Evolution du rôle et des propriétés biochimiques de LEAFY

Edwige Moyroud

UNIVERSITE DE GRENOBLE
École doctorale Chimie et Science du Vivant

THESE DE DOCTORAT DE BIOLOGIE
Spécialité : Biologie Végétale

Présentée par
Edwige MOYROUD

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITE DE GRENOBLE

ÉVOLUTION DU ROLE ET DES PROPRIETES BIOCHIMIQUES DE

LEAFY

Un régulateur central du développement floral

Thèse dirigée par Dr. François PARCY et Dr. Charlie SCUTT

Soutenue le mercredi 6 octobre 2010 devant le jury composé de :

Pr. Stefan NONCHEV	Président du jury
Dr. Miguel BLÁZQUEZ	Rapporteur
Dr. Michiel VANDENBUSSCHE	Rapporteur
Dr. Sylvie MAZAN	Examinateur

Remerciements

« I think we have all experienced passion that is not in any sense reasonable »

Stephen Fry

Ces années de thèse auront représenté de longues heures passées à rêver, s'enflammer, monter les escaliers… et certainement pas assez de temps à remercier. J'ai eu la chance de partager mes journées avec une équipe en or massif et je dois à chacun un merci bien particulier. J'espère que ces quelques phrases traduiront toute la reconnaissance et l'affection profondes que j'ai pour ceux qui, de près ou de loin, se sont retrouvés embarqués dans cette aventure.

Tout d'abord, un grand merci à **Marilyn Vantard** pour m'avoir accueilli au sein de son laboratoire et pour avoir toujours fait preuve d'une grande disponibilité.

Un merci très chaleureux également à **Sylvie Mazan**, **Miguel Blazquez**, **Michiel Vandenbussche** et **Stefan Nonchev**, qui ont très gentiment accepté d'évaluer ce travail.

François, je m'étais dit que j'apprendrai vite à tes côtés et c'est vrai, tu m'as appris à manipuler, à me tromper et à tout recommencer. Bien sûr je pourrais te remercier pour ton écoute, ta patience, ta disponibilité…tout cela à beaucoup compté mais pas autant que le plaisir partagé. Merci d'avoir échangé sans compter tes idées, ta curiosité, tes envies tout en étant la raison qui m'a si souvent manqué. Merci d'avoir eu l'envie et l'énergie de dompter le 'bébé biologiste' que j'étais. Je ne sais pas si beaucoup de directeurs de thèse m'auraient laissé m'exprimer autant que tu l'as fait. Ces quatre années se sont envolées et je sais que je vais devoir filer. Nous avons encore quelques histoires à terminer et j'espère en avoir bientôt d'autres à commencer ailleurs mais le coeur du moteur risque de me manquer ! Faire de la science c'est bien, avec toi c'est mieux.
Mon plus grand merci te revient.

Un merci tout particulier à **Cécile** qui m'a mis le pied à l'étrier quand je suis arrivée et avec qui j'ai partagé ma paillasse, mon bureau et mes souvenirs des deux premières années. On a tous travaillé pour faire grandir ton 'petit LFY' et j'espère que tu seras là cet automne pour écouter ses progrès.

Mille mercis à '**Manou'**, notre 'lab manager bien-aimé' qui fait tourner notre labo avec beaucoup de brio et de bonne humeur. Quand tu es là, tout va ! Travailler à tes côtés est toujours un immense plaisir et tu vas beaucoup me manquer.

Un merci du fond du cœur à **Sandrine**, mon acolyte de Selex, de café, de paillasse, de bureau et mon amie. Sans toi, le Selex serait toujours handicapé. Merci d'être aussi bon public, de toujours rire à mes bêtises même si elles sont loin d'être réussies. Je sais que Romain doit attendre impatiemment de revoir sa maman tous les soirs, mais je suis bien contente qu'il nous la prête un petit peu pendant la journée.

Muchas gracias **Eugenio** ! Le plus difficile pour te remercier, c'est de savoir par où commencer ! Depuis ton arrivée, tu nous as beaucoup apporté : des gâteaux, des photos, du vocabulaire inventé et un usage intensif du mot 'en effet' ….et bien sûr des programmes Pythons, des équations et beaucoup de discussions animées et souvent très drôles. Merci d'avoir choisi notre équipe pour t'exiler quelque temps de ton Espagne adorée, travailler à tes côtés à été un réel plaisir.

Un gros merci 'bien cristallisé' à **Renaud**. Te voir aussi 'mordu de LEAFY' suffit à me faire sourire et je ne crois pas avoir passé une journée sans me dire combien j'étais contente que tu aies rejoins notre équipe…même lorsque tu siffles toute la journée. Merci d'être là aussi dans les moments les moins drôles.

Un sincère merci à Super-**Gabi** qui, entre une UE à préparer, des manips à avancer et des enfants à cajoler, a toujours trouvé le temps de discuter, de m'écouter et…de relire une bonne partie de ce manuscrit ! Autant te l'annoncer, je ne partirai pas sans avoir eu droit à un cours accéléré de 'démarche sautillante'. C'est tellement joli, que moi aussi je veux essayer !

Camille et **Marie**, vous commencez juste votre thèse et pourtant je crois avoir plus appris de vous que vous de moi. Merci pour me prouver tous les jours qu'on peut avancer sans être survolté. PpLFY et ses petits copains sont entre de bonnes mains. Votre calme et votre douceur m'ont beaucoup apporté, surtout au cours de ces derniers mois un peu chargés ! Camille, il faut que tu me donnes le secret de ta technique du Koala et Marie, j'aimerais vraiment, juste pour satisfaire ma curiosité, que tu me montres rien qu'une fois ce que ça donne une 'Marie énervée' !

Un immense Merci aux 'Supermen' de l'actine, **Laurent** et **Christophe**, pour nous avoir tant aidé à faire démarrer ce projet. Votre attention constante au cours de cette thèse m'a été très précieuse. Vos éclats de rire manquent beaucoup au bâtiment C2.

Mylène, la liste des remerciements à te faire serait longue alors je vais me contenter du merci le plus personnel : merci de m'avoir aidé à rebaptiser les objets de notre nouveau labo, c'est quand même plus agréable d'appeler les choses par leur prénom !

Merci aussi à **Gilles** et son calme olympien, à **Sophie** et nos pauses-café et à **Cristel** pour nos discussions toujours animées.

Un gros merci à Sylvianne : je voulais compter le nombre de fois où tu m'as secourue d'une CAI (catastrophe administrative imminente) parce que j'avais en quelque sorte 'oublié' de remplir quelques papiers très importants…grâce à toi, je fais des progrès !

A Lyon, mille mercis à **Charlie** pour ses conseils, sa disponibilité, sa gentillesse et son humour ainsi qu'à **Mathieu** sans qui les heures de SPR auraient semblé bien longues. Merci aussi à **Claudia** et **Pierre**, mes 'parents' du RDP pour m'avoir accueilli à bras ouvert chaque fois que j'ai eu besoin d'un toit !

Un merci tout particulier à **Nelu**, **Christophe** et **Patrice** du laboratoire RDP. Je n'oublierai jamais le premier jour, où vous m'avez expliqué que 'oui, effectivement, il faut mettre un cône au bout de la pipette'....je confirme, ça marche beaucoup mieux comme ça !

All my admiration and affection to Mike Frohlich. Without you, '*Welwitschia*' would have meant nothing but a rude word to me. I feel very lucky to be able to learn from you. A big 'Thank you' to Nicole too, who looked after me so well when I was in London and keep an eye on me even when I am away, I owe you a lot.

To Simon, Willem, Lewis, Stefan and Alice, who waited for me patiently.

Merci à mes parents d'être tout simplement contents que j'aime autant ce que je fais et à Lauranne et Emmanuel pour avoir si bien entouré leur grande sœur pas toujours facile à supporter.

Je suis venu à au laboratoire PCV pour jouer avec LEAFY et pourtant c'est elle que j'aurai le moins de mal à quitter.

Abréviations

35S: promoteur constitutif fort de l'*ARN 35S* du virus de la mosaïque du chou-fleur
aa: acides aminés
ADN: Acide DésoxyriboNucléique
AF: Anisotropie de Fluorescence
AG: AGAMOUS
ANA: Amborellaceae-Nymphaeales-Austrobaileyales
AP1/AP2/AP3: APETALA1, APETALA2, APETALA3
ARN: Acide RiboNucléique
ChIP-Seq: Chromatine ImmunoPrécipitation Sequencing, Précipitation de la chromatine suivie d'un séquençage massif
EDTA: Acide Ethylène-Diamine-Tétraacétique
EMSA: Electrophoretic Mobility Shift Assay, experience de retard sur gel
FLO: FLORICAULA (homologue de LEAFY chez le muflier).
GFP: Green Fluorescent Protein, protéine fluorescente verte
GST: Gluthatione S-Transferase
HTH: Helix-Turn-Helix, motif de liaison à l'ADN de type hélice-tour-hélice
IPTG: IsoPropyl-beta-Thio-Galactoside
K_D^{App}: constante de dissociation apparente
kDa: kiloDalton
LFY: LEAFY
LFY-C: domaine C-terminal de LFY
MW: Molecular Weight, poids moléculaire
NLY: NEEDLY, paralogue de LFY chez les gymnospermes
nt: nucleotide
pb: paire de bases
PI: PISTILLATA
PpLFY: homologue de LFY chez la mousse *Physcomitrella patens*
PSSM: *Position-Specific Scoring Matrix*
PWM: *Position Weight Matrix*
SDS-PAGE: Sodium DodecylSulfate PolyAcrylamide Gel Electrophoresis, gel de polyacrymide en conditions dénaturantes
SEC: Size Exclusion Chromatography, chromatographie d'exclusion de taille
Selex: Systematic Evolution of Ligands by Exponential Enrichment, sélection de sites de liaison par enrichissement exponentiel
SEP1/SEP2/SEP3/SEP4: SEPALLATA1, SEPALLATA2, SEPALLATA3, SEPALLATA4
SPR: Surface Plasmon Resonance, résonance, plasmonique de surface
TEV: Protéase du Virus Tobacco Etch
TFL1: TERMINAL FLOWER1
u.a.: Unité Arbitraire
UFO: UNUSUAL FLORAL ORGANS
VP16: facteur de transcription du virus de l'Herpès possédant un domaine activateur constitutif de la transcription
WUS: WUSCHEL

SOMMAIRE

Introduction

Introduction

« Il y a des fleurs partout pour qui veut bien les voir »

Matisse

Les fleurs comptent parmi les créations les plus visibles et spectaculaires de l'évolution. Immobiles et silencieuses, elles sont si familières à l'œil humain qu'il est pourtant facile de les côtoyer sans les remarquer. La fleur distingue les angiospermes ou plantes à fleurs des autres plantes terrestres. Riches de plus de 350000 espèces, les angiospermes constituent le plus grand groupe d'organismes vivants après les arthropodes (Bramwell, 2002; Paton et al., 2008). Outre leur attrait esthétique, les plantes à fleurs fournissent à l'homme de quoi se nourrir, se vêtir, se soigner ou s'abriter.

Redoutablement efficace pour attirer les pollinisateurs animaux et produire les fruits renfermant les graines qui assurent le maintien et la propagation des espèces, la fleur a joué un rôle prépondérant dans le succès évolutif des angiospermes. C'est pourquoi les bases génétiques des traits floraux ont depuis longtemps suscité l'intérêt des horticulteurs et des biologistes de l'évolution. Cependant, les gènes bâtisseurs de l'architecture florale n'ont été identifiés que récemment grâce à une poignée d'espèces modèles. Ces connaissances ouvrent désormais des perspectives nouvelles pour explorer les mécanismes à l'origine de cette structure unique.

1 La fleur, un 'abominable' mystère

La fleur est une structure familière qui peut adopter une infinie variété de morphologies mais son origine évolutive demeure méconnue.

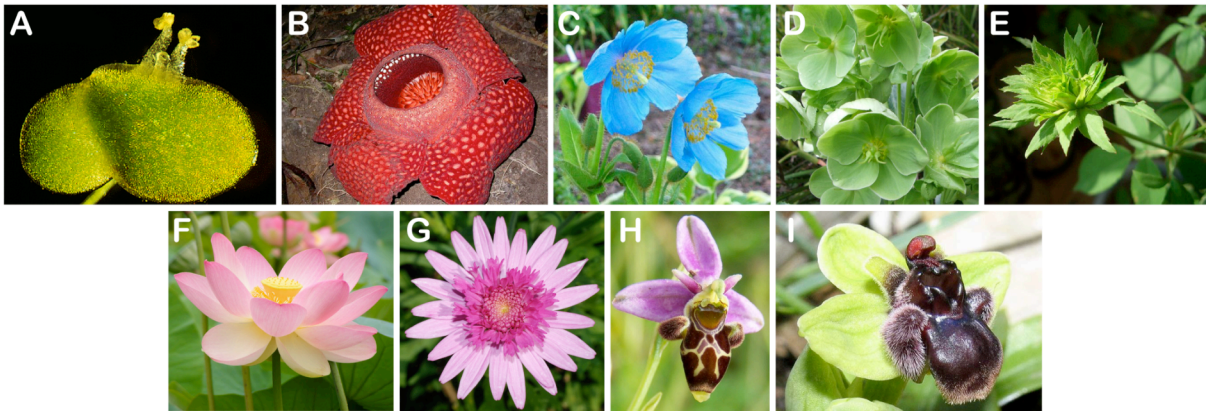
1.1 Qu'est ce qu'une fleur ?

La fleur est la structure reproductrice emblématique des angiospermes, pourtant il est difficile de lui trouver une définition consensuelle (Bateman et al., 2006). Extrêmement plastique, la fleur adopte en effet une variété remarquable de morphologies : les fleurs de la lentille d'eau (*Lemna minor*) par exemple, ne mesurent que deux à trois millimètres tandis que celle de la rafflésie (*Rafflesia arnoldi*) avec ses 7 kg et son mètre de circonférence appartient aux fleurs géantes (Fig.1A-B). Si la plupart se distinguent par leurs couleurs vives, certains spécimens sont à peine visibles, les fleurs étant aussi vertes que le feuillage (Fig.1C-E). Alors que la longue tige du lotus (*Nelumbo nucifera*) porte une fleur unique, de nombreuses espèces

exhibent une multitude de fleurs disposées le long d'un axe pour constituer une inflorescence. Chez les Astéracées, la famille des marguerites et des tournesols par exemple, les fleurs combinées au sein d'une même inflorescence sont de plusieurs types : chacune adopte une morphologie caractéristique de sa position, au centre ou en périphérie du groupe, conférant ainsi à l'inflorescence l'aspect d'une grande fleur unique (Fig.1F-G). Enfin, certaines espèces sont capables d'élaborer des structures complexes: en plus de mimer l'aspect des femelles de l'espèce pollinisatrice, plusieurs fleurs d'orchidées synthétisent un cocktail de phéromones attirant les mâles, s'assurant ainsi une pollinisation efficace (Fig.1H-I).

Figure 1 | Diversité des fleurs.

(A) La fleur minuscule de la lentille d'eau (*Lemna minor*) ; (B) La plus grande fleur du monde, *Rafflesia arnoldii* ; (C) La corolle bleu vif du pavot de l'Himalaya (*Meconopsis sp.*) contraste avec les fleurs vertes de l'hellébore *Helleborus viridis* (D) ou celle (E) d'une espèce ancienne de rose, *Rosa chinensis viridiflora*. (F) La fleur simple du lotus (*Nelumbo nucifera*) et (G) l'inflorescence du chrysanthème (*Argyranthemum frutescence*) où chaque languette correspond à une fleur. Ici deux types de fleurs sont visibles : les fleurs à grande corolle rose pâle de la périphérie et les fleurs du centre plus petites et plus foncées. Deux exemples de mimétisme floral : une partie de la corolle d'*Ophrys scolopax* (H) et d'*Ophrys bombyliflora* (I) mime le corps des insectes femelles pollinisateur en adoptant sa couleur et son aspect ('poils', 'pattes' et 'yeux' visibles).



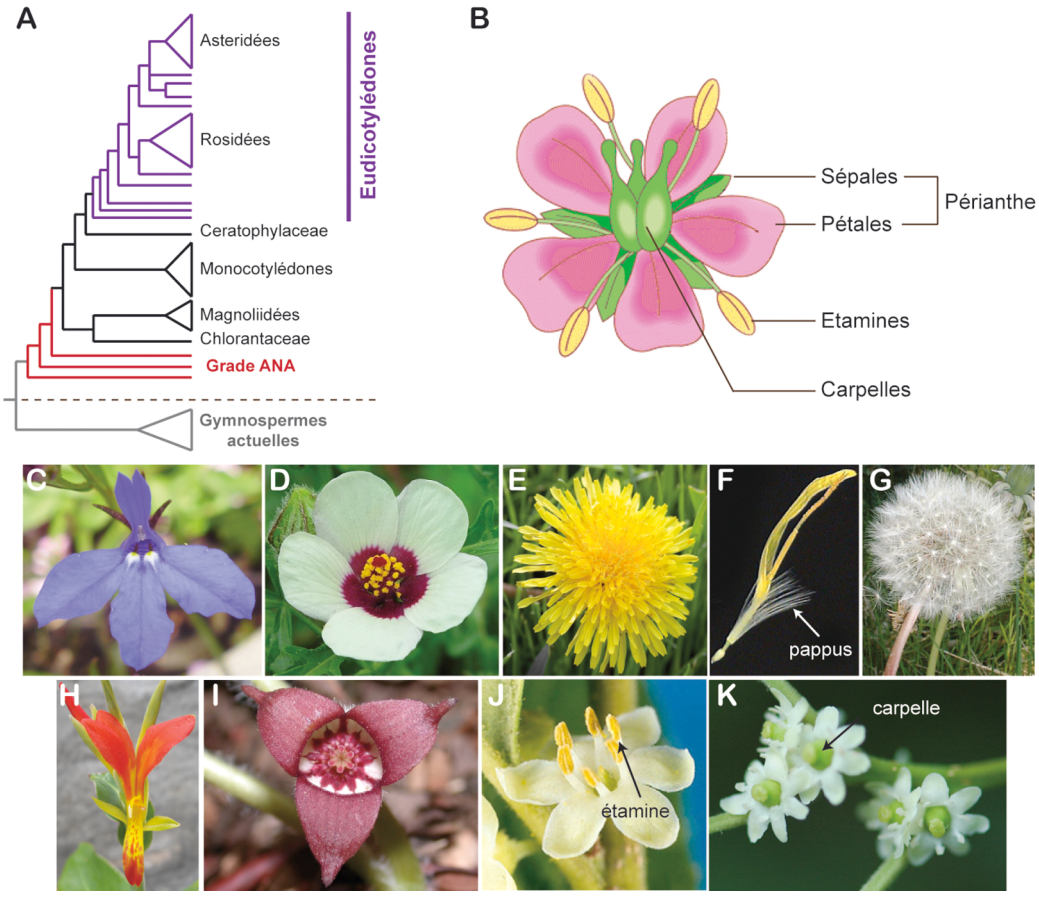
Malgré les apparences, il existe en réalité des limites environnementales et développementales à l'évolution de certaines formes (Barrett, 2008; Endress and Doyle, 2009): les morphologies adoptées doivent être pleinement fonctionnelles sous peine de disparaître et les fleurs élaborées, aussi complexes soient elles, doivent pouvoir émerger d'un petit groupe de cellules indifférenciées, le méristème floral, grâce à un développement ordonné. Satisfaire à ces deux contraintes restreint la gamme des formes possibles si bien qu'il existe un plan d'organisation simple partagé par l'ensemble des fleurs et témoignant de leur origine commune.

