHP1	S-----	-----V	RGTIERYKK-	--ACSDAVN-	PPSVTEANT-	-----
SHP2	S-----	-----V	RGTIERYKK-	--ACSDAVN-	PPTITEANT-	-----
AG	S-----	-----V	KGTIERYKK-	--AISDNSN-	TGSSVAEINA-	-----
AGL11	N-----	-----I	RSTIERYKK-	--ACSDSTN-	TSTVQEINA-	-----
AGL12	GT-----	-----M	EGMIDKYMK-	--CTGGGRG	SSSATFTAQE	QLQPPNL---
AGL70	DN-----	-----M	SKIIDRYEI-	--HHADELKA	L-----	-----
MAF2I	DN-----	-----M	SKIIDRYEI-	--HHADELEA	L-----	-----
MAF1	DD-----	-----I	SKIIDRYEI-	--QHADELRA	L-----	-----
MAF5I	DS-----	-----M	AKIISRFKI-	--QQADDPET	L-----	-----
MAF4I	DS-----	-----M	AKILSRYEL-	--EQADDLKT	L-----	-----
AGL17	S-----	-----V	KSTIERFNT-	--AKMEEQEL	MNPASEV---	-----
AGL21	S-----	-----M	KSVIDRYNK-	--SKIEQQQL	LNPASEV---	-----
ANR1	SS-----	-----M	KTIERYNR-	--VKEEQHQL	LNHASEI---	-----
AGL16	S-----	-----M	KSVIERYS-	--AKGETSSE	NDPASEI---	-----
AGL15	G-----	-----M	KQTLSTRYGN-	--HQSSSASK	A-----	-----
AGL18	C-----	-----M	EQILSRYGY-	--TTASTEHK	QOREHQLLIC	ASHGNEA---
SVP	S-----	-----M	KEVLERHNL-	--QSKNLEKL	DQPSLELQ--	-----
AGL24	R-----	-----M	RDILGRYSL-	--HASNINKL	MDPPSTHL--	-----
AGL19	S-----	-----I	AATIERYQR-	--RIKEIGNN	HKRNDNS---	-----
AGL14	SS-----	-----I	PKTVERYQK-	--RIQDLGSN	HKRNDNS---	-----
SOC1	N-----	-----M	QDTIDRYLR-	--HTKDRVST	KPVSEENM---	-----
AGL72	D-----	-----I	RNTIKRYAE-	--YKREYFVA	ETHPIEQYV-	-----
AGL71	Q-----	-----M	EKIIDRYGK-	--FSNAFYVA	ERPQVERYL-	-----
AGL42	D-----	-----M	QKTIERYRK-	--YTKDHETS	NHDSQIHL--	-----
PI	SM-----	-D-----	L	GAMLDQYQK-	--LS-G-KKL	WDAKH-----
AP3	NT-----	-T-----	T	KEIVDLYQT-	--IS-D-VDV	WATQY-----
TT16	QN-----	-R-----	M	PQLIDRYLH-	--TN-G-LRL	-PDHDDQ-----
AGL63	ST-----	-S-----	M	ENLMRYQK-	--EKEGQTTA	EHSFHSQCS-----
AGL104	TR-----	-----I	EDVFSRFINL	PKQERESALY	FPDQNRPPDI	QNK-----
AGL66	TRFFVTFSLD	TSIFLIKNER	SKLTFQGFNR	FSKDIP-LIL	FPDQSRPPI	SRAKRYVIRL
AGL67	TR-----	-----I	EDVLARYINL	PDQERENAIV	FPDQSKRQGI	QNK-----
AGL94	HS-----	-----I	GEVIAKFAQL	SPQERAKRKL	ENLEALRKT	MKANHD-----
AGL30	RR-----	-CFSFESSEL	EENFPKVGSR	CKYTRIYSLK	DLS-----	-----
AGL65	HR-----	-YNYQNHSYA	LKKTFFKLDH	DVNIHDFLGA	RNQTIE----	-----

	K Domain						
	
	125	135	145	155	165	175	
SEP1	-----	ENSY	REYLKLGKRY	EN-LQRQQRN	LLGEDLGPLN	SKELEQLERQ	LDGSLKQVRS
SEP2	-----	ENSY	REYLKLGKRY	EN-LQRQQRN	LLGEDLGPLN	SKELEQLERQ	LDGSLKQVRC
SEP3	-----	LSSQ	QEYLKLKERY	DA-LQRTQRN	LLGEDLGPLS	TKEESLERQ	LDSSLKQIRA
AGL3	-----	QDKY	QDYKLLKSRV	EI-LQHSQRH	LLGEELESEM	VNELEHLERQ	VDAASLRQIRS
AGL13	-----	QGLR	QEVTKLKCKY	ES-LLRTHRN	LVGEDLEGMS	IKELQTLERQ	LEGALSATRK
AGL6	-----	QSWC	QEVTKLKSKY	ES-LVRTNRN	LLGEDLGEMG	VKELQALERQ	LEAALTATRQ
CAL	-----	TNWS	MEYSRLKAKI	EL-LERNQRH	YLGEELPMS	LKDLQNLQEQ	LETALKHIRS
AP1	-----	TNWS	MEYNRLKAKI	EL-LERNQRH	YLGEDLQAMS	PKELQNLQEQ	LDTALKHIRT
FUL	-----	ENWV	LEHAKLKARV	EV-LEKNRNR	FMGEDLDLSL	LKELQSLEHQ	LDAAIKSIRS
AGL79	-----	GECS	TECSKLLRMI	DV-LQRSLRH	LRGEEVDGLS	IRDLQGVEMQ	LDTALKKTRS
FLF	-----	DHQ	SKALNYGSHY	EL-LELVDSK	LVGSNVKNVS	IDALVQLEEH	LETALSVTRA
SHP1	-----	QYYQ	QEASKLRRQI	RD-IQNSNRH	IVGESLGLSN	FKELKNLEGR	LEKGISRVRS
SHP2	-----	QYYQ	QEASKLRRQI	RD-IQNLNRH	ILGESLGLSN	FKELKNLESR	LEKGISRVRS
AG	-----	QYYQ	QESAKLRQOI	IS-IQNSNRQ	LMGETIGSMS	PKELRNLEGR	LEKISITRIRS
AGL11	-----	AYYQ	QESAKLRQOI	QT-IQNSNRN	LMGDSLSSLS	VKELKQVENR	LEKALSRIRS
AGL12	-----	DPK	DEINVLKQEI	EM-LQKGISY	MFGGGDGAMN	LEELLLLEKH	LEYWISQIRS
AGL70	-----	DLA	EKIRNYLPHK	EL-LEIVQSK	LEESNVDNVS	VDSLISMEEQ	LETALSVIRA
MAF2I	-----	DLA	EKTRNYLPLK	EL-LEIVQSK	LEESNVDNAS	VDTLISLEEQ	LETALSVTRA
MAF1	-----	DLE	EKIQNYLPHK	EL-LETVQSK	LEEPNVDNVS	VDSLISLEEQ	LETALSVSRA

MAF51	-----DLE	DKTQDYLSHK	EL-LEIVQRK	IEEAKGDNVS	IESLISMEEQ	LKSALSVIRA
MAF41	-----DLE	EKTLNLYLSHK	EL-LETIQCK	IEEAKSDNVS	IDCLKSLEEQ	LKTALSVTRA
AGL17	-----KFWQ	REAETLRQEL	HS-LQENYRQ	LTGVELNGLS	VKELQNISSQ	LEMSLRGIRM
AGL21	-----KFWQ	REAAVLRQEL	HA-LQENHRQ	MMGEQLNGLS	VNELNSLENQ	IEISLRGIRM
ANR1	-----KFWQ	REVASLQQQL	QY-LQECHRK	LVGEELSGMN	ANDLQNLQEDQ	LVTSLKGVRL
AGL16	-----QEMY	IVTLEKYAYS	EE-L-VLDRQ	MMGEELSGLS	VEALQNLQENQ	LELSLRGVRM
AGL15	-----EEDC	AEVDILKDQL	SK-LQEKHLQ	LQGGKLNPLT	FKELQSLEQQ	LYHALITVRE
AGL18	-----VLRN	DDSMK--GEL	ER-LQLAIER	LKGKELEGMS	FPDLISLENQ	LNESLHSVKD
SVP	-----LVEN	SDHARMSKEI	AD-KSHRLRQ	MRGEELQGLD	IEELQOLEKA	LETGLTRVIE
AGL24	-----RLEN	CNLSRLSKEV	ED-KTKQLRK	LRGEDLDGLN	LEELQRLEKL	LESGLSRVSE