Longtemps obscures, les relations de parenté (phylogénie) entre les différentes espèces sont aujourd'hui mieux comprises (Fig.2A) : les plantes à fleurs incluent deux larges clades ou

branches, les monocotylédones et les eudicotylédones, un grade basal constitué de trois groupes plus petits (*Amborella*, Nymphaeales et Austrobaileyales) et un clade de taille intermédiaire, les magnoliidées. Deux familles, les Chloranthaceae et les Ceratophyllaceae viennent compléter l'embranchement, cependant leur positionnement est encore débattu (Soltis et al., 2005; Qiu et al., 2006; Jansen et al., 2007; Moore et al., 2007; APGIII, 2009). Ainsi, les angiospermes dérivent toutes d'un ancêtre commun, par conséquent la fleur telle que nous la connaissons n'a été inventée qu'une seule fois au cours de l'évolution.

Figure 2 | Organisation de la fleur type des angiospermes et variation du plan de base.

(A) Phylogénie simplifiée des angiospermes d'après (Jansen et al., 2007; Moore et al., 2007; APGIII, 2009) (B) Dessin de la fleur-type montrant les différentes pièces florales disposées sur les 4 verticilles. Fleur zygomorphe de Lobélia (*Lobelia erina*) (C) et corolle régulière d'Hibiscus (*Hibiscus trionum*) (D). Le pappus du pissenlit (*Taxacum sp.*), visible à la base de chaque fleur individuelle, (E) représente une modification extrême des sépales (F). Le pappus forme les aigrettes blanches lorsque la corolle disparaît (G). La corolle du Canna (*Canna brasiliensis*) est composée d'étamines transformées (H). Fleur de gingembre (*Asarum canadense*) dépourvue de pétales (I) ou les fleurs mâles (J, carpelles absents) ou femelles (K, étamines absentes) du Houx (*Ilex verticillata*) (J).



La fleur typique des angiospermes rassemble au sein d'une même structure organisée en territoires concentriques ou verticilles, les organes reproductifs mâles (étamines) et femelles

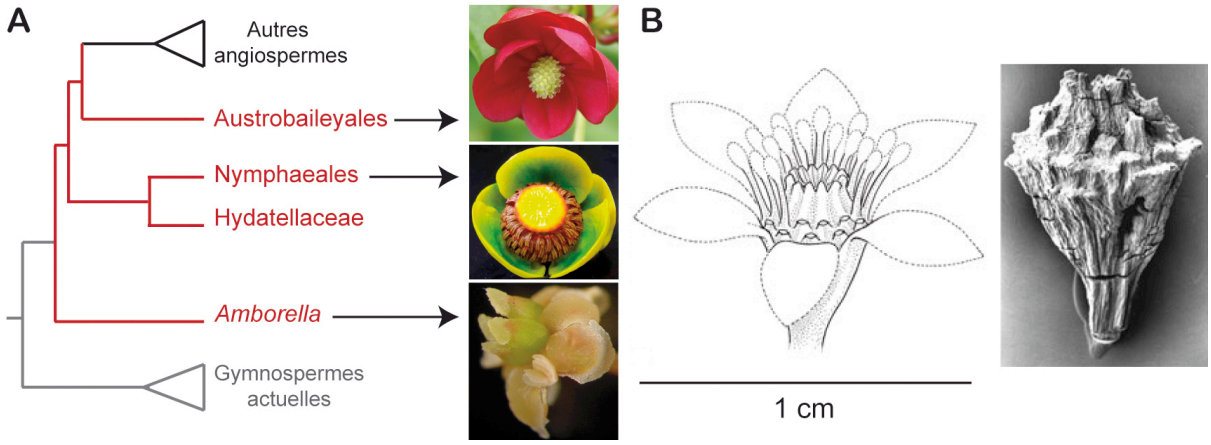
(carpelles renfermant les ovules) entourés des pièces ornementales et protectrices constituant le périanthe: une corolle de pétales participe à l'attraction des pollinisateurs et un ensemble de sépales, le calice, protège le bouton floral (Fig.2B). Cet assemblage assure une reproduction sexuée efficace (Regal, 1977) et favorise l'isolement reproductif et donc la spéciation (Grant, 1971). À une exception près (*Lacandonia schismatica*,(Ambrose et al., 2006)), l'architecture florale est rigoureusement conservée chez les eudicotylédones et monocotylédones qui représentent à elles seules plus de 90% des angiospermes connues (Drinnan et al., 1994). La diversité naît de variations à l'intérieur de ce plan de base. En effet, le nombre et l'aspect des pièces florales varient énormément d'une famille à l'autre. Au sein d'un même verticille, chaque organe peut aussi adopter une morphologie propre: contrairement aux fleurs régulières, la corolle des fleurs dites zygomorphes présente une symétrie bilatérale due à l'assemblage de différents pétales (Fig2.C-D). La modification des pièces florales est parfois extrême et l'organe d'origine devient difficilement identifiable (Fig.2E-G). De même, les conversions entre organes de verticilles voisins sont fréquentes: les sépales des tulipes par exemple sont identiques aux pétales tandis que la 'corolle' des fleurs de Cana est en fait constituée d'étamines modifiées, les vraies pétales étant à peine visibles (Fig.2H). Enfin, des modifications drastiques sont observées lorsqu'un organe est perdu (Fig.2I-K)
Ainsi, la simplicité du plan d'organisation commun aux fleurs leur confère la flexibilité nécessaire pour générer la diversité observée dans la nature.

A quoi ressemblaient les premières fleurs ? Ces structures seraient apparues il y a près de 140-180 millions d'années (Bell et al., 2005; Frohlich, 2006) et les fossiles découverts indiquent qu'une grande diversité de fleurs existait très précocement dans l'histoire évolutive des angiospermes (Friis et al., 2006; Crepet and Nicklas, 2009). Cette variabilité ancestrale est toujours visible aujourd'hui : le grade « ANA » (Qiu et al., 2006; Friis and Crane, 2007; Saarela et al., 2007; APGIII, 2009) compte moins de 200 espèces, pourtant les fleurs de ces « angiospermes basales » sont très différentes les unes des autres. Plusieurs études comparatives de ces espèces couplées aux reconstructions tridimensionnelles de fossiles récemment découverts (Gandolfo et al., 2004; Friis et al., 2010) ont permis cependant de dégager des caractéristiques partagées, probablement héritées de l'ancêtre commun (Endress, 2001; Endress and Doyle, 2007).Les premières fleurs étaient de petites structures bisexuelles, possédant certainement un nombre non strictement fixé d'organes arrangés en spirale même si plusieurs spécimens fossiles et Nymphaeales actuelles démontrent que l'organisation en verticilles a été adoptée très tôt (Endress and Friis, 1991). Le périanthe ancestral était

vraisemblablement indifférencié, sans sépales et pétales distincts, et les transitions morphologiques progressives d'une pièce florale à une autre le long de la spirale étaient fréquentes (Kim et al., 2005). Les carpelles de ces fleurs étaient simplement fermés par des sécrétions et abritaient un nombre très restreint d'ovules (Endress and Doyle, 2007; Endress, 2010).

Figure 3 | Les premières plantes à fleurs.

(A) Diversité des fleurs du groupe ANA : *Amborella trichopoda* (Photo : Sangtae Kim), *Nuphar advena* et *Schisandra rubriflora* (Photo : Scott Zona). Reconstruction (B) et spécimen fossile (C) d'une espèce de Nymphaeales du Crétacé inférieur. Adapté de (Friis et al., 2001)) La fleur n'excédait probablement pas 1 cm.



Les progrès de la phylogénétique supportent l'idée que la structure reproductrice des angiospermes n'est apparue qu'une seule fois. Malgré une diversité apparente, les fleurs partagent toutes un plan d'organisation simple assurant la mise en place successive des 4 types d'organes caractéristiques de la fleur qui distinguent les angiospermes des autres plantes terrestres.

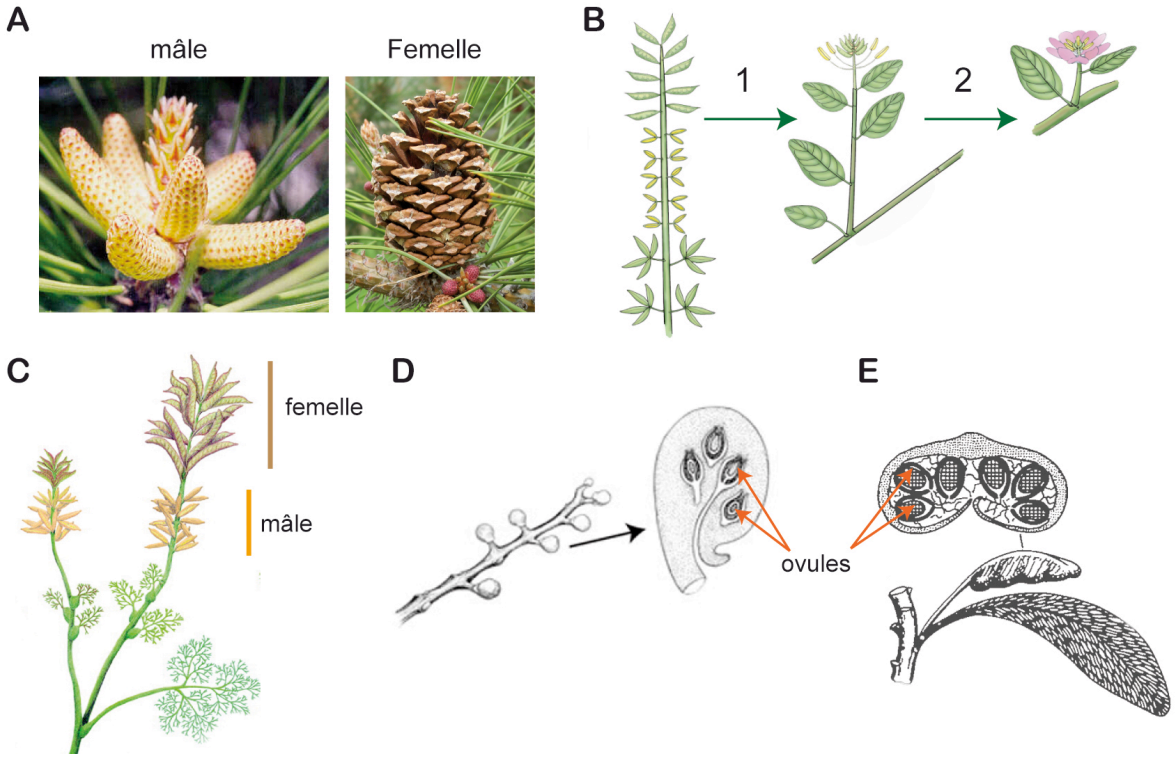
1.2 L'origine des angiospermes

La fleur rassemble sur une même structure des organes mâles et femelles, protège les ovules dans une structure fermée, le carpelle (d'où le nom de l'embranchement du grec *angeion* « le vase » et *sperma* « la graine », signifiant littéralement « graine dans un récipient ») et associe aux organes reproducteurs des pièces stériles, les pétales et les sépales, capables d'attirer les pollinisateurs animaux. Ces caractéristiques combinées à d'autres non mentionnées ici, expliquent pourquoi la fleur est souvent décrite comme « un complexe d'innovations » qui a conféré au groupe un net avantage sélectif par rapport aux autres plantes terrestres (Baum and Hileman, 2006). En effet, les gymnospermes, dont les conifères sont les principaux représentants actuels, sont les végétaux les plus proches des plantes à fleurs mais leurs organes reproducteurs sont très différents : ceux-ci sont portés par des structures allongées ou

cônes, mâles ou femelles, qui ne possèdent pas de périanthe mais des feuilles modifiées ressemblant à des écailles supportant les pièces reproductrices. Les organes producteurs de pollen sont très différents des étamines et les ovules sont nus (d'où le nom du groupe *gymnos* « nu » *sperma* « la graine ») car le carpelle est absent (Frohlich, 2003; Frohlich and Chase, 2007; Rudall and Bateman, 2010). Créer la première fleur à partir d'une structure de type cône a nécessité de rassembler les organes des deux sexes, d'ajouter un périanthe et de raccourcir les entrenoeuds entre les différents types d'organes pour aboutir à une structure compacte (Baum and Hileman, 2006). De plus, comme les organes mâles et femelles des deux groupes ne sont pas équivalents, il a fallu profondément les modifier ce qui complique l'identification de l'ancêtre commun et explique qu'aujourd'hui encore on ignore l'origine des plantes à fleurs (Specht and Bartlett, 2009).

Figure 4 | Construction de la fleur à partir du cône des gymnospermes.

(A) Cônes mâle et femelle de Pin (*Pinus sp.*) (B) L'apparition de la fleur a nécessité (1) le rapprochement des organes mâle et femelle sur une même structure, (2) la compaction de l'axe et (3) le développement du périanthe [adapté de (Baum and Hileman, 2006)]. (C) *Archaefructus* (Crétacé) est en fait une inflorescence moderne puisque les 'organes' mâles et femelles qui lui donnait l'apparence de chaînon manquant sont en fait des fleurs individuelles unisexuée [d'après (Sun et al., 2002)]. Les fossiles de (D) *Caytonia* et (E) *Glossopteris* (Jurassique) possèdent une structure qui pourrait ressembler à l'ancêtre du carpelle [d'après (Retallack and Dilcher, 1981; Frohlich and Chase, 2007)].



Les gymnospermes actuelles partageant elles aussi une origine commune (Doyle, 2006), il n'existe pas d'intermédiaire vivant entre les deux groupes : l'ancêtre commun des plantes à graine est une espèce aujourd'hui disparue qu'il faut donc rechercher parmi les fossiles (Chaw et al., 2000; Doyle, 2008). Or, les gisements fossiles les plus anciens renferment de nombreuses structures florales déjà élaborées mais aucune forme intermédiaire qui témoignerait d'une transition entre le cône des gymnospermes et la fleur (Frohlich and Chase, 2007). Charles Darwin lui-même a donc qualifié l'origine et la diversification des angiospermes « d'abominable mystère » (Darwin, 1903).De nombreux fossiles dont le célèbre *Archaefructus* ont initialement été proposés comme chaînon manquant mais ces spécimens ont *in fine* été rattachés à des groupes modernes de plantes à fleurs, leur aspect « primitif » étant dû à des simplifications secondaires (Sun et al., 2002; Friis et al., 2003). On a donc recherché du côté des fossiles gymnospermes. La majorité des espèces gymnospermes ayant aujourd'hui disparu, les fossiles du groupe présentent une diversité morphologique bien supérieure à celle des espèces survivantes. Plusieurs spécimens fossiles (glossopterids, *Bennetitales*, *Caytonia*) ont été proposés comme précurseur de la fleur. Malheureusement, aucun de ces spécimens ne présente à lui seul l'ensemble des caractéristiques nécessaires pour générer la première fleur (Doyle, 2006; Frohlich and Chase, 2007).

Comment, en l'absence de données fossiles suffisantes, accéder aux étapes intermédiaires ayant précédé l'apparition de la fleur ? Une solution consiste à changer d'échelle d'observation en se rappelant que les réseaux génétiques qui contrôlent le développement floral dérivent d'un réseau ancestral. Reconstituer ce réseau permettrait alors de comprendre les modifications nécessaires à l'émergence des angiospermes.

1.3 La stratégie évo-dévo

La morphologie d'un individu est le résultat de l'organogenèse qui permet de créer une forme complexe à partir d'un groupe de cellules indifférenciées. Pour générer des structures nouvelles, transmises au travers des générations, il faut altérer la composante héréditaire du développement, c'est-à-dire modifier les réseaux génétiques qui gouvernent la mise en place du plan d'organisation des êtres vivants et la formation de leurs différents organes (Carroll et al., 2001; Wilkins, 2002). La génétique évolutive du développement ou évo-dévo aborde de manière inédite les mécanismes à l'origine des innovations morphologiques (Gould, 1977). En mettant en évidence qu'il existe des outils moléculaires et cellulaires conservés par tous

les organismes d'un grand groupe et que l'utilisation différentielle de ces outils au cours de l'évolution est source de diversité, cette approche parfois qualifiée de « paléontologie sans fossile » a éclairé des questions jusque-là sans réponse. À titre d'exemple, on comprend désormais comment deux systèmes complexes, l'œil à facette des insectes et l'œil camérulaire des vertébrés, pourtant d'apparence si différente, ont été engendrés à partir d'un système visuel ancestral (Arendt, 2003).

Or, plusieurs milliers de gènes sont activés lors de la floraison (Schmid et al., 2003) et plus de 1300 gènes sont différentiellement exprimés lors du développement des poils à la surface des ailes de drosophiles (Ren et al., 2005). Comment identifier parmi tous ces candidats les molécules à l'origine des modifications phénotypiques ?

En dépit de la complexité apparente, ces milliers de gènes sont eux-mêmes contrôlés par un nombre restreint de loci codants pour des protéines régulatrices, les facteurs de transcription. Capables de reconnaître spécifiquement certains motifs du génome (les élèments *cis*) pour réguler l'expression des gènes environnants, les facteurs de transcription sont au cœur des circuits génétiques du développement. Des modifications, même mineures du lieu et du moment d'expression des gènes encodant ces protéines régulatrices ou des éléments *cis* de leurs gènes cibles peuvent donc conduire à des changements phénotypiques importants (Chen and Rajewsky, 2007; Wagner, 2007). Puisqu'ils jouent le rôle d'architecte du vivant, ces régulateurs ont très souvent été recrutés par l'évolution et la domestication pour créer de la nouveauté. Par exemple, la corne portée par la tête de différentes espèces de scarabées au sein d'une même famille aurait été inventée plusieurs fois indépendamment en exprimant ectopiquement un à trois facteurs de transcription contrôlant normalement le développement des pattes (Moczek and Rose, 2009). Réciproquement, la domestication du maïs à partir de son ancêtre la téosinte a nécessité la fixation d'une combinaison particulière d'élement-*cis* dans la région régulatrice du gène *teosinte branched 1* (Clark et al., 2006).

Dès lors, identifier les facteurs qui gouvernent le développement floral et étudier leur devenir au travers de l'évolution apparaît comme une alternative séduisante pour explorer les origines mystérieuses de la fleur.

2 LEAFY, chef d'orchestre du développement floral

Le passage de l'état juvénile à l'état reproducteur de la plante permet la mise en place de groupes de cellules indifférenciées, les méristèmes floraux. Quels sont les éléments indispensables au niveau génétique pour générer à partir de ces cellules non spécialisées, les différents organes de la fleur ?