AGL19	-----QQAR	DETSGLTKKI	EQ-LEISKRK	LLGEGIDACS	IEELQOLENQ	LDRSLSRIRA
AGL14	-----QSK	DETYGLARKI	EH-LEISTRK	MMGEGLDASS	IEELQOLENQ	LDRSLMKIRA
SOC1	-----QHLK	YEAANMMKKI	EQ-LEASKRK	LLGEGIGTCS	IEELQOIEQQ	LEKSVKCIRA
AGL72	-----QGLK	KEMVTMVKKI	EV-LEVHNRK	MMGQSLDSCS	VKELSEIATQ	IEKSLHMVRL
AGL71	-----QELK	MEIDRMVKKI	DL-LEFVHRK	LLGQGLDSCS	VTELQEIDTQ	IEKSLRIVRS
AGL42	-----QQLK	QEASHMITKI	EL-LEFHKRK	LLGQGIASCS	LEELQEIDSQ	LQRSLG----
PI	-----ENLS	NEIDRIKKEN	DS-LQLELRH	LKGEDIQSLN	LKNLMAVEHA	IEHGLDKVRD
AP3	-----ERMQ	ETKRKLEETN	RN-LRTQIKQ	RLGECLDELD	IQELRRLEDE	MENTFKLVRE
TT16	-----EQLH	HEMELLRRET	CN-LELRLRP	PHGHDLASIP	PNELDGLERQ	LEHSVLKVRE
AGL63	-----DCVK	-TKESMMREI	EN-LKLNQL	YDGHGLNLLT	YDELLSFEHL	LESSLQHARA
AGL104	-----ECLL	RILQQLKTEN	DIALQVTNPA	AINSDVEELE	HE-VCRLOQQ	LQ--MA----
AGL66	SILCMCHYLL	RTLQQLKAEN	DIALQLLYNY	VICNNSEELE	HE-VYKLQQQ	LL--MA----
AGL67	-----EYLL	RTLEKLIKIED	DMALQINIDF	GMEMEQULEN	FS-WVRTDEN	MN--VPI----
AGL94	-----IDIS	KFLDRISTPT	VE-VCTTTMM	LINDSEVSSN	SFRDTGIFRC	LAKKSDSY--
AGL30	-----TQAR	ILQARISEIH	GR-LSYWTPE	DKINNVEHLG	QLEIS-IRQS	LDQ-LRAH--
AGL65	-----VWID	HL-RFMNFLG	YFLLSCWTNI	DRIENTEHL	LLEES-LRKS	IER-IQIH--

	C terminal region					
	
	185	195	205	215	225	235
SEP1	IKTQYMLDQL	SDLQNKQML	LETNRALAMK	LD-----	-----D--	-MI----GVR
SEP2	IKTQYMLDQL	SDLQKKEHIL	LDANRALSMK	LE-----	-----D--	-MI----GVR
SEP3	LRTQFMLDQL	NDLQSKERML	TETNKTLLRL	LA-----	-----D-G	YQM---P-L
AGL3	TKARSMLDQL	SDLKTKEEML	LETNRDLRRK	LE-----	-----D-S	DAAL---T--
AGL13	QKTQVMMEQM	EELRRKEREL	GDINNKLKLE	TE-----	-----D--H	DFK----
AGL6	RKTQVMMEEM	EDLRKKERQL	GDINKQLKIK	FE-----	-----TEGH	AFK---T--
CAL	RKNQLMNESE	NHLQRKEKEI	QEENSMITKQ	IK-----	-----E--	--R---E-N
AP1	RKNQLMYESI	NELQKKEKAI	QEQNSMLSKQ	IK-----	-----E--	--R---E-K
FUL	RKNQAMFESI	SALQKKDKAL	QDHNNSLK	IK-----	-----E--R	EKK---T--
AGL79	RKNQLMVESI	AQLQKKEKEL	KELKKQLTKK	VK-----	-----A-G	ER-----E--
FLF	KKTELMLKLV	ENLKEKEKML	KEENQVLASQ	ME-----	-----	-----N--
SHP1	KKNELLVAEI	EYMOKREMEL	QHNNMYLRAK	IA-----	-----EG	ARL---NPD
SHP2	KKHEMLVAEI	EYMOKREIEL	QNDNMYLRSK	IT-----	-----ER	TGL---Q--
AG	KKNELLFSEI	DYMOKREVDL	HNDNQILRAK	IA-----	-----EN	ERN---N--
AGL11	KKHELLLVEI	ENAQKREIEL	DNENIYLRTK	VA-----	-----EV	ER-----
AGL12	AKMDVMLQEI	QSLRNKEGVL	KNTNKYLLDK	IE-----	-----NN	-----
AGL70	KKTELMMEDM	KSLQEREKLL	IEENQILASQ	VG-----	-----	-----K--
MAF2I	RKTELMMGEV	KSLQKTENLL	REENQTLASQ	VG-----	-----	-----K--
MAF1	RKAELMMEYI	ESLKEKEKLL	REENQVLASQ	MG-----	-----	-----K--
MAF5I	RKTELLMELV	KNLQDKKELL	KEKNQVLASE	VG-----	-----	-KL---K--
MAF4I	RKTELLMELV	KTHQEKEKLL	REENQSLTNQ	LI-----	-----KM	GKM---K--
AGL17	KREQILTNEI	KELTRKRNLV	HHENLELSRK	VQ-----	-----RI	-----
AGL21	RKEQLLTQEI	QELSQRNLI	HQENLDLSRK	VQ-----	-----RI	-----
ANR1	KKDQMLTNEI	RELNRKGQII	QKENHELQNI	VD-----	-----IM	RKE---N--
AGL16	KKDQMLIEEI	QVLNREGNLV	HQENLDLHKK	VN-----	-----LM	-----
AGL15	RKERLLTNQL	EESRLKEQRA	ELENETLRRQ	VQ-----	-----EL	RSFL-P----
AGL18	QKTQILLNQI	ERSRIQEKA	LEENQILRQ	V-----	-----	-EML-G----
SVP	TKSDKIMSEI	SELQKGMQL	MDENKRLRQ	VC-----	-----VL	-----
AGL24	KKGECVMSQI	FSLEKRGSEL	VDENKRLRDK	LE-----	-----TL	-----


AGL19	KKYQLLREEI	EKLKAEERNL	VKENKDLKEK	WL-----	-----GM	-----
AGL14	KKYQLLREET	EKLKEKERNL	IAENKMLMEK	CE-----	-----MQ	-----
SOC1	RKTQVFKEQI	EQLKQKEKAL	AAENKLSSEK	WG-----	-----SH	-----
AGL72	RKAKLYEDEL	QKLKAKEREL	KDERVRLSLK	KT-----	-----IY	-----
AGL71	RKAELYADQL	KKLKEKEREL	LNE--RKRLLE	EE-----	-----VN	-----
AGL42	-KALFKKEQL	EKLKAKEKQL	LEENVKLHQK	NV-----	-----IN	-----
PI	HQMEILISK-	---RRNEKMM	AEEQRQLTFQ	LQ-----	-----Q-	-----
AP3	RKFKSLGNQI	ETTKKKKNSQ	QDIQKNIHIE	LE-----	-----L-	-----R-
TT16	RKNELMQQQ	ENLSRKRRL	EEDNNMYRW	LH-----	-----EH	RAAM-EFQQ-
AGL63	RKSEFMHQQQ	QQ-QTDQKLK	GKEKGQSSW	EQ-----	-----L	M-----
AGL104	-----EEL	RRYEPDP--I	RFTTMEEYEV	SE--KQLLDT	LTHVVQRRDH	LMSN-HLSSY
AGL66	-----EEL	MKYEPDP--I	RFTTMEEYET	CE--KQLMDT	LTRVNQRREH	ILSQDQLSSY
AGL67	-----EED	PNLQLHH--M	YKDITCSASS	AL-----	-----GNV	SGLF-SKSSD
AGL94	-----WTDV	DNIDSVD--V	LQQLEHSLRQ	SLAQIYGRKA	SMPQRQQQQL	MSSQ-CKNQT
AGL30	-----KMQ	DGI-QIP--L	EQQLQSMWI	LN-----	-----	-----
AGL65	-----KEH	YRK-NQL--L	PIECAFHSGI	QLPMFGSYPG	YFGT-----	-----

	245	255	265	275	285	295
SEP1	SHHMGG--WE	GGEQ-----	-NVTYAHHQ	Q-SQGL-YQP	-----	LECNPTLQMG