2.1 Les bâtisseurs moléculaires de la fleur

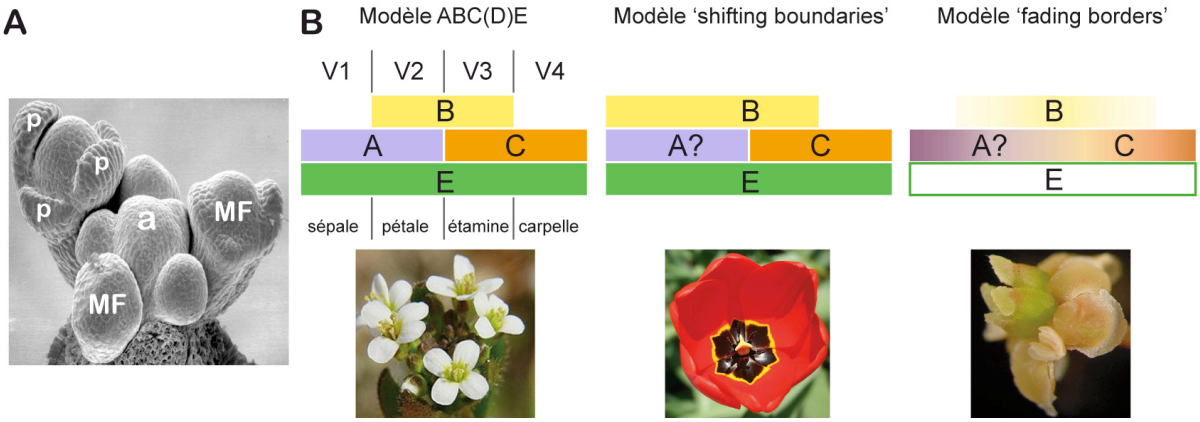
La transition florale marque la fin du développement végétatif de la plante en installant le long de la tige ou hampe florale des groupes de cellules pluripotentes dédiés à la construction de la fleur : les méristèmes floraux (Blazquez et al., 2006; Liu et al., 2009a). Comme les fleurs partagent un même plan d'organisation, il doit exister un circuit génétique commun capable d'engendrer à partir de ce méristème et quels que soit l'espèce et l'aspect final de chaque organe, l'organisation typique de la fleur des angiospermes. Le fonctionnement du méristème génère les différentes pièces florales sous forme de primordium tout en spécifiant l'identité qu'elles doivent adopter (Fig.4A). Une des premières étapes de l'organogenèse florale consiste donc à établir quatre sous-domaines au sein du méristème floral afin de générer les quatre types d'organes de la fleur. Les études réalisées chez deux espèces modèles, *Arabidopsis thaliana*, une petite herbacée de la famille du chou et de la moutarde et *Antirrhinum majus*, le muflier, fréquemment rencontré dans les jardins, ont établi un modèle génétique, le modèle ABC, permettant d'expliquer comment ces différents territoires sont spécifiés (Bowman and Meyerowitz, 1991; Coen and Meyerowitz, 1991; Irish, 2010).

Le modèle ABC repose sur l'expression combinée d'une poignée de gènes homéotiques organisés en trois classes A, B et C. Chaque verticille de la fleur exprime une combinaison unique des activités A, B et C, ce qui détermine l'identité adoptée par ses primordia (Fig.4B). Chez *Arabidopsis*, les gènes *APETALA1* (*AP1*) et *APETALA2* (*AP2*) (classe A), sont impliqués dans l'identité des sépales et pétales ; *APETALA3* (*AP3*) et *PISTILLATA* (*PI*) (classe B), contribuent à l'identité des pétales et étamines ; enfin le gène *AGAMOUS* (*AG* ; classe C) est responsable de l'identité des pièces reproductrices mâles (si coexpression avec les gènes de classe B) ou femelles (classe C exprimée seule) (Fig.5B). Plus récemment, deux groupes supplémentaires ont été ajoutés au modèle: la classe D incarnée par le gène *SEEDSTICK* spécifie les ovules et les gènes *SEPALLATA1, 2*, 3 et 4 (*SEP1-4*) formant la classe E participent à l'identité de tous les organes. À l'exception d'*AP2*, tous les gènes du modèle ABC appartiennent à la grande famille des gènes à boîte MADS (Ng and Yanofsky, 2001; Lohmann and Weigel, 2002). Les protéines correspondantes sont des facteurs de transcription capables de s'assembler selon le modèle quartet (Egea-Cortines et al., 1999;

Theissen and Saedler, 2001). En fonction du complexe formé, le répertoire de gènes régulés varie, permettant ainsi d'orienter la différentiation du primordium pour élaborer un sépale, un pétale, une étamine ou un carpelle. Ce modèle s'est révélé largement conservé, en particulier pour les fonctions B et C, chez l'ensemble des eudicotylédones même s'il a été suggéré chez plusieurs espèces que la fonction A ne serait pas remplie par les homologues d'*AP1* et *AP2* mais par un autre gène ou groupe de gènes inconnus (Litt and Kramer, 2010).

Figure 5 | Fonctionnement du méristème floral et modèles ABCDE.

(A) L'apex (a) génère les méristèmes floraux (MF) qui initient les primordia (p) d'organes (Photographie: J.Bowman) (B) Modèle ABC(D)E classique d'*Arabidopsis thaliana* et ses variations principales : le modèle 'sliding boundaries' explique pourquoi sépales et pétales sont morphologiquement identiques chez la tulipe (*Tulipa sp.*) et le modèle 'fading borders' a été proposé comme applicable aux angiospermes du grade ANA et à certaines Magnoliidées. La fonction E n'a pas été étudiée en détail chez ces espèces d'où l'absence de coloration. Pour simplifier, la fonction D qui spécifie les ovules chez les eudicotylédones centrales n'a pas été représentée. Le point d'interrogation signifie que les gènes assurant la fonction A chez ces espèces n'ont pas été clairement identifiés.



Le modèle ABCDE permet aussi d'expliquer quelques-unes des différentes morphologies florales observées dans la nature (Theissen and Melzer, 2007; Litt and Kramer, 2010). Les fleurs de tulipe ou d'orchidée par exemple possèdent des sépales pétaloïdes identiques en apparence aux 'vrais' pétales issus du second verticille et il a été montré que les gènes de classe B s'exprimaient également dans le premier verticille des fleurs de ces espèces ce qui expliquerait la morphologie de type pétale (Fig.4B, (Bowman, 1997; Kramer et al., 2003))

Le modèle ABCDE était-il en place chez les premières plantes à fleurs ? Le portrait de 'fleur primitive' actuellement privilégié implique que les pièces florales disposées en spirale, présentaient un changement graduel de leur identité le long de cette spirale. Une telle organisation est, à première vue, incompatible avec le modèle ABCDE qui nécessite

l'existence de territoires concentriques bien délimités. Or, chez plusieurs espèces du grade ANA, les homologues des gènes B, C et E sont généralement exprimés dans des territoires plus larges que ceux prédits par le modèle ABC de sorte que la superposition des différentes classes se fait ici de manière graduelle (Buzgo et al., 2004; Kim et al., 2005; Soltis et al., 2007). L'absence de limites franches entre un territoire exprimant C et un territoire exprimant B par exemple, expliquerait qu'une gamme de morphologies soit observée le long de la spirale (Fig.4B).

Chez les premières plantes à fleurs, une combinaison d'homologues des gènes homéotiques du modèle ABCDE contrôlait probablement déjà la mise en place des différentes pièces florales. Une restriction des différents territoires d'expression associée à une duplication des gènes des classes ABCDE serait survenue plus tard dans l'histoire des angiospermes permettant la mise en place contiguë d'organes radicalement différents d'un verticille à l'autre. Comprendre comment le modèle ABCDE s'est mis en place et donc comprendre l'origine de la fleur demande d'identifier les gènes qui gouvernent l'expression combinée de ces facteurs d'identité.

2.2 *LEAFY*, le régulateur des régulateurs

Au début des années 90, le gène *LEAFY* (LFY) a été identifié quasi-simultanément chez l'Arabette et le muflier grâce à l'étude de mutants (Coen et al., 1990; Schultz and Haughn, 1991; Weigel et al., 1992). Chez les plantes porteuses d'une version défectueuse du gène, le développement floral est fortement perturbé voire complètement aboli et les organes floraux sont remplacés par des feuilles d'où le nom du gène (« feuillu » en anglais).

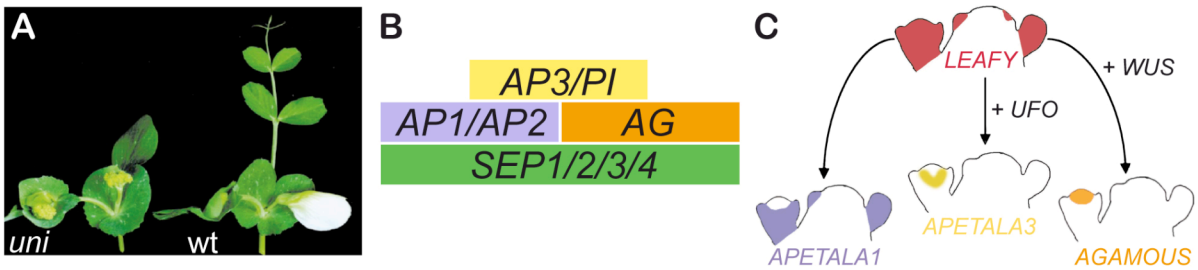
L'étude des patrons d'expression du gène ont démontré que *LFY* était principalement voire exclusivement activé sur les flancs du méristème apical de la plante délimitant ainsi les secteurs qui donneront les méristèmes floraux (Coen et al., 1990; Weigel et al., 1992). Cette activation précoce du gène est le résultat de l'intégration sur son promoteur de plusieurs voies qui contrôlent la floraison : LFY répond positivement par exemple à une horloge développementale et à la signalisation par les phytohormones gibbérellines, d'où l'appellation d'intégrateur floral, c'est-à-dire un gène percevant les différents stimuli de floraison pour déclencher l'organogenèse florale (Blazquez et al., 1998; Nilsson et al., 1998; Blazquez and Weigel, 2000; Parcy, 2005). Plusieurs facteurs ont été identifiés comme capables de réguler directement ou indirectement l'expression de ce gène mais les régions régulatrices de *LFY* ne

sont pas caractérisées en détail (Putterill et al., 1995; Blazquez et al., 1998; Achard et al., 2004; Moon et al., 2005; Eriksson et al., 2006; Lee et al., 2008; Yamaguchi et al., 2009; Preston and Hileman, 2010).

LFY code pour un facteur de transcription essentiellement présent dans le noyau mais aussi capable de se déplacer de cellule en cellule au travers des plasmodesmes (Sessions et al., 2000). Ce régulateur confère leur identité aux méristèmes floraux en réprimant les gènes d'identité de tige (ex :*AGL24*, *TFL1*, (Weigel and Nilsson, 1995; Yu et al., 2004)) et l'expression de LFY est ensuite maintenue au travers du méristème floral via une boucle d'autorégulation (ex : boucle d'autorégulation AP1-CAL, (Bowman et al., 1993; Liljegren et al., 1999)) lorsque les primordia des 4 types d'organes floraux sont générés. Une découverte majeure a eu lieu lorsque plusieurs analyses génétiques ont démontré que LFY activait directement l'expression d'*AP1* (classe A), d'*AP3* et *PI* (classe B) et d'*AG* (classe C) (Parcy et al., 1998; Busch et al., 1999; Lohmann et al., 2001; Lamb et al., 2002). Comme *LFY* est exprimé au travers de l'ensemble du méristème floral, cette régulation nécessite la présence de cofacteurs exprimés localement et permettant une induction « verticille spécifique » des gènes d'identités. Ainsi, les gènes *AP3* et *AG* sont induits par LFY en coopération avec *UNUSUAL FLORAL ORGANS* (*UFO*) et *WUSCHEL* (*WUS*) respectivement (Levin and Meyerowitz, 1995; Lohmann et al., 2001) et SEPALLATA3 (SEP3) (Liu et al., 2009b).

Figure 6 | Activation des gènes homéotiques floraux par *LEAFY* et ses homologues.

(A) Plante de pois (*Pisum sativum*) sauvage ou présentant une mutation dans le gène *UNIFOLIATA*, homologue de *LEAFY* chez cette espèce (Photo : (Hofer et al., 1997)) (B) Gènes ABCE d'*Arabidopsis thaliana* : *APETALA1* (*AP1*), *APETALA2* (*AP2*), *APETALA3* (*AP3*), *PISTILLATA* (*PI*), *AGAMOUS* (*AG*) et *SEPALLATA1/2/3* et *4* (*SEP1/2/3/4*) (C) Régulation directe d'*AP1*, *AP3* et *AG* par LEAFY (LFY). L'activation d'*AP3* est limitée aux verticilles 2 et 3 grâce à *UNUSUAL FLORAL ORGANS* (*UFO*), *WUSCHEL* est un cofacteur de LFY pour activer *AG*. Récemment, il a été montré que *SEP3* est nécessaire à l'activation d'*AP3* et *AG* par LFY (non représenté ici, (Liu et al., 2009b)).



Ainsi, LFY pilote le réseau gouvernant l'organogenèse florale et constitue donc un candidat de premier choix pour essayer de comprendre les mécanismes génétiques qui ont permis l'émergence de la fleur.

2.3 Un seul gène, plusieurs fonctions: nouveaux rôles de LEAFY

LFY jouant un rôle clé lors du développement floral, les homologues du gène ont été intensément recherchés chez les autres angiospermes. Les études réalisées chez plusieurs eudicotylédones mais aussi monocotylédones ont montré que les fonctions du gène étaient largement conservées chez ces espèces. Des travaux plus récents ont également mis à jour des fonctions insoupçonnées de *LFY*. Afin de faire le point sur l'état des connaissances, nous avons réalisé une synthèse du rôle de *LFY* chez les plantes à fleurs. Ce travail a abouti à la rédaction d'une première revue présentée ici (Moyroud et al., 2009).

L'existence de *LFY* précède l'apparition des fleurs puisque des homologues du gène ont été identifiés chez plusieurs espèces de mousses, fougères et gymnospermes, les trois autres groupes de plantes terrestres. Ces espèces ne produisant pas de fleur, la fonction de ce gène chez ces organismes reste en grande partie mystérieuse. Plusieurs scénarios évolutifs ont proposé des mécanismes moléculaires pour expliquer la création de la première fleur à partir d'un ancêtre gymnosperme (Frohlich, 2000; Albert et al., 2002; Theissen and Becker, 2004; Baum and Hileman, 2006) et tous attribuent à LFY un rôle important voire central. Dans une seconde revue, nous avons rassemblé les données portant sur les homologues non-angiospermes de *LFY* et proposé un scénario intégratif de l'évolution de ce facteur (Moyroud et al., 2010).

Ces deux revues récapitulent de manière synthétique l'état des connaissances sur LFY et sont fondamentales pour la compréhension du travail de thèse réalisé. A ce titre, elles font partie intégrante de l'introduction et nous renvoyons le lecteur à ces deux articles avant de terminer la lecture de l'introduction.

J. Plant Biol. (2009) 52:177–185
DOI 10.1007/s12374-009-9028-8

REVIEW

The LEAFY Floral Regulators in Angiosperms: Conserved Proteins with Diverse Roles

Edwige Moyroud · Gabrielle Tichtinsky · François Parcy

Received: 3 March 2009 / Accepted: 9 March 2009 / Published online: 12 May 2009

Abstract Genetic analyses in model angiosperms have shown that the LEAFY/FLORICAULA transcription factor plays a central role in flower development. In *Arabidopsis*, LEAFY (LFY) triggers the development of floral meristems and controls their patterning through the activation of floral organ identity genes. Several recent reports enlighten the structure and function of this conserved protein but also illustrate the variety of roles it plays in different angiosperms.

Keywords LEAFY · Angiosperms · Flower · Evolution · Architecture

Introduction

Flowering plants add to nature's beauty and supply many of the resources needed for human life. Molecular genetics of model angiosperms identified the *FLORICAULA/LEAFY* gene, thereafter named *LEAFY* (*LFY*), as a central regulator of floral development. The *LFY* gene is found throughout terrestrial plant genomes, including from groups such as mosses, ferns, or gymnosperms, predating the origin of flowering plants. Work in *Arabidopsis* and several other plants identified LFY as a novel type of transcription factor responsible for the regulation of genes controlling floral meristem and floral organs development (Benlloch et al. 2007; Blazquez et al. 2006; Parcy 2005). LFY possesses several intriguing features: (1) its sequence does not resemble any known transcription factor and its origin remains elusive; (2) LFY exhibits two domains with high levels of sequence conservation from mosses to angiosperms; (3) as opposed to most transcription factors, LFY did not form a multigene family and remained at very low copy number in the genome with no obvious signs of subfunctionalization.

Because of its essential function in flower development and its presence in plant genomes before the appearance of flowers, LFY stands at the center of several evolutionary scenarios attempting to explain the origin of angiosperms (Frohlich and Chase 2007; Theissen and Melzer 2007). In recent years, major progress has been made on our knowledge about the target genes of LFY, its structure, its mode of action, its interacting partners, and its roles in different species. In this review, we discuss these advances with a focus on the evolutionary implications of the properties and molecular activity of this peculiar protein in flowering plants.

Genetic Analysis in Model Angiosperms

Early Work in *Arabidopsis* and *Antirrhinum*

Two mutants (*floricaula* in *Antirrhinum majus* and *leafy* in *Arabidopsis thaliana*) provided the first genetic evidence of the involvement of the *FLO*/*LFY* gene in floral meristem identity (Carpenter and Coen 1990; Coen et al. 1990; Schultz and Haughn 1991; Weigel et al. 1992). In the snapdragon *flo* mutant, flowers are replaced by shoots

E. Moyroud · G. Tichtinsky · F. Parcy (✉)
Laboratoire Physiologie Cellulaire Végétale, UMR5168,
Centre National de la Recherche Scientifique, Commissariat à l'énergie atomique, Institut National de la Recherche Agronomique, University Joseph Fourier,
17 rue des Martyrs, bât. C2,
38054 Grenoble Cedex 9, France
e-mail: francois.parcy@cea.fr

(Carpenter and Coen 1990; Coen et al. 1990). In the *Arabidopsis lfy* mutant, the most basal flowers are also converted into shoots, but then, flower/shoot intermediates and abnormal flowers are formed at more apical positions (Huala and Sussex 1992; Schultz and Haughn 1991; Weigel et al. 1992). The cloning of *FLO* and *LFY* genes revealed their homology (Coen et al. 1990; Weigel et al. 1992). From now on, we will use *LFY* as a generic name for the *FLO*/*LFY* genes by simplicity, not underestimating the historical importance of *FLO*. Since *lfy* mutants display abnormal development of floral organs, the expression of the *ABC* floral organ identity genes (Lohmann and Weigel 2002) was analyzed in these backgrounds. *FLO* and *LFY* were found to be required for proper *B* and *C* genes regulation (Hantke et al. 1995; Weigel and Meyerowitz 1993) (Fig. 1). In *Arabidopsis*, *LFY* is also needed for the expression of the *APETALA1* (*AP1*) *A*-class gene in early floral meristems (Fig. 1), but *AP1* can also be activated in a *lfy*-independent manner (Ruiz-Garcia et al. 1997) which is also true for its snapdragon ortholog *SQUAMOSA* (*SQUA*) (Carpenter and Coen 1995). The regulation of *ABC* genes by LFY was further corroborated using various gain-of-function transgenic plants. The overexpression of *LFY* in transgenic *Arabidopsis* is sufficient to induce *AP1* in young leaves and the use of an inducible version of LFY (*35S:LFY-GR*) showed that *AP1* regulation by LFY is direct (Parcy et al. 1998; Wagner et al. 1999). Moreover, the expression of LFY fused to the VP16 activation domain demonstrated the capacity of LFY to regulate the *C* gene *AGAMOUS* (*AG*) and the concomitant overexpression of *LFY* and its *UFO* coregulator described later resulted in a precocious activation of the *B* gene *APETALA3* (*AP3*) (Parcy et al. 1998).

In addition to its role in patterning the floral meristem by the local induction of the *A*, *B*, and *C* genes, the increasing LFY levels in leaves primordia prior to flower formation appear to contribute to the control of the flowering time (Blazquez et al. 1997; Weigel and Nilsson 1995).

Large-Scale Experiments Identified New Potential Target Genes

After the initial identification of a few target genes, large-scale experiments were performed that identified a battery of additional target genes. These experiments either used the LFY-GR inducible version (Wagner et al. 2004; William et al. 2004) or used shifts from noninductive short days to inductive long days in wild-type and mutant plants followed by expression analysis at the genomic scale (Maizel et al. 2005; Schmid et al. 2003). Many of these potential target genes still await further detailed analysis while the identity of a few of them guaranteed the functional relevance of their activation by LFY. For example, the *CAULIFLOWER* (*CAL*) gene, known to share meristem identity function with *AP1*, was shown to be directly activated by LFY. *CAL* is also regulated by another recently identified LFY target, *LATE MERISTEM IDENTITY1* (Saddic et al. 2006; William et al. 2004), revealing a regulatory mechanism by which LFY induces some primary targets that subsequently reinforce its role as a gene expression regulator (Fig. 1). Another example of functionally coherent targets for LFY are the *SEPALLATA* (*SEP*) genes (Robles and Pelaz 2005), which are responsible for the E function (Schmid et al. 2003; William et al. 2004). Experiments using the inducible LFY-GR fusion coupled to a translational inhibitor strongly suggested that LFY might directly regulate *SEP2* and *SEP3* (William et al. 2004) (Fig. 1).

Genes Downregulated by LFY

LFY not only activates gene expression but also represses genes controlling the identity of the vegetative or inflorescence meristem (Fig. 1), such as *EMF1* (Chen et al. 1997; Chou et al. 2001) and *TFL1* (Liljegren et al. 1999; Parcy et al. 2002; Ratcliffe et al. 1998) or *AGL24* (Yu et al. 2004; Yu et al. 2002). Genetic analyses showed that LFY is linked

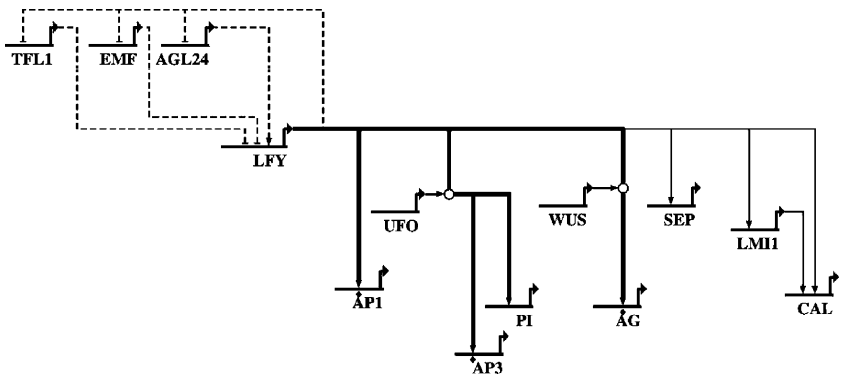


Fig. 1 *LFY* gene regulatory network in *Arabidopsis*. Positive regulations are indicated with *arrows* whereas *bars* indicate negative ones. *Diamonds* indicate the presence of LFY binding sites. Synergistic interactions between gene products are indicated with *bubbles*. For clarity, only the main regulatory relationships, mentioned in the text, are shown. This figure has been generated using the Biotapestry program (Longabaugh et al. 2009)

with these genes through mutual negative feedback loops, a general mechanism shown to be important for switches between different developmental fates. The precise nature of these regulations (direct or indirect) remains to be established.

Mechanisms of Target Gene Regulation by LFY

Identification of DNA Binding Sites for *Arabidopsis* LFY

The FLO and LFY protein sequences did not immediately reveal their function because these proteins did not show any similarity to other known regulators. In 1998, the LFY protein from *Arabidopsis* was demonstrated to be nuclear and to bind DNA elements present in the *AP1* promoter both in vitro and by yeast one-hybrid assays (Parcy et al. 1998). Later, additional binding sites were also identified in the *AG* regulatory intron and the *AP3* promoter (Busch et al. 1999; Lamb et al. 2002; Lohmann et al. 2001). These data identified LFY as a novel type of transcription factor. To understand its mode of action, it is important to note that LFY was able to activate transcription in yeast only when fused to a heterologous activation domain (Parcy et al. 1998), suggesting that it may lack an intrinsic capacity for transcriptional activation. It might, however, not be the case for all LFY proteins, as suggested by FLO sequence analysis and by the phenotype of *Arabidopsis* plants expressing the rice *LFY* ortholog, *RFL* (Chujo et al. 2003; Coen et al. 1990).

The alignment of the binding sites present in *AP1* and *AG* regulatory regions identified the pseudopalindromic CCANTGG/T sequence as the consensus recognized by *Arabidopsis* LFY, although the binding sites found in the *AP3* promoter match poorly with this motif (they display only CCNNNG) (Lamb et al. 2002). Thus, the current definition of the LFY binding site has thus little predictive value and more work is needed to establish a position weight matrix capable of accurately predicting the presence of LFY binding sites in a given DNA stretch.

Structural Analysis of LFY DNA Binding Domain

Once the LFY capacity to bind specific DNA sequences was established, the question of the nature and the origin of this novel transcription factor became more acute. The answer was obtained with the crystallographic structure of LFY DNA binding domain (LFY-DBD) in complex with *AP1* or *AG* binding sites (Hames et al. 2008). This structure revealed a novel protein fold, made of seven alpha helices, which is not found in any other protein structures. Within the seven helices, three form a helix-turn-helix (HTH) motif, frequently found in proteins interacting with nucleic acids (Aravind et al. 2005). Interestingly, structural comparisons showed that LFY possesses similarities with DNA binding proteins such as the Tc3 transposase, paired, or homeodomain transcription factors (Hames et al. 2008). Like these proteins, LFY interacts with both DNA grooves: the HTH contacts conserved bases in the major groove, whereas a N-terminal extension with an arginine residue enters into the minor groove. The contacts between LFY and DNA extend farther than anticipated from the consensus *cis*-element.

The DNA binding mode of LFY was also elucidated (Hames et al. 2008). Consistent with the semipalindromic nature of the LFY binding site, LFY-DBD was found to bind DNA as a dimer. However, LFY-DBD appears to be monomeric in the absence of DNA and to dimerize upon DNA binding following a cooperative mode of DNA binding: the binding of the first monomer to DNA favors the binding of the second one. At the atomic level, the cooperativity was explained by the presence of several H bonds between both monomers. This cooperative binding mechanism was proposed to contribute to the sharp induction of flowering.

Further analysis of the LFY protein should determine whether this mechanism is valid for the entire protein and how the presence of the conserved N-terminal domain contributes to LFY functional properties.

Interaction of LFY with Coregulators

As indicated previously, LFY is thought to be a neutral transcription factor, at least in *Arabidopsis*, requiring coactivators to activate the transcription of its target genes in different domains. Two of these coregulators have been identified: WUSCHEL (WUS) in the case of *AG* activation and UNUSUAL FLORAL ORGANS (UFO) for *AP3*. WUSCHEL is a homeodomain transcription factor required for meristem homeostasis (Laux et al. 1996) and expressed in the center of shoot and flower meristems. WUS binding sites have been identified in close proximity to LFY binding sites on *AG* regulatory intron, and yeast assays demonstrated the capacity of these proteins to synergistically activate transcription when coexpressed (Lohmann et al. 2001). However, the complex containing LFY and WUS bound together to DNA has never been observed in vitro, and the recent crystallographic structure suggests that a LFY dimer might not fit together with WUS on *AG* binding sites. It is, therefore, not clear whether LFY and WUS could form a heterodimer or whether they need to bind alternatively to recruit complementary members of the transcription machinery. Moreover, the expression domain of *WUS* overlaps only partially with that of *AG*, suggesting either that WUS capacity to act at a distance widely extends its action domain or that other proteins (such as members of

the WOX family) (Breuninger et al. 2008) might also contribute to *AG* regulation together with LFY.

In the case of *AP3* activation, the UFO protein was shown to be required as LFY coregulator. UFO is not a transcription factor but an F-BOX protein involved in protein ubiquitination through an SCF complex (SKP1, CULLIN, F-Box). Recently, two independent studies in *Arabidopsis* and *Petunia hybrida* demonstrated a direct interaction between UFO and LFY (DOT and ALF in petunia) (Chae et al. 2008; Souer et al. 2008). Although interaction data are not entirely consistent between the two studies (interaction seems to occur through LFY N terminus in petunia and C terminus in *Arabidopsis*), the evidence supporting this interaction is very convincing. The fusion between UFO and the EAR repression domain, triggering a LFY loss-of-function phenotype, nicely demonstrates that UFO is expected to be recruited in LFY regulatory complexes not only at *AP3* promoter (Chae et al. 2008). Based on work on several transcription factors, different models have been proposed to explain how ubiquitination might promote transcription factor activity (Conaway et al. 2002; Kodadek et al. 2006; Lipford and Deshaies 2003). Until now, examples of such regulation in the plant kingdom are rare and it will be interesting to investigate how LFY fits with the models arising from other kingdoms.

Coregulators and precise mechanisms for other LFY-dependent genes such as *SEP* or *TFL1* genes have not been identified. Furthermore, despite the wealth of data regarding *ABC* genes activation, the reason why each gene is induced in a specific spatiotemporal expression domain has not been elucidated. Whether this information lies in the LFY binding site themselves, in their vicinity, or in the chromatin environment is still a matter of investigation.

Evolution of the Protein

Origin

The origin of the ancestral *LFY* gene is unknown. *LFY* is found in all terrestrial plants but has not so far been identified in algae. The DNA binding domains of LFY and the Tc3 transposase show some structural similarity (Hames et al. 2008) suggesting that, as many other transcription factors, LFY might be derived from a transposon (Breitling and Gerber 2000; Feschotte 2008). LFY could have been brought to plants early on by viral or bacterial transfer and would have drifted until acquiring a new essential function. Unfortunately, sequence similarity with Tc3 transposase is too weak to suggest a common origin and confirm this attractive hypothesis. The sequencing of new genomes might help answering the question in the future.

Why not a family

As opposed to most developmental regulators in angiosperms, *LFY* is not part of an extended gene family (e.g., Bharathan et al. 1999; Martinez-Castilla and Alvarez-Buylla 2003). *LFY* homologues have been cloned in more than 200 species and *LFY* is mostly found as a single copy gene. It is clear that *LFY* experienced duplication as any other genes since there are traces of copies being eliminated (Aagaard et al. 2006; Baum et al. 2005; Bomblies and Doebley 2005; Bomblies and Doebley 2006; Southerton et al. 1998). Moreover, several species exhibit two or three *LEAFY*-like genes but the phylogeny studies demonstrate that the paralogs are recent copies (Archambault and Bruneau 2004; Shu et al. 2000; Southerton et al. 1998; Wada et al. 2002; Wang et al. 2008; Yoon and Baum 2004). They either result from polyploidy, as in *Nicotiana tabaccum*, or from smaller-scale duplication events, as the two paralogs did not persist long enough to be inherited by multiple species (Baum et al. 2005; Kelly et al. 1995). There are only two documented cases (Maize and Lamiales) where a second copy seems to have been kept unusually long (Aagaard et al. 2006; Bomblies et al. 2003). The reason why *LFY* copies are not more often maintained is not understood.

It has been suggested that extra copies might be detrimental (Baum et al. 2005; Cronk 2001). For instance, increased *LFY* expression might affect plant architecture and reduce the number of progeny, as *35S:LFY* does in *Arabidopsis* (Weigel and Nilsson 1995). It has also been proposed that hub proteins, which contain several distinct interaction surfaces with coregulators, are less prone to form extended families (Kim et al. 2006). This could apply to LFY, although only two interaction surfaces have thus far been identified (for dimerization and interaction with DNA). Testing this hypothesis will thus require some more experimental evidence.

Evolution of the Sequence

LFY contains two domains of high conservation (Maizel et al. 2005). Recent structural data showed that amino acids from the C terminus with side chains interior to the protein or facing the DNA are extremely conserved whereas there is more variation on the protein surface opposite to DNA (Fig. 2). According to this structural model, there is thus no reason to imagine major changes in the DNA recognition in angiosperms (with the exception of the *Brownea* genus where several amino acids in direct contact with DNA are modified; Hames et al. 2008). This prediction is consistent with the complementation experiments showing that *LFY* from several angiosperms partially complement the *Arabidopsis lfy* mutant phenotype. However, careful experiments

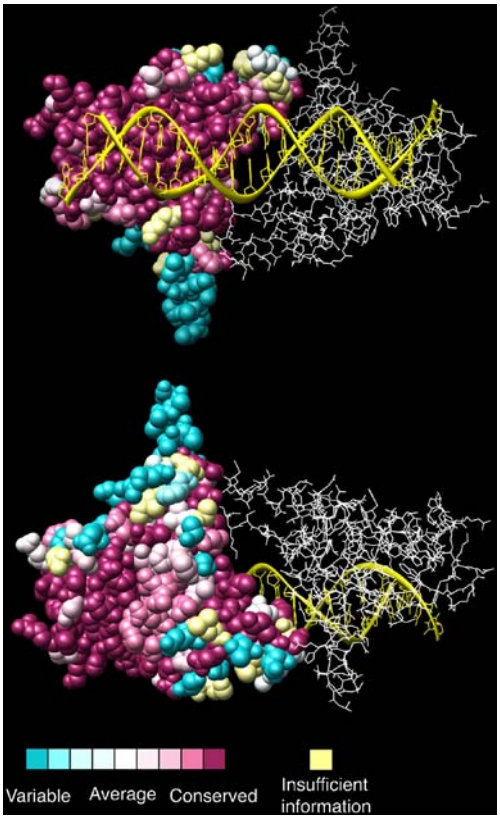


Fig. 2 Conservation of LFY amino acid sequence in angiosperms. Two LFY monomers are shown bound to the DNA (in *yellow*). One monomer is shown in *white sticks* whereas the other one is shown as *spheres* and color-coded according to amino acid sequence conservation using the consurf program (Landau et al. 2005). Conservation color scale is indicated

using microarrays clearly showed that an apparent phenotypic complementation does not guarantee that gene expression is fully restored (Maizel et al. 2005).

Evolution of the Role

The studies in the model plant *Arabidopsis* established that *LFY*'s expression level is critical to trigger flowering and that LFY subsequently controls the development and patterning of newly formed floral meristems. LFY loss-of-function mutants (or transgenic plants) are available for an increasing number of angiosperms species [snapdragon (*A. majus*), *Arabidopsis (A. thaliana*), tomato (*Lycopersicon esculentum*), tobacco (*N. tabaccum*), petunia (*P. hybrida*), *Lotus japonicus*, pea (*Pisum sativum*), *Medicago truncatula*, maize (*Zea mays*), and rice (*Oryza sativa*)]. It thus becomes possible to investigate the evolution of *LFY* function during flowering plants history. As we will elaborate below, with the data available from this reduced number of species, at least one conclusion emerges: in most cases, *LFY* is required for a normal flower development but does not necessarily have the role established in *Arabidopsis* and *Antirrhinum* (Fig. 3).

In tomato, the two *lfy* mutants available [*falsiflora* (*fa*) and *leafy inflorescence* (*lfi*)] flower late and exhibit a complete conversion of flowers to shoots (Kato et al. 2005; Molinero-Rosales et al. 1999). In maize, the plants mutant for the two *LFY* homologs, known as *ZFL* genes, are late flowering and display a loss of floral meristem identity (Bomblies and Doebley 2005; Bomblies and Doebley 2006; Bomblies et al. 2003), demonstrating that these *LFY* functions are relevant outside of the dicots. Both *FA* from tomato and *ZFL* from maize are expressed in all floral primordia and appear to regulate the expression of *ABC* genes: *FA* promotes the induction of *TDR6* (group *B* gene) and *TAG1* (group *C* gene) and the few flowers that develop in maize double mutants are highly modified with sterile carpels or stamens, suggesting that at least B and C functions are altered when ZFL is not active (Bomblies et al. 2003). There is, however, no clear evidence that LFY controls the A function gene in these species.

In petunia, LFY activity is required for normal flower development but LFY does not seem to be the limiting factor in floral initiation. A mutation in *LFY*'s ortholog *ABERRANT LEAF AND FLOWER* (*ALF*) causes a leafy shoot to form instead of flowers. However, *ALF* is already strongly expressed in the organogenetic zone of the shoot apical meristem long before the transition to flowering and overexpressing *ALF* (or *Arabidopsis LFY*) in petunia produces no apparent phenotypic effect (Souer et al. 1998, 2008). The limiting factor in this species has been identified as *DOUBLE TOP* (*DOT*), the homolog of *UFO*: the *DOT* loss-of-function leads to the same phenotype as an *alf* mutation and *DOT* overexpression leads to precocious flowering and transformation of inflorescence to a solitary flower. These results both suggest that *ALF* is not the key factor controlling where and when flowers are produced. It might also be the case in other species such as impatiens (*Impatiens balsamifera*) or tobacco where *LFY* expression is already detected in the vegetative apical meristem and does not increase upon flowering (Kelly et al. 1995; Pouteau et al. 1997). Along the same line, the overexpression of *LFY*'s orthologs in tobacco and poplar does not accelerate flowering in these species as opposed to what is observed in *Arabidopsis* (Ahearn et al. 2001; Rottmann et al. 2000).