SEP2	HHHIGGG-WE	GGDQQ-----	-NIAYGHPQA	H-SQGL-YQS	-----	LECDPTLQIG
SEP3	QLNPNQEEVD	HYGRH-----	-HHQQQHH--	--SQAF-FQP	-----	LECEPILQIG
AGL3	---QSFWG	SSAAEQQQQH	QQQQQGMSSY	Q-SNPP-IQE	A-----	GFFKP-LQ-G
AGL13	--GF-QDLLL	NPVLTAG---	--CSTDFSL-	Q-STHONYIS	D-----	CNLGYFYRLG
AGL6	---F-QDLWA	NSAASVAG--	DPNNSFPVE	P-SHPN--VL	D-----	CNTEPFLQIG
CAL	ILKTKQTQCE	---QLNRSV	DDVP-QPQPF	Q--HPLYMI	A-----	HQTSPLNMG
AP1	ILRAQQEQWD	-QQNQGHMNP	PPLPPQHQI	Q--HP--YML	S-----	HQPSPLNMG
FUL	--G--QEGQ	LVQCSNSS--	-SVL-LPQYC	VTS-----	-----	SRDGFVERVG
AGL79	--DF-QTQNL	SHDLASLATP	PFESPHELRR	TIS-P-----	-----	---PPPPLSSG
FLF	--N--HHVGA	EAEMEMS---	---PAGQIS	DNLP-----	-----	---VTLPLLN--
SHP1	QQE--SSVIQ	GTTVYES---	---GVSSHD	QSQH-----	-----	---YNNRYIPVN
SHP2	QQE--SSVIH	QGTVYES---	---GVTSSH	QSGQ-----	-----	---YNNRYIAVN
AG	-PS--ISLMP	GGSNYEQ---	---LMPPPQ	TQSQPF-----	-----	---DSRNYFQVA
AGL11	--Y--QQHHH	QMVSGSE---	---INAIEA	LASRN-----	-----	---YFAHSIMTAG
AGL12	-N--SILDA	NFAVMET---	---NYSYP	LTMP-----	-----	---SEIFQF--
AGL70	--K--TFLVI	EGDRGMS---	---RENGSG	NKVP-----	-----	---ETLSLLK--
MAF2I	--K--TFLVI	EGDRGMS---	---WENGSG	NKVR-----	-----	---ETLPLLK--
MAF1	--N--TLLAT	DDERGMF---	---PGSSSG	NKIP-----	-----	---QLTPLLN--
MAF5I	--K--ILETG	DERAVMS---	---PENSSG	HSPP-----	-----	---ETLPLLK--
MAF4I	--K--SVEAE	DARAMS---	---PESSSD	NKPP-----	-----	---ETLPLLK--
AGL17	--H--QENVE	LYKKAYG---	---TSNTNG	LGHH-----	ELVDAYVE	SHAQVRLQLS
AGL21	--H--QENVE	LYKKAYG---	---MANTNG	FTHR-----	EVAVADDE	SHTQIRLQLS
ANR1	IKL--QKKVH	GRTNAIEG--	---NSSVDP	ISNG-----	TT	TYAPPQLQLI
AGL16	--H--QQNME	LHEKVSE---	---VEGVKI	ANKNSLLT--	NGLDMRDT	SNEHVHLQLS
AGL15	-SF--THYVP	SYIKCFAID--	---PKNALI	NHDSKCS---	-----LQN	TDSDTLQLG
AGL18	-RG--SGPKV	LNERPDSS---	---P-----	ADPE-----	SSS	SEEDENDNEE
SVP	-P--SLLIT	NPFLST---	---INVHTP	KFNP-----	-----	---QLSTHMFHD
AGL24	--E--RAKLT	TLKEALE---	---TESVTT	NVSS-----	-----	---YDSGTFLEDD
AGL19	--G--TATIA	SSQSTLS---	---SSEVNI	DDN-----	-----	---MEVETGLFIG
AGL14	-G--RIIG	RISSSSS---	---TSELDI	DDNE-----	-----	---MEVVTDLFIG
SOC1	--E--SEVWS	NKNQEST---	---GRGDEE	SSPS-----	-----	---SEVETQLFIG
AGL72	--T--HLCQV	GERPMGM---	---PSG---	SKEK-----	-----	---EDVETDLFIG
AGL71	--M--HSSK	GNTGEGH---	---R-----	TKHS-----	-----	---SEVETDLFIG
AGL42	--P--WRGSS	TDQQQEK---	---YKV---	IDLN-----	-----	---LEVETDLFIG
PI	-----E	MAIASNARGM	MMRDHDGQ--	-----	-----	---FGYR
AP3	-----AEDP	HYGLVDNG--	-G-----DYDS	V-LGYQIE--	-----	---GSRAYALR
TT16	-----A	GIDTKPEYQ	QFIEQLQCY-	-----	-----	---KP
AGL63	--W--QAERQ	MMTCQRQ---	---KDPAPA	NEGG-----	-----	---VPFLRWG

AGL104	EAS--TMQPN	IGGPFVNDVV	EGWLPENGTN	QTHLFDASAH	SNQLRELSSA	MYEPLLQSS
AGL66	EASALQQQS	MGGPFVNDVV	GGWLTENGN	EAHLFDASAH	SAMYET----	----LLQSS
AGL67	ILQ--KLETG	SIPGTSADPN	QQFSNLSFLN	DQKQLQAEW	NLLG-SPADY	YVSQILEASY
AGL94	EID--AMGGN	SSMQEAHMS	WLPDNDHQQT	ILPGDSSFLP	HREMDGSIPV	YSSCFFESTK
AGL30	--S--NTTNI	VTEEHNS---	--IPQREVEC	SAS-----	-----	SSEPRFIPH-
AGL65	-----GK	SPEMTIPQE	TSF-LDELNT	GQ--LKQDTS	SQQQFTNNNN	ITAYNPNLHN

	305	315	325	335	345	355
SEP1	YDNPVCS-E-	--QITATTQA	Q-----	----AQPNG	YIP-GWML--	-----
SEP2	YSHPVCS-E-	--QMAVTVOG	Q-----	----SQQGNG	YIP-GWML--	-----
SEP3	YQGQDQG-M-	--GAGPSV--	-----	-----NN	YML-GWLPYD	TNSI-----
AGL3	NVALQMSSHY	N----HNPA	N-----	ATN	SATTSQNVNG	FFP-GWMV--
AGL13	FNN-----	--TMSKVKDL	R-----	-----	-----	-----
AGL6	FQQ-----	--HYVQEGEG	S-----	SV	SKSNVAGETN	FVQ-GWVL--
CAL	GLYQ-----	--GEDQTAMR	R-----	NN	LDLTLEPIYN	Y-L-GCYAA-
AP1	GLYQ-----	--EDDPMAMR	-----	ND	LELTLEPVYN	CNL-GCFAA-
FUL	GENGG-----	--ASSLTEPN	S-----	-----	LLPA--	----WMLRP
AGL79	DTS-----	----QRDGV	G-----	EV	AAGTLIRRTN	ATLPHWMPQL
FLF	-----	-----	-----	-----	-----	-----
SHP1	LLEPN-----	----QFSG	Q-----	DQ	PP-----	LQ
SHP2	LLEPN-----	----QNSSN	Q-----	DQ	PP-----	LQ
AG	ALQPNN-----	--HHYSSAGR	Q-----	DQ	TA-----	LQ
AGL11	SGS-----	----GNGGS	YS-----	DP	DK-----	KI
AGL12	-----	-----	-----	-----	-----	-----
AGL70	-----	-----	-----	-----	-----	-----
MAF2I	-----	-----	-----	-----	-----	-----
MAF1	-----	-----	-----	-----	-----	-----
MAF5I	-----	-----	-----	-----	-----	-----
MAF4I	-----	-----	-----	-----	-----	-----
AGL17	QPE-----	----QSHYK	T-----	SS	NS-----	-----
AGL21	QPE-----	----HSDYD	T-----	PP	RANE-----	-----
ANR1	QLQP-----	----APREK	S-----	-----	IRLG	LQLS-----