In Fabaceae (*P. sativum*, *L. japonicus*, and *Medicago sativa*), *lfy* mutant plants do develop flowers but they are highly modified and display indeterminate growth: sepals and abnormal carpels form, but the region that normally generates stamens and petals initiates new abnormal flowers instead. In *L. japonicus lfy* mutants, *ABC* genes expression is strongly affected but the expression of the *C* gene initiates almost normally in the center of the meristem

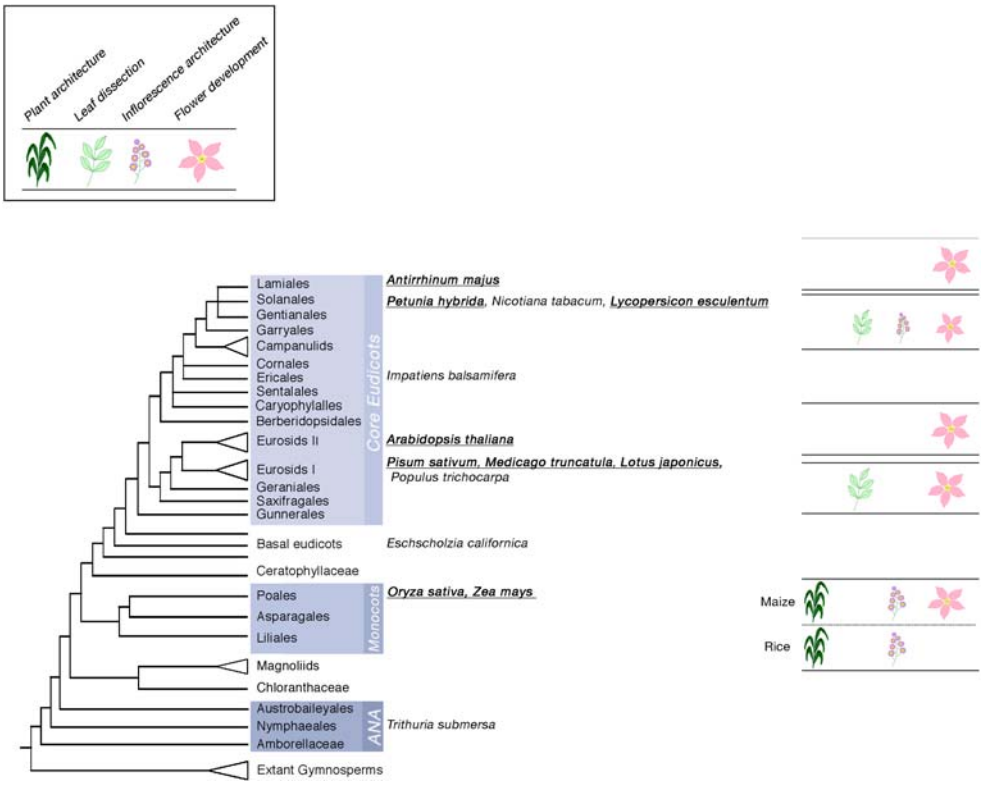


Fig. 3 The variety of roles fulfilled by LFY homologs in angiosperms. The summary phylogenetic tree of major lineages of angiosperms is based on the analyses of Jansen et al. (2007) and Moore et al. (2007). The species named in the text are *listed*, and those where *lfy* mutant or loss-of-function have been described are *underlined*. The processes involving a LFY homolog in these species are indicated on the *right*

indicating a partially *LFY*-independent regulation (Dong et al. 2005). That *C* gene expression could be *LFY*-independent is also suggested by the analysis of several species: in impatiens, tobacco, or the basal eudicot Californian poppy (*Eschscholzia californica*), LFY's expression pattern does not coincide with *C* gene expression and is absent from the center of the floral meristem (Busch and Gleissberg 2003; Kempin et al. 1993; Ordidge et al. 2005). In the absence of *lfy* mutant in these species, a distant action of LFY cannot be excluded (Sessions et al. 2000) but it is also possible that LFY's ability to regulate C activity was acquired long after the appearance of the angiosperms.

A Role in Promoting Indeterminacy

The Fabaceae species are particularly interesting because they show that LFY can also control leaf shape. In wild-type pea plants, for example, the leaves are compound; their limb is divided into leaflet and tendrils, combined together to form a typical dissected leaf. In contrast, mutants in the *UNIFOLIATA (UNI*) gene, the *LFY* homolog, generate simple leaves, with no tendrils and a reduced number of visible leaflets. Similar defects are also observed in two other species of the family, *Lotus* and *Medicago*, but also, to a lesser extent, in tomato, a member of the Solanaceae (Dong et al. 2005; Molinero-Rosales et al. 1999; Wang et al. 2008). The expression pattern studies confirmed that *UNI* is expressed at the margin of the leaves during leaflet formation and it has been suggested that UNI may maintain cells in a transient indeterminate state to facilitate the formation of a compound leaf (Hofer et al. 1997).

Maintaining an indeterminate state is a function often fulfilled by genes of the *KNOTTED* family (Blein et al. 2008) and appears difficult to reconcile with the LFY function in *Arabidopsis* that rather consists in promoting a determinate differentiation of cells on the flanks of the apical meristem. Such a role might actually not be restricted to compound leaf development as indicated by a very interesting recent study in rice (Rao et al. 2008). In this worldwide-cultivated cereal, the *RFL* gene (*LFY* from rice) plays a clear role in flowering time as plants with a compromised *RFL* level show a strong flowering delay. However, once flowering occurs, the architecture of the inflorescence (panicle) of these plants is deeply modified with a reduced number of branches demonstrating that *RFL* plays a role in the generation of outgrowths from the inflorescence meristem and the maintenance of its indeterminacy (Rao et al. 2008). Surprisingly, the few flowers produced in these plants have a normal structure and are fertile, suggesting that LFY does not regulate floral organ identity in rice (even if *RFL* does so when expressed in *Arabidopsis*).

Focusing mostly on the small number of plants where *lfy* mutants or loss-of-function are available already reveals that the highly conserved LFY protein plays a variety of roles in different angiosperms. It is a difficult task,

therefore, to propose an ancestral role for LFY in such a context. A recent study in *Trithuria submersa* (Rudall et al. 2009), a species from one of the earliest extant angiosperm lineages (ANA grade; Fig. 3), clearly demonstrates that the LFY protein is mainly localized in the reproductive organs (stamens and carpels primordia) of this early divergent flowering plant. However, functional studies have not yet been carried out within the ANA grade, and there is an urgent need to develop tools offering us the possibility to unveil the role of *LFY* in significant groups. Indeed, given the high level of conservation, complementation in heterologous systems (such as *Arabidopsis*) is a nice way to test the conservation of biochemical properties, but is not very informative regarding the role of the protein in its own species. An integrative analysis of a few basal and gymnosperm plants will be key to understanding the evolution of LFY's role and may, in turn, shed light on the mysterious origin of flowering plants.

Acknowledgments We thank M. Blazquez, E. Gomez-Minguet, M. Monniaux, T. Spencer, and S. Perry for the critical reading of the manuscript. E.M. is supported by a Ph.D. grant from the French Ministry of Research. The work in our laboratory is supported by the ANR-07-BLAN-0211-01 ("Plant TF-Code") and ATIP+ from the CNRS.

References

Aagaard JE, Willis JH, Phillips PC (2006) Relaxed selection among duplicate floral regulatory genes in Lamiales. J Mol Evol 63:493–503

Ahearn KP, Johnson HA, Weigel D, Wagner DR (2001) *NFL1*, a *Nicotiana tabacum LEAFY*-like gene, controls meristem initiation and floral structure. Plant Cell Physiol 42:1130–1139

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 29:231–262

Archambault A, Bruneau A (2004) Phylogenetic utility of the *LEAFY/FLORICAULA* gene in the Caesalpinioideae (Leguminosae): gene duplication and a novel insertion. Syst Bot 29:609–626

Baum DA, Yoon HS, Oldham RL (2005) Molecular evolution of the transcription factor LEAFY in Brassicaceae. Mol Phylogenet Evol 37:1–14

Benlloch R, Berbel A, Serrano-Mislata A, Madueno F (2007) Floral initiation and inflorescence architecture: a comparative view. Ann Bot (Lond) 100:659–676

Bharathan G, Janssen BJ, Kellogg EA, Sinha N (1999) Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. Mol Biol Evol 16:553–563

Blazquez MA, Soowal LN, Lee I, Weigel D (1997) *LEAFY* expression and flower initiation in *Arabidopsis*. Development 124:3835–3844

Blazquez MA, Ferrandiz C, Madueno F, Parcy F (2006) How floral meristems are built. Plant Mol Biol 60:855–870

Blein T, Pulido A, Vialette-Guiraud A, Nikovics K, Morin H, Hay A, Johansen IE, Tsiantis M, Laufs P (2008) A conserved molecular framework for compound leaf development. Science 322:1835–1839

Bomblies K, Doebley JF (2005) Molecular evolution of *FLORICAULA/LEAFY* orthologs in the Andropogoneae (Poaceae). Mol Biol Evol 22:1082–1094

Bomblies K, Doebley JF (2006) Pleiotropic effects of the duplicate maize *FLORICAULA/LEAFY* genes *zfl1* and *zfl2* on traits under selection during maize domestication. Genetics 172:519–531

Bomblies K, Wang RL, Ambrose BA, Schmidt RJ, Meeley RB, Doebley J (2003) Duplicate *FLORICAULA/LEAFY* homologs *zfl1* and *zfl2* control inflorescence architecture and flower patterning in maize. Development 130:2385–2395

Breitling R, Gerber JK (2000) Origin of the paired domain. Dev Genes Evol 210:644–650

Breuninger H, Rikirsch E, Hermann M, Ueda M, Laux T (2008) Differential expression of WOX genes mediates apical-basal axis formation in the *Arabidopsis* embryo. Dev Cell 14:867–876

Busch A, Gleissberg S (2003) *EcFLO*, a *FLORICAULA*-like gene from *Eschscholzia californica* is expressed during organogenesis at the vegetative shoot apex. Planta 217:841–848

Busch MA, Bomblies K, Weigel D (1999) Activation of a floral homeotic gene in *Arabidopsis*. Science 285:585–587

Carpenter R, Coen ES (1990) Floral homeotic mutations produced by transposon-mutagenesis in *Antirrhinum majus*. Genes Dev 4:1483–1493

Carpenter R, Coen ES (1995) Transposon induced chimeras show that *floricaula*, a meristem identity gene, acts non-autonomously between cell layers. Development 121:19–26

Chae E, Tan QK, Hill TA, Irish VF (2008) An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. Development 135:1235–1245

Chen L, Cheng JC, Castle L, Sung ZR (1997) EMF genes regulate *Arabidopsis* inflorescence development. Plant Cell 9:2011–2024

Chou ML, Haung MD, Yang CH (2001) EMF genes interact with late-flowering genes in regulating floral initiation genes during shoot development in *Arabidopsis thaliana*. Plant Cell Physiol 42:499–507

Chujo A, Zhang Z, Kishino H, Shimamoto K, Kyozuka J (2003) Partial conservation of LFY function between rice and *Arabidopsis*. Plant Cell Physiol 44:1311–1319

Coen ES, Romero JM, Doyle S, Elliott R, Murphy G, Carpenter R (1990) *floricaula*: a homeotic gene required for flower development in *Antirrhinum majus*. Cell 63:1311–1322

Conaway RC, Brower CS, Conaway JW (2002) Emerging roles of ubiquitin in transcription regulation. Science 296:1254–1258

Cronk QC (2001) Plant evolution and development in a post-genomic context. Nat Rev Genet 2:607–619

Dong ZC, Zhao Z, Liu CW, Luo JH, Yang J, Huang WH, Hu XH, Wang TL, Luo D (2005) Floral patterning in *Lotus japonicus*. Plant Physiol 137:1272–1282

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405

Frohlich MW, Chase MW (2007) After a dozen years of progress the origin of angiosperms is still a great mystery. Nature 450:1184–1189

Hames C, Ptchelkine D, Grimm C, Thevenon E, Moyroud E, Gerard F, Martiel JL, Benlloch R, Parcy F, Muller CW (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. EMBO J 27:2628–2637

Hantke SS, Carpenter R, Coen ES (1995) Expression of *floricaula* in single cell layers of periclinal chimeras activates downstream homeotic genes in all layers of floral meristems. Development 121:27–35

Hofer J, Turner L, Hellens R, Ambrose M, Matthews P, Michael A, Ellis N (1997) UNIFOLIATA regulates leaf and flower morphogenesis in pea. Curr Biol 7:581–587

Huala E, Sussex IM (1992) LEAFY interacts with floral homeotic genes to regulate *Arabidopsis* floral development. Plant Cell 4:901–903

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC,

Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104:19369–19374

Kato K, Ohta K, Komata Y, Araki T, Kanahama K, Kanahama Y (2005) Morphological and molecular analyses of the tomato floral mutant *leafy* inflorescence, a new allele of falsiflora. Plant Sci 169:131–138

Kelly AJ, Bonnlander MB, Meeks-Wagner DR (1995) *NFL*, the tobacco homolog of *FLORICAULA* and *LEAFY*, is transcriptionally expressed in both vegetative and floral meristems. Plant Cell 7:225–234

Kempin SA, Mandel MA, Yanofsky MF (1993) Conversion of perianth into reproductive organs by ectopic expression of the tobacco floral homeotic gene *NAG1*. Plant Physiol 103:1041–1046

Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314:1938–1941

Kodadek T, Sikder D, Nalley K (2006) Keeping transcriptional activators under control. Cell 127:261–264

Lamb RS, Hill TA, Tan QK, Irish VF (2002) Regulation of *APETALA3* floral homeotic gene expression by meristem identity genes. Development 129:2079–2086

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–302

Laux T, Mayer KFX, Berger J, Jürgens G (1996) The *WUSCHEL* gene is required for shoot and floral meristem integrity in *Arabidopsis*. Development 122:87–96

Liljegren SJ, Gustafson-Brown C, Pinyopich A, Ditta GS, Yanofsky MF (1999) Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate. Plant Cell 11:1007–1018

Lipford JR, Deshaies RJ (2003) Diverse roles for ubiquitin-dependent proteolysis in transcriptional activation. Nat Cell Biol 5:845–850

Lohmann JU, Weigel D (2002) Building beauty: the genetic control of floral patterning. Dev Cell 2:135–142

Lohmann JU, Hong RL, Hobe M, Busch MA, Parcy F, Simon R, Weigel D (2001) A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. Cell 105:793–803

Longabaugh WJ, Davidson EH, Bolouri H (2009) Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. Biochim Biophys Acta 1789:363–374

Maizel A, Busch MA, Tanahashi T, Perkovic J, Kato M, Hasebe M, Weigel D (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science 308:260–263

Martinez-Castilla LP, Alvarez-Buylla ER (2003) Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. Proc Natl Acad Sci U S A 100:13407–13412

Molinero-Rosales N, Jamilena M, Zurita S, Gomez P, Capel J, Lozano R (1999) *FALSIFLORA*, the tomato orthologue of *FLORICAULA* and *LEAFY*, controls flowering time and floral meristem identity. Plant J 20:685–693

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A 104:19363–19368

Ordidge M, Chiurugwi T, Tooke F, Battey NH (2005) LEAFY, TERMINAL FLOWER1 and AGAMOUS are functionally conserved but do not regulate terminal flowering and floral determinacy in *Impatiens balsamina*. Plant J 44:985–1000

Parcy F (2005) Flowering: a time for integration. Int J Dev Biol 49:585–593

Parcy F, Nilsson O, Busch MA, Lee I, Weigel D (1998) A genetic framework for floral patterning. Nature 395:561–566

Parcy F, Bomblies K, Weigel D (2002) Interaction of LEAFY, AGAMOUS and TERMINAL FLOWER1 in maintaining floral meristem identity in *Arabidopsis*. Development 129:2519–2527

Pouteau S, Nicholls D, Tooke F, Coen E, Battey N (1997) The induction and maintenance of flowering in impatiens. Development 124:3343–3351

Rao NN, Prasad K, Kumar PR, Vijayraghavan U (2008) Distinct regulatory role for RFL, the rice LFY homolog, in determining flowering time and plant architecture. Proc Natl Acad Sci U S A 105:3646–3651

Ratcliffe OJ, Amaya I, Vincent CA, Rothstein S, Carpenter R, Coen ES, Bradley DJ (1998) A common mechanism controls the life cycle and architecture of plants. Development 125:1609–1615

Robles P, Pelaz S (2005) Flower and fruit development in *Arabidopsis thaliana*. Int J Dev Biol 49:633–643

Rottmann WH, Meilan R, Sheppard LA, Brunner AM, Skinner JS, Ma C, Cheng S, Jouanin L, Pilate G, Strauss SH (2000) Diverse effects of overexpression of *LEAFY* and *PTLF*, a poplar (*Populus*) homolog of *LEAFY/FLORICAULA*, in transgenic poplar and *Arabidopsis*. Plant J 22:235–245

Rudall PJ, Remizowa MV, Prenner G, Prychid CJ, Tuckett RE, Sokoloff DD (2009) Nonflowers near the base of extant angiosperms? Spatiotemporal arrangement of organs in reproductive units of Hydatellaceae and its bearing on the origin of the flower. Am J Bot 96:67–82

Ruiz-Garcia L, Madueno F, Wilkinson M, Haughn G, Salinas J, Martinez-Zapater JM (1997) Different roles of flowering-time genes in the activation of floral initiation genes in *Arabidopsis*. Plant Cell 9:1921–1934

Saddic LA, Huvermann B, Bezhani S, Su Y, Winter CM, Kwon CS, Collum RP, Wagner D (2006) The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of *CAULIFLOWER*. Development 133:1673–1682

Schmid M, Uhlenhaut NH, Godard F, Demar M, Bressan R, Weigel D, Lohmann JU (2003) Dissection of floral induction pathways using global expression analysis. Development 130:6001–6012

Schultz EA, Haughn GW (1991) *LEAFY*, a homeotic gene that regulates inflorescence development in *Arabidopsis*. Plant Cell 3:771–781

Sessions A, Yanofsky MF, Weigel D (2000) Cell–cell signaling and movement by the floral transcription factors LEAFY and APETALA1. Science 289:779–782

Shu G, Amaral W, Hileman LC, Baum DA (2000) LEAFY and the evolution of rosette flowering in violet cress (*Jonopsidium acaule*, Brassicaceae). Am J Bot 87:634–641

Souer E, van der Krol A, Kloos D, Spelt C, Bliek M, Mol J, Koes R (1998) Genetic control of branching pattern and floral identity during petunia inflorescence development. Development 125:733–742

Souer E, Rebocho AB, Bliek M, Kusters E, de Bruin RA, Koes R (2008) Patterning of inflorescences and flowers by the F-Box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of petunia. Plant Cell 20:2033–2048

Southerton SG, Strauss SH, Olive MR, Harcourt RL, Decroocq V, Zhu X, Llewellyn DJ, Peacock WJ, Dennis ES (1998) Eucalyptus has a functional equivalent of the *Arabidopsis* floral meristem identity gene LEAFY. Plant Mol Biol 37:897–910

Theissen G, Melzer R (2007) Molecular mechanisms underlying origin and diversification of the angiosperm flower. Ann Bot (Lond) 100:603–619

Wada M, Cao QF, Kotoda N, Soejima J, Masuda T (2002) Apple has two orthologues of *FLORICAULA/LEAFY* involved in flowering. Plant Mol Biol 49:567–577

Wagner D, Sablowski RW, Meyerowitz EM (1999) Transcriptional activation of APETALA1 by LEAFY. Science 285:582–584

Wagner D, Wellmer F, Dilks K, William D, Smith MR, Kumar PP, Riechmann JL, Greenland AJ, Meyerowitz EM (2004) Floral induction in tissue culture: a system for the analysis of LEAFY-dependent gene regulation. Plant J 39:273–282

Wang H, Chen J, Wen J, Tadege M, Li G, Liu Y, Mysore KS, Ratet P, Chen R (2008) Control of compound leaf development by *FLORICAULA/LEAFY* ortholog *SINGLE LEAFLET1* in *Medicago truncatula*. Plant Physiol 146:1759–1772

Weigel D, Meyerowitz EM (1993) *LEAFY* controls meristem identity in *Arabidopsis*. In: Amasino R (ed) Cellular communications in plants. Plenum, New York, pp 115–122

Weigel D, Nilsson O (1995) A developmental switch sufficient for flower initiation in diverse plants. Nature 377:495–500

Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM (1992) LEAFY controls floral meristem identity in *Arabidopsis*. Cell 69:843–859

William DA, Su Y, Smith MR, Lu M, Baldwin DA, Wagner D (2004) Genomic identification of direct target genes of LEAFY. Proc Natl Acad Sci U S A 101:1775–1780

Yoon HS, Baum DA (2004) Transgenic study of parallelism in plant morphological evolution. Proc Natl Acad Sci U S A 101:6524–6529

Yu H, Xu Y, Tan EL, Kumar PP (2002) AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals. Proc Natl Acad Sci U S A 99:16336–16341

Yu H, Ito T, Wellmer F, Meyerowitz EM (2004) Repression of *AGAMOUS-LIKE 24* is a crucial step in promoting flower development. Nat Genet 36:157–161

Review

LEAFY blossoms

Edwige Moyroud[1], Elske Kusters[2], Marie Monniaux[1], Ronald Koes[2] and François Parcy[1]

[1] Laboratoire de Physiologie Cellulaire Végétale, UMR5168, Centre National de la Recherche Scientifique, Commissariat à l'Énergie Atomique, Institut National de la Recherche Agronomique, Université Joseph Fourier, 17 av. des Martyrs, bât. C2, 38054 Grenoble, France
[2] Department of Molecular Cell Biology, Graduate School of Experimental Plant Sciences, VU-University, de Boelelaan 1085, 1081HV Amsterdam, The Netherlands

The *LEAFY* (*LFY*) gene of *Arabidopsis* and its homologs in other angiosperms encode a unique plant-specific transcription factor that assigns the floral fate of meristems and plays a key role in the patterning of flowers, probably since the origin of flowering plants. *LFY*-like genes are also found in gymnosperms, ferns and mosses that do not produce flowers, but their role in these plants is poorly understood. Here, we review recent findings explaining how the LFY protein works and how it could have evolved throughout land plant history. We propose that LFY homologs have an ancestral role in regulating cell division and arrangement, and acquired novel functions in seed plants, such as activating reproductive gene networks.