AGL16	QP-----	----QHDHE	T-----	HS	KAIQLNYFSF	IA-----
AGL15	LPGEAHD-RR	TNEGERESPS	SDSVTTNTSS	ETAERGDQSS	LANSPPPEAKR	QRFSV-----
AGL18	HHSDSL-QL	G---LSSTG	YCTKRKKPKI	ELVC-----	DNSGSQVASD	-----
SVP	TV-----	----R--	-----	-----	-----	-----
AGL24	SD-----	----TSLK	L-----	GL	PSWE-----	-----
AGL19	PPE-----	----TRQS	K-----	KF	PPQN-----	-----
AGL14	PPE-----	----TRHF	K-----	KF	PPSN-----	-----
SOC1	LPC-----	----SSRK-	-----	-----	-----	-----
AGL72	FLK-----	----NRP-	-----	-----	-----	-----
AGL71	LPV-----	----TRL-	-----	-----	-----	-----
AGL42	LP-----	----NRNC	-----	-----	-----	-----
PI	VQPI-----	----QPN	-----	LQ	-----	EKIMS
AP3	FHQ-----	----HHHYYPN	-----	HGLH	APSASDIITF	HLE-----
TT16	--GE-----	----YQQ	-----	FL	-----	EQQQQ
AGL63	TTHR-----	----RSSPP	-----	-----	EQQQQ	QPNSVLQLAT
AGL104	SSSNQNN-MS	ECHVTNHNGE	MFPEWAQAYS	SSALFASMQQ	QHEGVGPSIE	EMMPAQQSDI
AGL66	SSSNQNNIMG	ESNVSNHNGD	MFQEWQAQYN	STAHNPSTL	FPPMQHQHGL	VVDPNIEEIE
AGL67	KPQIGGK-NN	GASSETLPYV	AVFDDPLYFW	VNNGLFIIHL	FSKLCYWSC	FADCF-----
AGL94	PEDQICS-NP	GQQFEQLEQQ	GNGCLGLQQL	GEEYSYPTPF	GTTLMMEEDQ	EKKIKSEMEL
AGL30	-----	-----	-----	-----	-----	-----
AGL65	DMNHHTLPP	PPLPLTLPHA	QVIYIPMQR	YHMNGFFEAP	PPDSSAYNDN	TNQTRFGSSS

	365	375	385	395	405	415
SEP1	-----	-----	-----	-----	-----	-----
SEP2	-----	-----	-----	-----	-----	-----
SEP3	-----	-----	-----	-----	-----	-----
AGL3	-----	-----	-----	-----	-----	-----
AGL13	-----	-----	-----	-----	-----	-----
AGL6	-----	-----	-----	-----	-----	-----
CAL	-----	-----	-----	-----	-----	-----
AP1	-----	-----	-----	-----	-----	-----
FUL	-----	-----	-----	-----	-----	-----
AGL79	-----	-----	-----	-----	-----	-----
FLF	-----	-----	-----	-----	-----	-----
SHP1	-----	-----	-----	-----	-----	-----
SHP2	-----	-----	-----	-----	-----	-----
AG	-----	-----	-----	-----	-----	-----
AGL11	-----	-----	-----	-----	-----	-----
AGL12	-----	-----	-----	-----	-----	-----
AGL70	-----	-----	-----	-----	-----	-----
MAF2I	-----	-----	-----	-----	-----	-----
MAF1	-----	-----	-----	-----	-----	-----
MAF5I	-----	-----	-----	-----	-----	-----
MAF4I	-----	-----	-----	-----	-----	-----