LEAFY: a master regulator of flower development

Mutations, such as *floricaula* (*flo*) in snapdragon (*Antirrhinum majus*) and *leafy* (*lfy*) in *Arabidopsis*, allowed the identification of a unique type of transcription factor that specifies floral identity of meristems and controls the very first steps in the formation of a flower (Figure 1). Following the isolation of *FLO* and *LFY*, homologs have been identified from numerous species, including gymnosperms and non-seed plants. Here, we summarize the recent findings enlightening the key role of *LFY* genes in multiple species with a variety of different floral architectures, as well as novel data illustrating how this unique protein works.

In flower meristems, LFY acts as a master regulator orchestrating the whole floral network: it activates downstream genes that give their unique identities to the floral meristem and the floral organ primordia [1–6]. It also controls the expression of several additional genes of unknown function, some of which might specify other floral traits such as whorled phyllotaxis or absence of internode elongation [7,8].

LFY is a plant-specific transcription factor that directly binds to the regulatory region of its target genes through a helix–turn–helix motif buried within a unique protein fold [9]. Surprisingly, *LFY* is found as a single gene in most land plant species, even in monocots where several events of gene duplications occurred [4]. LFY is uniformly expressed in floral buds and, therefore, regional coregulators are needed to induce distinct target genes in specific subdomains (Figure 2). The MADS box gene *SEPALLATA3 (SEP3)*, which is expressed in the three central floral whorls, acts as such a coactivator for the induction of B (*APETALA3*, *AP3*) and C (*AGAMOUS*, *AG*) floral organ identity genes [10]. The expression of *AG* is restricted to the flower center by several additional coactivators and corepressors [3,6]. To activate *AP3* in whorls 2 and 3, LFY binds to an F-box protein, known as UFO (UNUSUAL FLORAL ORGANS) in *Arabidopsis*, which is part of an SCF (Skp1–cullin–F-box protein)-type ubiquitin ligase [11]. How SCF^{UFO} promotes LFY activity is unclear, but is likely to be similar to the activation of a variety of transcription factors in yeast by the ubiquitination–proteasome system [12]. Several recent findings indicate that UFO is involved in the activation of many other LFY targets: mutation of the *UFO* homolog *FIMBRIATA* (*FIM*) in snapdragon reduces expression of both B and C genes [13] and in petunia (*Petunia hybrida*), tomato (*Solanum lycopersicon*) and lotus (*Lotus japonicus*), it causes a nearly complete loss of floral meristem identity and strongly perturbs the expression of all floral organ identity genes [14–16]. In addition, gain-of-function experiments have shown that constitutive expression of both *ABERRANT LEAF AND FLOWER* (*ALF*, petunia *LFY*) and *DOUBLE TOP* (*DOT*, petunia *UFO*) ectopically activates a wide spectrum of B-, C-, D- and E-type organ identity genes in petunia [15]. The relatively weak phenotype of *Arabidopsis ufo* and snapdragon *fim* mutants is probably due to functional redundancy, because LFY activity is fully dependent on coexpression of *UFO* or *DOT* when expressed in petunia [15]. Moreover, expression of a dominant negative form of *UFO* in *Arabidopsis* causes a strong *lfy*-like phenotype, including loss of floral meristem identity, which is similar to petunia and tomato *ufo* mutants [11].

LFY is also involved in grass flower development. In maize (*Zea mays*), the *LFY* homologs *ZFL1* and *ZFL2* are required for proper expression of B and C genes in flowers [17], whereas in rice (*Oryza sativa*), the *LFY* ortholog *RFL* is not expressed in floral meristems (Figure 2), and flowers appear fertile even when *RFL* is silenced [18,19]. However, the phenotype of these flowers has not been precisely described, and, because the *UFO* ortholog *ABERRANT PANICLE ORGANIZATION 1* (*APO1*) controls flower development in rice [20,21], RFL and APO1 might perform this task together. The *LFY* gene could therefore have had its floral function before the divergence of dicots and monocots, but this role was partially lost in rice and was taken over by another factor (e.g. the MADS box factor MOSAIC FLORAL ORGANS1 [22]). In Californian poppy

Corresponding authors: Koes, R. (ronald.koes@falw.vu.nl); Parcy, F. (francois.parcy@cea.fr).

 1360-1385/$ – see front matter doi:10.1016/j.tplants.2010.03.007 Trends in Plant Science 15 (2010) 346–352

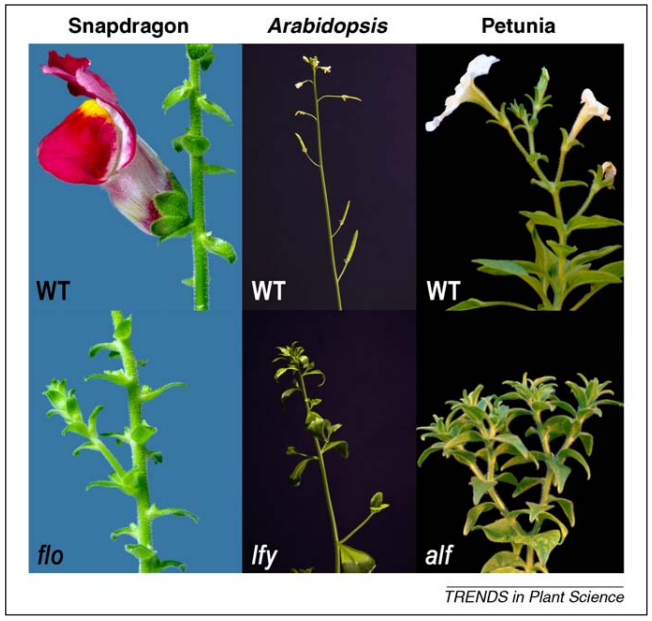


Figure 1. Phenotypes of wild type inflorescences in snapdragon, *Arabidopsis* and petunia and the corresponding *flo*, *lfy* and *alf* mutant.

(*Eschscholzia californica*), the expression pattern of the *LFY* ortholog *EcoFLO* does not coincide with expression of *AG* orthologs [23], suggesting that LFY might not regulate C genes in early branching eudicots. Studies in basal angiosperms will be crucial to understand whether LFY already controlled the floral network as a whole in the most recent common ancestor of flowering plants or whether ABC genes progressively came under LFY regulation as angiosperms diversified.

Role of *LFY* in the patterning and evolution of inflorescences

Angiosperm inflorescences consist of either solitary flowers or clusters of flowers with a variety of architectures [24,25] (Figure 3). In (open) racemes, the apical meristem grows indefinitely (i.e. it is indeterminate) and flowers develop from lateral meristems. Cymes show an opposite mode of development because the apical meristem terminates by forming a flower and growth continues from a lateral ('sympodial') meristem that generates the next unit. Panicles take an intermediate position since both apical and lateral meristems form flowers after several branching events. Given the importance of *LFY* in specifying floral meristem identity, it is not surprising that the spatiotemporal regulation of *LFY* activity is a major factor that determines when (flowering time) and where (inflorescence architecture) flowers are formed.

In *Arabidopsis*, which develops an open raceme, the time and the place where flowers form is indeed primarily regulated via the transcription of *LFY*. *LFY* is expressed during the vegetative phase in leaf primordia at steadily

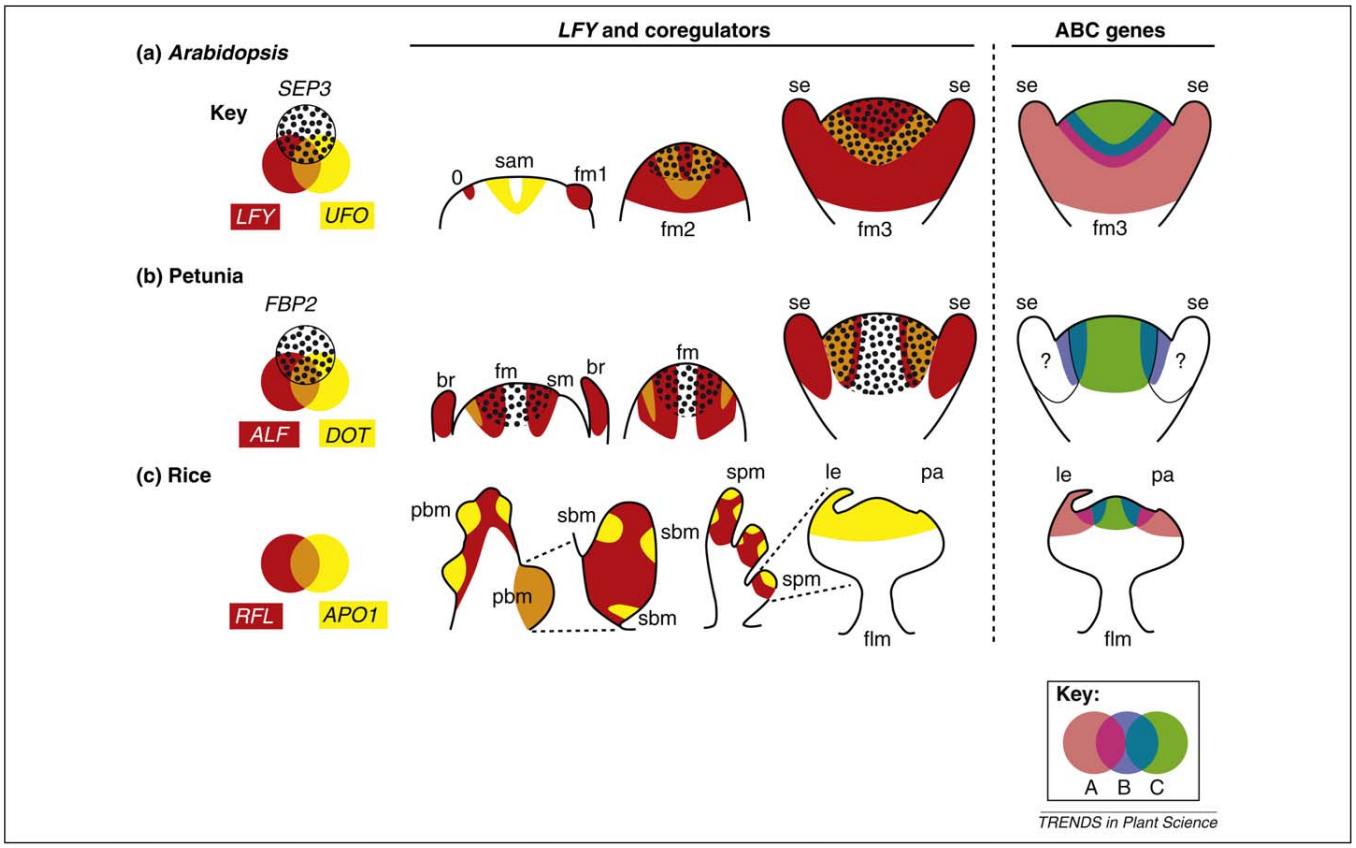


Figure 2. Expression patterns of *LEAFY* genes, some of their coregulators and their main target genes in **(a)** *Arabidopsis*, **(b)** petunia and **(c)** rice. Developing structures leading to early floral bud are schematically described. The color codes are explained in the keys. In rice, glumes have been omitted and only florets are shown. Flower meristems of stages 1, 2 and 3 (fm1, fm2 and fm3, respectively) are numbered according to Ref. [70]. Expression patterns for *LFY* and *UFO* genes have been depicted according to Refs [2,15,18,20,21,31,32]. The expression patterns of the ABC genes have been adapted from Refs [71–74]. Abbreviations: br, bract; flm, floret meristem; fm, floral meristem; le, lemma; pa, palea; pbm, primary branch meristem; sam, shoot apical meristem; sbm, secondary branch meristem; se, sepal; sm, sympodial meristem; spm, spikelet meristem.

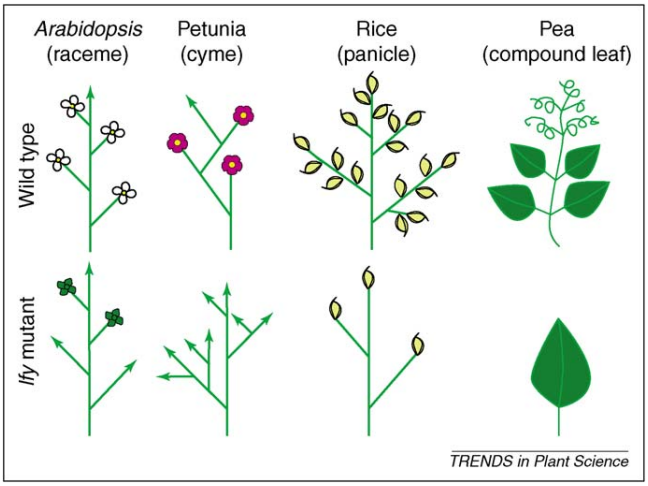


Figure 3. Inflorescence structures in *Arabidopsis*, petunia and rice and in pea leaf shape in wild type and *leafy* mutant plants. Green arrows indicate shoots topped with an indeterminate inflorescence meristem. Green flowers in *Arabidopsis* indicate a shoot or flower intermediate and abnormal flowers.

increasing levels until a threshold is reached, and flowering is triggered via the direct activation of *APETALA1* (*AP1*) [1,26]. Within the inflorescence, *LFY* is expressed in lateral (floral) meristems, whereas it is repressed in the apical inflorescence meristem by *TERMINAL FLOWER 1* (*TFL1*) [27]. Constitutive *LFY* expression converts both apical and axillary meristems into terminal flowers, indicating that the transcriptional regulation of *LFY* is the limiting factor defining when and where flowers are produced [28].

In cymes, floral identity is first specified in apical floral meristems, whereas it is transiently repressed in lateral sympodial meristems, which eventually also acquire floral identity after generating the next lateral primordium. Cymes are common among Solanaceae, and the *LFY* orthologs *NFL* from tobacco (*Nicotiana tabacum*), *FALSIFLORA* (*FA*) from tomato and *ALF* from petunia are expressed in different patterns than *LFY* and *FLO* in the racemose inflorescences of *Arabidopsis* and snapdragon. During early vegetative stages, *NFL*, *FA* and *ALF* are already highly expressed in (incipient) leaf primordia, whereas in the inflorescence meristems their mRNAs appear in the apical region and with a slight delay in the emerging lateral meristem that forms the next sympodial inflorescence unit [29–31]. However, these distinct solanaceous *LFY* transcription patterns do not account for the distinct cymose inflorescence architecture, because constitutive expression of either *LFY* or *ALF* does not alter the inflorescence or flowering time of petunia [15]. Here, LFY activity is restricted in time and space via the *UFO* homolog: *DOT* mRNA is only expressed during flowering and first appears in the apical (floral) meristem, whereas expression in the sympodial meristem is delayed, much more than that of *ALF* (Figure 2). Moreover, ectopic expression of either *DOT* or *UFO* is sufficient to trigger precocious flowering and convert the cymose inflorescence into a solitary flower [15]. In contrast, *Arabidopsis* meristems at the apex of embryos, vegetative shoots and inflorescences express *UFO*, but do not acquire floral identity, because the transcription of *LFY* is limiting, and, for the same reason, constitutive expression of *UFO* does not alter flowering time or inflorescence architecture in *Arabidopsis* [28,32]. Thus, the divergent spatiotemporal expression of floral meristem identity in these cymes and racemes seems to be largely the result of differences in both *LFY* and *UFO* homologs expression patterns (Figures 2 and 3) and is associated with a shift in the restriction of LFY activity in time and space between transcriptional and post-translational control. The directionality of this shift (back, forth or back and forth multiple times) cannot be inferred from the current data on petunia (a Rosid) and *Arabidopsis* (an Asterid) only and requires analysis of additional species across smaller phylogenetic distances.

It is thought that alterations in the expression patterns of developmental genes are a major factor driving the morphological divergence of organisms, but controversy exists as to whether this results mostly from changes in upstream regulatory proteins or, in another species, in the *cis*-regulatory DNA elements that control transcription [33–36]. Current data indicate that both types of alterations could have contributed to the divergence of *LFY* expression in angiosperms. Several species of the Brassicaceae, to which *Arabidopsis* belongs, express their *LFY* homologs within the apical meristem, which, nevertheless, remains indeterminate and does not convert into a flower [37,38]. Whether that is because these species do not express their *UFO* homologs in the apical meristem has not been examined. Studies in which promoter reporter gene constructs were introduced into *Arabidopsis* indicated that the expression pattern of the homologs of *LFY* in three other Brassicaceae are divergent from that of *LFY* in *Arabidopsis*, either because of an alteration in *cis*-regulatory elements (possibly a loss of the element responding to *TFL1*) or because of a yet unidentified difference in the upstream regulatory network [38,39]. The latter difference could involve any of the recently discovered regulators of *LFY* expression [40–44].

LEAFY function during grass inflorescence branching and legume leaf development

In rice, the *LFY* homolog, *RFL*, also controls inflorescence structure, but in a different way than in dicots. The rice inflorescence architecture (a panicle, Figure 3) is made of primary and secondary branches that hold the flowering structures (spikelets). Therefore, there are additional meristem identities in grasses that are absent in model eudicot inflorescences. *RFL* is expressed in the early panicle meristem and is required, together with *APO1*, to maintain its indeterminacy: downregulation of *RFL* or mutation of *APO1* reduce branch number and trigger precocious branch termination with spikelets [18–21]. *RFL* also controls branching at the whole plant level because silencing of *RFL* abolishes the development of tillers (secondary shoots growing from the base of the plant) [19]. As in rice, *ZFL* genes control the branching of the male inflorescence of maize (called the tassel) [17]. These findings show that *LFY* orthologs are able to promote meristematic (indeterminate) growth in grass inflorescences.

Interestingly, a related role was observed in a specific clade of legumes including pea (*Pisum sativum*), lotus

(*Lotus japonicus*) and alfalfa (*Medicago sativa*) [45]. These species generate compound leaves, each consisting of several leaflets, a process that involves the maintenance of a transient meristematic state at the margin of the developing leaves and in some species requires *LFY* activity. For example, when *UNIFOLIATA* (*UNI*, the pea *LFY* ortholog) is mutated [46], the leaves are simpler, sometimes with only one leaflet, showing that *UNI* is required, together with the *UFO* ortholog *Stamina pistilloida* [47], to promote the transient meristematic state required to form multiple leaflets (Figure 3). In other plants with compound leaves, leaf dissection is controlled by *KNOX* genes [45,48], which are also known to play an important role in apical meristem growth.

LFY genes, therefore, appear to promote an indeterminate and meristematic growth in both grass inflorescences and the leaves of some legumes. This could either represent the acquisition of a new function of *LFY* in angiosperms or an ancient role that exists (although sometimes hidden) in a wide array of plants species. Recent results in *Arabidopsis* support the latter hypothesis. Combining mutations in *PENNYWISE* (*PNY*) and *POUND–FOOLISH* (*PNF*) genes led to the production of cauline leaves devoid of axillary meristems. This phenotype is, at least in part, due to the strong *LFY* downregulation because constitutive *LFY* expression in such plants is able to direct flower development in the axils of leaves, as it does in the wild type [40]. Moreover, *pny pnf*/+ *lfy* plants show a high proportion of cauline leaves lacking axillary meristem, a phenotype not seen in a *pny pnf*/+ background [40]. These data nicely show that, in *Arabidopsis* too, *LFY* is able to stimulate meristem growth (in this case axillary meristems) and this function might be important during the early development of floral meristems before they are converted into flowers. We speculate that the role of *LFY* on the development of compound leaves, grasses inflorescence or *Arabidopsis* axillary meristems might reveal, in diverse manners, its capacity to stimulate meristematic growth. We thus propose that *LFY* possesses two functions: promoting meristem growth and conferring floral identity. Depending on the species, the two functions could be obvious (as in maize or pea), the first one might be cryptic (as in *Arabidopsis*) or the second one might be reduced (as in rice).

LEAFY in gymnosperms

In addition to the ortholog of *LFY*, the four groups of extant gymnosperms (ginkgo, cycads, gnetales and conifers) possess a related gene, *NEEDLY* (*NLY*) (first identified in a pine species, *Pinus radiata*, [49]), that was probably lost in the lineage leading to angiosperms [50]. The functions of the two proteins are unknown owing to the lack of gymnosperm mutants, but expression data exist that provide insights into the role of the family in non-flowering plants.

Gymnosperms do not form flowers but display their male and female organs on two distinct axes called strobili or cones, with the exception of some gnetalean taxa that have bisexual compound cones. *LFY* and/or *NLY* expression has been analyzed in six conifer species, one gnetale and *Ginkgo biloba* [49,51–57] and is commonly seen in vegetative tissues such as shoot apical meristem, stem and leaves or needles, but their role in these tissues is unknown. One interesting feature that is shared by all the gymnosperms that have been examined so far is the upregulation of *LFY* and/or *NLY* in axillary meristems, independently of their vegetative or reproductive fate. This suggests that the two proteins could be involved in the establishment of lateral meristems but are, in themselves, not sufficient to confer a reproductive status. *AGL6*-like genes, which form a clade that is sister to *SEP* genes [58] and have a role similar to *SEP3* in rice and petunia [22,59], are expressed in reproductive meristems of conifers and gnetales and, thus, are possible candidates to act together with LFY in the specification of reproductive identity in gymnosperms [51].

Do *LFY* genes also regulate MADS-box gene expression in gymnosperms? Whereas A genes and *UFO* orthologs have not been described in gymnosperms, orthologs of B and C genes do exist and are expressed in male and female gymnosperm structures, suggesting that they might, as in angiosperms, contribute to specify their identity [60]. Evidence that B and C orthologs are also regulated by LFY or NLY remains circumstantial. In early developmental stages of the reproductive meristem, expression of both *LFY* and *NLY* in nascent primordia generally precedes and encompasses that of B and C genes, consistent with a possible regulatory role. In later stages, *LFY* and *NLY* domains sometimes stop overlapping to become mutually exclusive. For example, in Norway spruce (*Picea abies*) male cones, *PaLFY* is detected in the sporogenous cells only, whereas *PaNLY* continues to be expressed in the surrounding tissues [57,61]. In female cones of at least three conifer species, *LFY* remains expressed in the ovule primordia coinciding with the C gene *DAL2*, whereas *NLY* is expressed in complementary tissues (cone axis and sterile scale) that are devoid of *DAL2* transcripts, suggesting that the two paralogs could regulate distinct sets of genes.

Gymnosperms with their two *LFY*-like genes that persisted for an extended period of time are an exception among land plants. This peculiarity could be explained by the distinct expression patterns of the two paralogs and possibly a different role. Indeed, *NLY* could have acquired a novel function and diverged slightly from LFY as suggested by the observation that *NLY* is not as efficient as gymnosperm *LFY* at complementing an *Arabidopsis lfy* mutant [52,56,62]. It is still unknown whether modifications in protein stability, affinity for DNA, sequence-specific recognition or interaction with coregulators account for differences between the two paralogs.

More research is needed to understand the role of *LFY* homologs in gymnosperms, but there is a growing body of evidence suggesting that a minimal network involving LFY and some ABC-type MADS-box proteins was already at work in the reproductive structures of the most recent common ancestor of seed plants. This is of crucial importance because the subsequent rearrangement of such a network after the divergence of gymnosperms is likely to be one of the causative forces of the origin of flowers in angiosperms [63]. Insights into the function of LFY before the evolution of the seed habit can be gained by studying living plants that have retained the free-sporing habit, for example, ferns and mosses.

Back to LEAFY origins: the situation in free-sporing land plants

The *LFY* gene is also present in free-sporing land plant [i.e. lycophytes, ferns and their allies and bryophytes (moss, hornworts and liverworts)] [64,65]. However, the reproductive organs of these plants are so different from flowers that it is difficult to compare the role of LFY in these groups with that in seed plants.

A single functional analysis has been performed in the moss *Physcomitrella patens*. The two close *PpLFY* paralogs are broadly expressed in both the sporophyte (diploid) and the gametophyte (haploid) [65]. When both *PpLFY* genes are mutated, the gametophyte develops normally, whereas the sporophyte is arrested at the first cell division stage right after fertilization. The few sporophytes that do form have general growth defects, suggesting that PpLFY proteins play an important role in the control of cell division during the diploid phase.

Based on the high amino acid conservation of the DNA binding domain, it is reasonable to propose that LFY also functions as a transcription factor in ferns and bryophytes. Indeed, *CrLFY* from the fern *Ceratopteris richardii* can partially complement *Arabidopsis lfy* mutants and the CrLFY protein binds a canonical LFY-binding site CCANT(G/T) *in vitro* [62]. By contrast, PpLFY is inactive in *Arabidopsis* and the protein lacks the capacity of binding this canonical LFY-binding site because an aspartic acid residue (D) replaces a conserved histidine (H) in the DNA binding domain. This feature appears to have been derived in the moss lineage given that the liverwort *Marchantia polymorpha*, which represents a lineage that is sister to all other land plants, resembles vascular plants in possessing the conserved H. It is thus likely that LFY acts as a transcription factor in free-sporing land plants with a slight variation of its mode of action; however, its targets remain unknown. Despite intensive searches, direct orthologs of A, B and C genes have not been identified in bryophytes, ferns and their allies [66–68]. The well-known 'floral' MADS-box genes probably arose later, during the expansion of the MADS family, and are thus specific to seed plants. More studies are needed to identify LFY target genes in early land plants. Such approaches could provide precious information about how LFY fulfilled its role 400 million years ago.

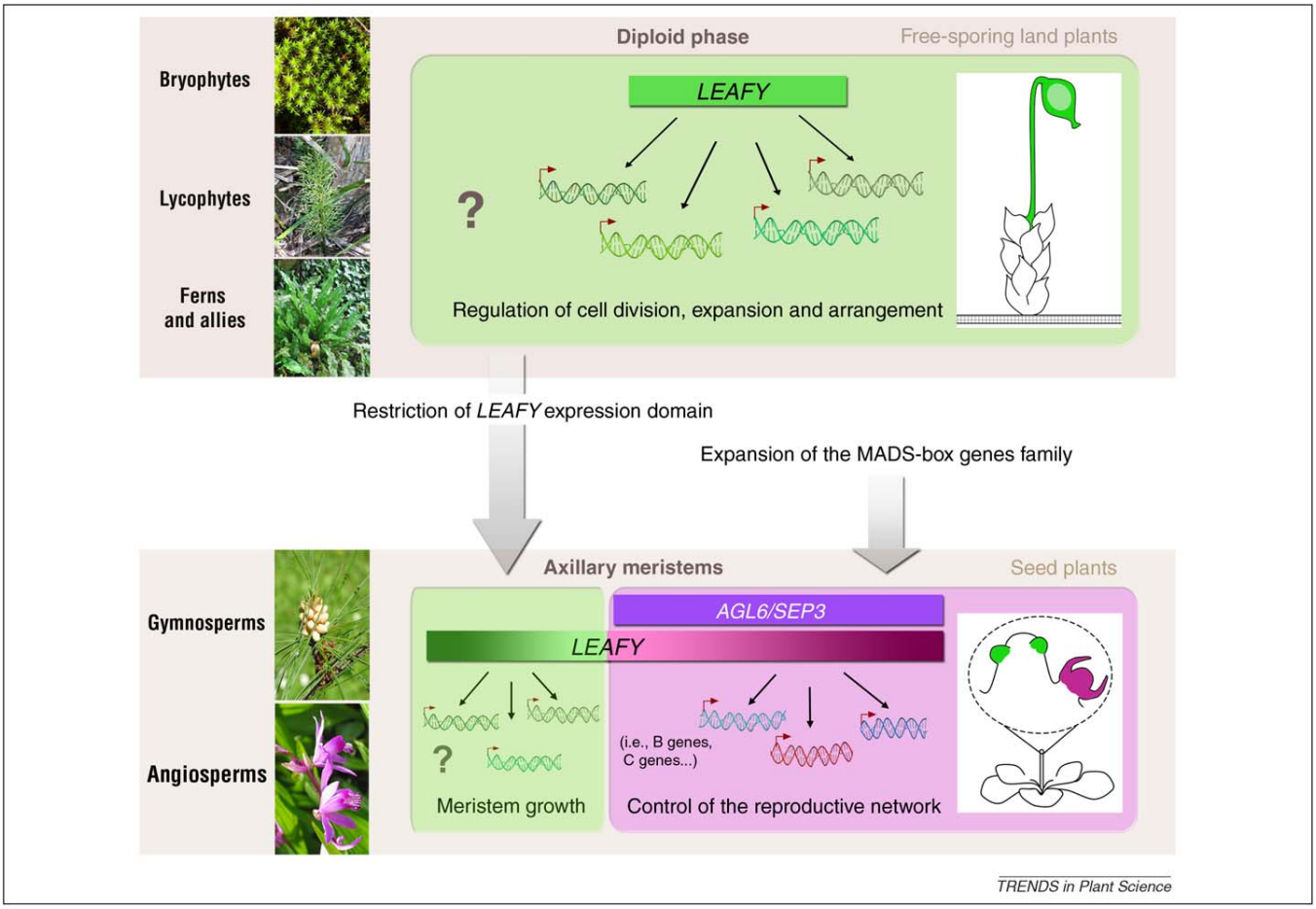


Figure 4. Speculative scenario describing the evolution of LEAFY function. In free-sporing plants such as bryophytes and ferns, *LFY* is broadly expressed throughout the diploid phase (sporophyte, colored in green) where it could act as a regulator of cell divisions, expansion and arrangement (ancestral function) but its targets are still unknown. In seed plants, *LFY* expression is detected at high levels in axillary meristems where it might still exert its control on cell division (young axillary meristems arising from the shoot apex, colored in green). The role of LFY in the inflorescence of grasses or the compound leaves of some legumes indicates that its ancestral function remains active in flowering plants. In the floral meristem of angiosperms (colored in pink), LFY also drives a reproductive network (modern function) that generates the different organs of the flower. This second function involves some 'seed plant specific' MADS box genes both as partners (*SEP3* and/or *AGL6*) or targets (ABC genes) of LFY. Some of the actors of the reproductive network are present in gymnosperms (*LEAFY*, B genes, C genes) and could fulfill a similar function, so a pre-floral version of the reproductive network is likely to exist in this group. However, regulations within this pre-floral network remain to be understood because so far *SEP* genes have only been found in angiosperms, whereas *AGL6*-like genes have also been found in gymnosperms.

The evolution of LEAFY functions: a speculative scenario

In early land plants, *LFY* homologs might have functioned in a minimal network regulating cell division and expansion throughout the sporophyte (Figure 4). *LFY* targets in such plants remain to be identified, but genes involved in cell cycle regulation, growth or differentiation, as well as hormonal signaling, are good candidates.

As vascular plants diversified, *LFY* expression patterns might have undergone a restriction, by changes in *cis*-regulatory elements or in the upstream regulatory network, so that in seed plants *LFY* genes remained inactive in most vegetative tissues and high levels of expression of these genes are first detected in axillary meristems arising on the flank of the shoot apex. Given that meristems are groups of undifferentiated cells that divide intensively, we hypothesize that LFY still exerts its ancestral role on the regulation of cell division in gymnosperms and angiosperms, but in a territory restricted to axillary meristems. This function, which is obvious in the grass inflorescence and more cryptic (but present) in *Arabidopsis* flowers, would have been recruited in some legumes to create compound leaves.

In seed plants, one novel LFY function would be to trigger a reproductive gene network (Figure 4). As shown in model angiosperms, this involves MADS-box transcription factors both as downstream LFY targets (ABC genes) and as LFY partners (AGL6 and/or SEP3). These MADS-box genes are absent from free-sporing plants and probably originated from a diversification of the MADS family in the lineage leading to seed plants. When exactly this modern network appeared remains elusive because its existence in gymnosperms has not been demonstrated. Its emergence probably involved changes in *cis*-elements of recruited targets, to place them under LFY control, as well as the establishment of novel protein–protein interactions. In most living angiosperms, both LFY functions would be present, and evolution has used all the inherent potential of the associated regulatory networks to create the wide variety of inflorescence structures observed in nature.

We are aware that this scenario is speculative, but our aim was to provide a framework that integrated the wealth of recent data. Now that we know more about the phylogeny of extant seed plants, developing new model species in gymnosperms as well as working with several well-placed angiosperm groups (e.g. *Amborella*, Nymphaeales, Piperales, Alismatales, Ranunculids) [69] is key to fully understanding the fascinating history of this peculiar gene and the evolution of the network that regulates floral identity.

Acknowledgements

We thank P. Laufs, M. Frohlich, J. Hofer, M. Blázquez and members of the Koes and Parcy laboratories for discussion, E. Coen and G. Tichtinsky for providing pictures, E. Dorcey for critical reading of the manuscript and the referees for constructive comments. Research in our laboratories is supported by funding from the Centre National de la Recherche Scientifique (CNRS, Action Thématique et Incitative sur Programme, F.P.), the Agence Nationale de la Recherche (ANR, Plant-TFcode, F.P.), the ANR and the Biotechnology and Biological Sciences Research Council (Flower Model, F.P.), and of the Netherlands Organisation for Scientific Research (NWO) to R.K.

References

1 Benlloch, R. *et al.* (2007) Floral initiation and inflorescence architecture: a comparative view. *Ann. Bot. (Lond.)* 100, 659–676
2 Blazquez, M.A. *et al.* (2006) How floral meristems are built. *Plant Mol. Biol.* 60, 855–870
3 Liu, C. *et al.* (2009) Coming into bloom: the specification of floral meristems. *Development* 136, 3379–3391
4 Moyroud, E. *et al.* (2009) The LEAFY floral regulators in Angiosperms: conserved proteins with diverse roles. *J. Plant Biol.* 52, 177–185
5 Wagner, D. (2009) Flower morphogenesis: timing is key. *Dev. Cell* 16, 621–622
6 Irish, V.F. (2010) The flowering of *Arabidopsis* flower development. *Plant J.* 61, 1014–1028
7 Schmid, M. *et al.* (2003) Dissection of floral induction pathways using global expression analysis. *Development* 130, 6001–6012
8 William, D.A. *et al.* (2004) Genomic identification of direct target genes of LEAFY. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1775–1780
9 Hamès, C. *et al.* (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *EMBO J.* 27, 2628–2637
10 Liu, C. *et al.* (2009) Regulation of floral patterning by flowering time genes. *Dev. Cell* 16, 711–722
11 Chae, E. *et al.* (2008) An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. *Development* 135, 1235–1245
12 Kodadek, T. *et al.* (2006) Keeping transcriptional activators under control. *Cell* 127, 261–264
13 Simon, R. *et al.* (1994) Fimbriata controls flower development by mediating between meristem and organ identity genes. *Cell* 78, 99–107
14 Lippman, Z.B. *et al.* (2008) The making of a compound inflorescence in tomato and related nightshades. *PLoS Biol.* 6, e288
15 Souer, E. *et al.* (2008) Patterning of inflorescences and flowers by the F-Box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of Petunia. *Plant Cell* 20, 2033–2048
16 Zhang, S. *et al.* (2003) *Proliferating Floral Organs* (*Pfo*), a *Lotus japonicus* gene required for specifying floral meristem determinacy and organ identity, encodes an F-box protein. *Plant J.* 33, 607–619
17 Bomblies, K. *et al.* (2003) Duplicate *FLORICAULA/LEAFY* homologs *zfl1* and *zfl2* control inflorescence architecture and flower patterning in maize. *Development* 130, 2385–2395
18 Kyozuka, J. *et al.* (1998) Down-regulation of *RFL*, the *FLO/LFY* homolog of rice, accompanied with panicle branch initiation. *Proc. Natl. Acad. Sci. U. S. A.* 95, 1979–1982
19 Rao, N.N. *et al.* (2008) Distinct regulatory role for *RFL*, the rice *LFY* homolog, in determining flowering time and plant architecture. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3646–3651
20 Ikeda, K. *et al.* (2007) Rice *ABERRANT PANICLE ORGANIZATION 1*, encoding an F-box protein, regulates meristem fate. *Plant J.* 51, 1030–1040
21 Ikeda-Kawakatsu, K. *et al.* (2009) Expression level of *ABERRANT PANICLE ORGANIZATION1* determines rice inflorescence form through control of cell proliferation in the meristem. *Plant Physiol.* 150, 736–747
22 Ohmori, S. *et al.* (2009) *MOSAIC FLORAL ORGANS1*, an *AGL6*-like MADS box gene, regulates floral organ identity and meristem fate in rice. *Plant Cell* 21, 3008–3025
23 Becker, A. *et al.* (2005) Floral and vegetative morphogenesis in California poppy (*Eschscholzia californica Cham.*). *Int. J. Plant Sci.* 166, 537–555
24 Prenner, G. *et al.* (2009) The key role of morphology in modelling inflorescence architecture. *Trends Plant Sci.* 14, 302–309
25 Prusinkiewicz, P. *et al.* (2007) Evolution and development of inflorescence architectures. *Science* 316, 1452–1456
26 Parcy, F. (2005) Flowering: a time for integration. *Int. J. Dev. Biol.* 49, 585–593
27 Bradley, D. *et al.* (1997) Inflorescence commitment and architecture in *Arabidopsis*. *Science* 275, 80–83
28 Weigel, D. and Nilsson, O. (1995) A developmental switch sufficient for flower initiation in diverse plants. *Nature* 377, 495–500
29 Kelly, A.J. *et al.* (1995) *NFL*, the tobacco homolog of *FLORICAULA* and *LEAFY*, is transcriptionally expressed in both vegetative and floral meristems. *Plant Cell* 7, 225–234