AGL17	-----	-----	-----	-----	-----	-----
AGL21	-----	-----	-----	-----	-----	-----
ANR1	-----	-----	-----	-----	-----	-----
AGL16	-----	-----	-----	-----	-----	-----
AGL15	-----	-----	-----	-----	-----	-----
AGL18	-----	-----	-----	-----	-----	-----
SVP	-----	-----	-----	-----	-----	-----
AGL24	-----	-----	-----	-----	-----	-----
AGL19	-----	-----	-----	-----	-----	-----
AGL14	-----	-----	-----	-----	-----	-----
SOC1	-----	-----	-----	-----	-----	-----
AGL72	-----	-----	-----	-----	-----	-----
AGL71	-----	-----	-----	-----	-----	-----
AGL42	-----	-----	-----	-----	-----	-----
PI	-----	-----	-----	-----	-----	-----
AP3	-----	-----	-----	-----	-----	-----
TT16	LQLAQP NLQN	DPTAQND---	-----	-----	-----	-----
AGL63	-----	-----	-----	-----	-----	-----
AGL104	PG-VTAETQV	DHEVSDYETK	VPQLSSQ---	-----	-----	-----
AGL66	IPVMKKDAQA	DHEVSDYDIR	MPQLSSQ---	-----	-----	-----
AGL67	-----	-----	-----	-----	-----	-----
AGL94	NNLQQQQQQQ	QQQQQQDPSM	YDPMANNNGG	CFQIPHDQSM	FVNDHHHHHH	HHHQNWVPDS
AGL30	-----	-----	-----	-----	-----	-----
AGL65	SSLPCSISMF	DEYLF SQMQQ	PN-----	-----	-----	-----

SEP1	-----	-----	-*
SEP2	-----	-----	-*
SEP3	-----	-----	-*
AGL3	-----	-----	-*
AGL13	-----	-----	-*
AGL6	-----	-----	-*
CAL	-----	-----	-*
AP1	-----	-----	-*
FUL	-----	-----	-*
AGL79	-----	-----	-*
FLF	-----	-----	-*
SHP1	-----	-----	-*
SHP2	-----	-----	-*
AG	-----	-----	-*
AGL11	-----	-----	-*
AGL12	-----	-----	-*
AGL70	-----	-----	-*
MAF2I	-----	-----	-*
MAF1	-----	-----	-*
MAF5I	-----	-----	-*
MAF4I	-----	-----	-*
AGL17	-----	-----	-*
AGL21	-----	-----	-*
ANR1	-----	-----	-*
AGL16	-----	-----	-*
AGL15	-----	-----	-*
AGL18	-----	-----	-*
SVP	-----	-----	-*
AGL24	-----	-----	-*
AGL19	-----	-----	-*
AGL14	-----	-----	-*
SOC1	-----	-----	-*
AGL72	-----	-----	-*
AGL71	-----	-----	-*
AGL42	-----	-----	-*
PI	-----	-----	-*
AP3	-----	-----	-*
TT16	-----	-----	-*
AGL63	-----	-----	-*
AGL104	-----	-----	-*
AGL66	-----	-----	-*
AGL67	-----	-----	-*
AGL94	MFGQTSYNQV	CVFTPPLELS	R*
AGL30	-----	-----	-*
AGL65	-----	-----	-*

The 5' of agamous and related sequences has been excluded.