30 Molinero-Rosales, N. *et al.* (1999) *FALSIFLORA*, the tomato orthologue of *FLORICAULA* and *LEAFY*, controls flowering time and floral meristem identity. *Plant J.* 20, 685–693
31 Souer, E. *et al.* (1998) Genetic control of branching pattern and floral identity during *Petunia* inflorescence development. *Development* 125, 733–742
32 Lee, I. *et al.* (1997) A *LEAFY* co-regulator encoded by *UNUSUAL FLORAL ORGANS*. *Curr. Biol.* 7, 95–104
33 Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25–36
34 Hoekstra, H.E. and Coyne, J.A. (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61, 995–1016
35 Pennisi, E. (2008) Evolutionary biology. Deciphering the genetics of evolution. *Science* 321, 760–763
36 Stern, D.L. and Orgogozo, V. (2008) The loci of evolution: how predictable is genetic evolution? *Evolution* 62, 2155–2177
37 Shu, G. *et al.* (2000) LEAFY and the evolution of rosette flowering in violet cress (*Jonopsidium acaule*, *Brassicaceae*). *Am. J. Bot.* 87, 634–641
38 Yoon, H.S. and Baum, D.A. (2004) Transgenic study of parallelism in plant morphological evolution. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6524–6529
39 Sliwinski, M.K. *et al.* (2007) The role of two *LEAFY* paralogs from *Idahoa scapigera* (Brassicaceae) in the evolution of a derived plant architecture. *Plant J.* 51, 211–219
40 Kanrar, S. *et al.* (2008) Regulatory networks that function to specify flower meristems require the function of homeobox genes *PENNYWISE* and *POUND-FOOLISH* in *Arabidopsis*. *Plant J.* 54, 924–937
41 Karim, M.R. *et al.* (2009) A role for *Arabidopsis PUCHI* in floral meristem identity and bract suppression. *Plant Cell* 21, 1360–1372
42 Lee, J. *et al.* (2008) SOC1 translocated to the nucleus by interaction with AGL24 directly regulates *LEAFY*. *Plant J.* 55, 832–843
43 Liu, C. *et al.* (2008) Direct interaction of AGL24 and SOC1 integrates flowering signals in *Arabidopsis*. *Development* 135, 1481–1491
44 Yamaguchi, A. *et al.* (2009) The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of *LEAFY*, *FRUITFULL*, and *APETALA1*. *Dev. Cell* 17, 268–278
45 Champagne, C.E. *et al.* (2007) Compound leaf development and evolution in the legumes. *Plant Cell* 19, 3369–3378
46 Hofer, J. *et al.* (1997) *UNIFOLIATA* regulates leaf and flower morphogenesis in pea. *Curr. Biol.* 7, 581–587
47 Taylor, S. *et al.* (2001) *Stamina pistilloida*, the pea ortholog of *Fim* and *UFO*, is required for normal development of flowers, inflorescences, and leaves. *Plant Cell* 13, 31–46
48 Blein, T. *et al.* (2008) A conserved molecular framework for compound leaf development. *Science* 322, 1835–1839
49 Mouradov, A. *et al.* (1998) *NEEDLY*, a *Pinus radiata* ortholog of *FLORICAULA/LEAFY* genes, expressed in both reproductive and vegetative meristems. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6537–6542
50 Frohlich, M.W. and Parker, D.S. (2000) The mostly male theory of flower evolutionary origins: from genes to fossils. *Syst. Bot.* 25, 155–170
51 Carlsbecker, A. *et al.* (2004) The MADS-box gene *DAL1* is a potential mediator of the juvenile-to-adult transition in Norway spruce (*Picea abies*). *Plant J.* 40, 546–557
52 Dornelas, M.C. and Rodriguez, A.P. (2005) A *Floricaula/Leafy* gene homolog is preferentially expressed in developing female cones of the tropical pine *Pinus caribaea* var. *caribaea*. *Genet. Mol. Biol.* 28, 299–307
53 Guo, C.L. *et al.* (2005) Expressions of LEAFY homologous genes in different organs and stages of *Ginkgo biloba*. *Yi Chuan* 27, 241–244
54 Mellerowicz, E.J. *et al.* (1998) *PRFLL* – a *Pinus radiata* homologue of *FLORICAULA* and *LEAFY* is expressed in buds containing vegetative shoot and undifferentiated male cone primordia. *Planta* 206, 619–629
55 Shindo, S. *et al.* (2001) Characterization of a *FLORICAULA/LEAFY* homologue of *Gnetum parvifolium* and its implications for the evolution of reproductive organs in seed plants. *Int. J. Plant Sci.* 162, 1199–1209
56 Shiokawa, T. *et al.* (2008) Isolation and functional analysis of the *CjNdly* gene, a homolog in *Cryptomeria japonica* of *FLORICAULA/LEAFY* genes. *Tree Physiol.* 28, 21–28
57 Vazquez-Lobo, A. *et al.* (2007) Characterization of the expression patterns of *LEAFY/FLORICAULA* and *NEEDLY* orthologs in female and male cones of the conifer genera *Picea*, *Podocarpus*, and *Taxus*: implications for current evo-devo hypotheses for gymnosperms. *Evol. Dev.* 9, 446–459
58 Zahn, L.M. *et al.* (2005) The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* 169, 2209–2223
59 Rijpkema, A.S. *et al.* (2009) The petunia *AGL6* gene has a *SEPALLATA*-like function in floral patterning. *Plant J.* 60, 1–9
60 Theissen, G. and Becker, A. (2004) Gymnosperms orthologs of class B floral homeotic genes and their impact on understanding flower origin. *Crit. Rev. Plant Sci.* 23, 129–148
61 Sundstrom, J. and Engstrom, P. (2002) Conifer reproductive development involves B-type MADS-box genes with distinct and different activities in male organ primordia. *Plant J.* 31, 161–169
62 Maizel, A. *et al.* (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. *Science* 308, 260–263
63 Frohlich, M.W. (2003) An evolutionary scenario for the origin of flowers. *Nat. Rev. Genet.* 4, 559–566
64 Himi, S. *et al.* (2001) Evolution of MADS-box gene induction by *FLO/LFY* genes. *J. Mol. Evol.* 53, 387–393
65 Tanahashi, T. *et al.* (2005) Diversification of gene function: homologs of the floral regulator *FLO/LFY* control the first zygotic cell division in the moss *Physcomitrella patens*. *Development* 132, 1727–1736
66 Münster, T. *et al.* (2002) Evolutionary aspects of MADS-box genes in the eusporangiate fern *Ophioglossum*. *Plant Biol.* 4, 474–483
67 Singer, S.D. *et al.* (2007) Clues about the ancestral roles of plant MADS-box genes from a functional analysis of moss homologues. *Plant Cell Rep.* 26, 1155–1169
68 Tanabe, Y. *et al.* (2003) Characterization of the *Selaginella remotifolia* MADS-box gene. *J. Plant Res.* 116, 71–75
69 Rudall, P.J. *et al.* (2009) Nonflowers near the base of extant angiosperms? Spatiotemporal arrangement of organs in reproductive units of Hydatellaceae and its bearing on the origin of the flower. *Am. J. Bot.* 96, 67–82
70 Smyth, D.R. *et al.* (1990) Early flower development in *Arabidopsis*. *Plant Cell* 2, 755–767
71 Krizek, B.A. and Fletcher, J.C. (2005) Molecular mechanisms of flower development: an armchair guide. *Nat. Rev. Genet.* 6, 688–698
72 Kyozuka, J. *et al.* (2000) Spatially and temporally regulated expression of rice MADS box genes with similarity to *Arabidopsis* class A, B and C genes. *Plant Cell Physiol.* 41, 710–718
73 Yamaguchi, T. and Hirano, H.Y. (2006) Function and diversification of MADS-box genes in rice. *Scientific World J.* 6, 1923–1932
74 Yamaguchi, T. *et al.* (2006) Functional diversification of the two C-class MADS box genes OSMADS3 and OSMADS58 in *Oryza sativa*. *Plant Cell* 18, 15–28

3 LEAFY, un facteur de transcription unique

Alors qu'il existe de nombreuses données sur la fonction de LFY et ses homologues chez les angiospermes, on connaît paradoxalement très peu de choses sur la façon dont la protéine remplit ses fonction.

LFY et ses homologues possèdent deux domaines en position N et C-terminale extrêmement conservés des mousses aux plantes à fleurs. Toutes les mutations non-synonymes identifiées lors de recherche de mutants d'*Arabidopsis* sont localisées dans l'un ou l'autre de ces deux domaines ce qui souligne leur importance fonctionnelle mais les propriétés de ces régions n'ont été que peu étudiées : la partie C-terminale de la protéine semble contenir le domaine de liaison à l'ADN (Maizel et al., 2005). Le domaine N-terminal en revanche n'a pas de fonction connue et il a été successivement proposé comme participant à la multimérisation de la protéine (Busch and Gleissberg, 2003) à l'interaction avec des partenaires (Souer et al., 2008) ou à la régulation de l'activité du domaine C-terminal (Maizel et al., 2005). La protéine ne possédant pas de pouvoir intrinsèque pour réguler la transcription, LFY sollicite des partenaires pour influer sur l'expression de ces gènes cibles (Parcy et al., 1998). Le partenariat LFY-UFO est fortement établi d'un point de vue génétique chez *Arabidopsis* et d'autres espèces comme le pétunia mais l'interaction physique entre les deux protéines est beaucoup moins claire puisque selon les auteurs UFO contacterait LFY au niveau du domaine N-terminal ou C-terminal (Lee et al., 1997; Parcy et al., 1998; Souer et al., 1998; Chae et al., 2008). De même, les conséquences moléculaires de cette interaction sont incomprises : *UFO* code pour une protéine à F-box impliquée dans l'ubiquitination des protéines et leur dégradation par le protéasome et il est difficile de réconcilier cette fonction avec l'activation des gènes B par LFY. *WUS* et *SEP3* ont également été décrits comme des cofacteurs de LFY (Lohmann et al., 2001; Liu et al., 2009b). Une interaction entre ces protéines et LFY fait sens car il s'agit également de facteurs de transcription et on peut imaginer que la formation d'hétérocomplexes LFY-WUS ou LFY-SEP3 permettrait de cibler différentiellement les gènes régulés. Malheureusement, l'existence de tels complexes n'a jamais été formellement démontrée.

Le fonctionnement de LFY est en grande partie incompris car les relations LFY-ADN sont elles-mêmes méconnues : LFY est capable de reconnaître plusieurs éléments *cis* des régions régulatrices d'*AP1*, *AP3* et *AG* mais les motifs nucléotidiques identifiés ne partagent que peu

de positions en commun, par conséquent les éléments reconnus par LFY dans le génome ne sont pas aisément identifiables (Parcy et al., 1998; Busch et al., 1999; Lamb et al., 2002). Des approches *in vivo* ont été menées pour tenter d'identifier le répertoire complet des gènes contrôlés par LFY mais établir une image exhaustive reste difficile : en 2003, Schmid et collaborateurs ont disséqué les voies moléculaires contrôlant la floraison en comparant le transcriptome de plantes sauvages et mutantes pour *LFY* au cours de la transition florale. Cette étude a pointé un certain nombre de gènes dérégulés lorsque *LFY* est inactif mais n'a pas permis de distinguer les cibles directement ou indirectement contrôlées par le facteur (Schmid et al., 2003). En 2004, Williams et al. ont expérimenté une méthode permettant d'identifier exclusivement les cibles directes de LFY : 14 gènes d'intérêt ont ainsi été identifiés mais leur contribution à l'organogenèse florale reste largement spéculative (William et al., 2004). Plusieurs fonctions autres que la floraison ont été attribuées à LFY mais là encore aucune cible directe du facteur n'a été identifiée, la fonction du gène est donc postulée à partir de l'étude des phénotypes mutants mais on ignore comment ces nouveaux rôles sont réalisés (Hofer et al., 1997; Tanahashi et al., 2005; Rao et al., 2008).

Les informations en provenance des espèces non angiospermes suggèrent une évolution du patron d'expression et des spécificités de liaison à l'ADN des homologues de LFY au cours de l'histoire évolutive de la lignée verte (Himi et al., 2001; Tanahashi et al., 2005). Comprendre les conséquences de ces changements sur la fonction de ce régulateur central du développement floral et déterminer comment des modifications affectant LFY ont pu contribuer à l'apparition des premières plantes à fleurs requiert donc une meilleure connaissance des propriétés de cette protéine chez l'organisme modèle et les espèces emblématiques de l'évolution des végétaux terrestres.

Objectifs

La naissance des plantes à fleurs est une énigme que la paléobotanique n'a toujours pas résolue. L'identification chez les espèces modèles des régulateurs essentiels du développement floral permet d'imaginer les étapes ayant conduit à la création de cette structure innovante. Ainsi, le gène LEAFY, régulateur central de l'organogenèse florale, pourrait être impliqué dans l'émergence même des angiospermes. Or, l'organisation structurale et le mode d'action de ce facteur de transcription original sont toujours inconnus.

Afin de déterminer le(s) rôle(s) évolutif(s) joué(s) par LFY nous avons cherché à comprendre, chez l'espèce modèle Arabidopsis thaliana*, le mode d'action de cette protéine unique puis, nous avons étudié comment ses propriétés biochimiques (patron d'expression, spécificité par rapport à un motif reconnu, multimérisation) ont évolué au cours de l'histoire des plantes terrestres et contribué à la création d'innovations.*

Les résultats obtenus au cours de ce travail de thèse sont présentés au travers de quatre chapitres de résultats qui visent donc à répondre aux deux grandes questions suivantes :

▸ Quelles sont les propriétés biochimiques de la protéine LEAFY d'*Arabidopsis thaliana* ?

Dans un premier temps, nous avons cherché à comprendre comment LFY contactait l'ADN (Chapitre 1). Puis, j'ai essayé d'élucider les règles qui gouvernent la reconnaissance spécifique des éléments-*cis* du génome par LFY (Chapitre 2), première étape dans le contrôle de l'expression des gènes.

▸ Comment les propriétés et le rôle de LEAFY et ses homologues ont évolué chez les plantes terrestres ?

LFY étant présent chez les plantes sans fleur, j'ai cherché à établir si les propriétés biochimiques de la protéine et ses homologues avaient varié au cours de l'histoire des plantes terrestres. Dans un premier temps, j'ai testé si les homologues de LFY chez les angiospermes basales, les gymnospermes et les mousses présentaient une spécificité de liaison différente par rapport à la protéine d'Arabidopsis (Chapitre 2). Comme les motifs ADN reconnus par LFY

appartiennent à des séquences régulatrices beaucoup plus grandes, le contexte promoteur peut influer sur la capacité d'un facteur de transcription à reconnaître l'ADN, nous avons donc voulu tester s'il est possible d'étudier expérimentalement l'interaction entre une protéine et un long fragment d'ADN (Chapitre 3). Enfin, j'ai exploré l'existence d'un réseau pré-floral, impliquant les homologues de LFY et des gènes homéotiques B et C, chez les gymnospermes pour déterminer si des modifications de ce réseau étaient survenues lors de l'apparition des angiospermes et identifier un rôle éventuel de LFY pour créer la première fleur (Chapitre 4).

Pour tenter de répondre à ces questions, nous avons dû mettre au point un certain nombre d'outils inédits, ces avancées techniques ont permis l'acquisition des données présentées et sont donc explicitées au travers des différents chapitres.

CHAPITRE 1

Structure et mode de formation du complexe LEAFY-ADN

*LEAFY est un régulateur central du développement floral, pourtant, le fonctionnement de cette protéine est peu connu. Nous avons donc cherché à élucider les mécanismes de l'interaction LEAFY-ADN par une caractérisation biochimique de la protéine d'*Arabidopsis thaliana*. La résolution de la structure du domaine de liaison de LEAFY complexé à un ADN cible (travail collaboratif mené par Cécile Hamès, étudiante en thèse dans notre équipe) a permis la publication d'un article dont je suis co-auteur. Les enseignements tirés de cette étude structurale ont ensuite été réexaminés à la lumière de l'évolution. Ces analyses ont généré de nouvelles interrogations à l'origine des travaux présentés dans les chapitres suivants.*

Introduction

Élucider les mécanismes à l'origine de l'interaction entre LEAFY (LFY) et l'ADN est indispensable pour comprendre les processus sous-jacents à la floraison et déterminer quelle fut la contribution de ce facteur à l'invention de la fleur.

En effet, la transition florale est un phénomène développemental brusque. Or, le niveau de transcription du gène *LFY* en revanche, augmente progressivement lors du passage de la phase végétative à la phase reproductive. Dès lors, il a été proposé que la floraison serait enclenchée à condition qu'un certain seuil de LFY soit franchi (Blazquez et al., 1997). Le mode d'action de ce facteur n'étant pas connu, rien ne permet d'expliquer cependant que LFY ne soit efficace qu'au-dessus d'une concentration critique. Par ailleurs, des homologues de ce gène sont systématiquement retrouvés chez les végétaux ayant divergé avant les angiospermes et donc dépourvus de fleurs. S'il est vrai que *LFY* a contribué à la création de la première fleur, cela implique par conséquent une modification de son comportement par rapport à la protéine ancestrale. Comprendre comment LFY fonctionne chez une plante de référence constitue alors un pré-requis nécessaire pour tester expérimentalement les scénarios évolutifs imaginés.

LFY et ses homologues possèdent deux domaines basiques extrêmement conservés tout au long de la lignée verte, mais ni l'un ni l'autre ne ressemblent à des domaines protéiques connus. Par conséquent, la structure de la protéine ne peut pas être inférée. Récemment, il a été montré que le domaine de liaison à l'ADN était porté par le domaine C-terminal de la protéine (Maizel et al., 2005). En revanche, la stœchiométrie du complexe LFY-ADN; c'est-à-dire le nombre de protéines impliquées dans la reconnaissance spécifique d'un élément *cis*, n'a jamais été décrite. De même, la nature précise des contacts entre les deux partenaires et la

façon dont ceux-ci s'établissent sont toujours inconnus. Afin d'obtenir des informations sur le fonctionnement de ce facteur unique, nous avons donc entrepris une étude biochimique et structurale du domaine de liaison à l'ADN de la protéine LFY d'*Arabidopsis thaliana*.

Résultats principaux

L'ensemble des résultats obtenus dans le cadre de notre étude est présenté en détail dans l'article qui suit. Seules les conclusions les plus marquantes sont résumées ici.

Pourquoi LEAFY est un interrupteur de la floraison ?

La stratégie expérimentale a été de surproduire en système bactérien le domaine C-terminal de la protéine LFY d'*Arabidopsis* (LFY-C). Après purification, nous avons observé que LFY-C est un monomère en solution mais lie l'ADN de façon coopérative sous forme dimérique: un premier monomère reconnaît l'ADN cible ce qui facilite la fixation d'un second monomère pour constituer un complexe stable. Or, il a été décrit qu'une liaison coopérative (couplée à des mécanismes d'autorégulation) favorise les transitions développementales brusques (Lebrecht et al., 2005; Gregor et al., 2007). LFY participant à une boucle de d'autorégulation avec AP1, la coopérativité de liaison révélée pourrait contribuer au caractère abrupt de la transition florale.

Comment s'organise l'interface LFY-ADN ?

La séquence consensus décrite jusqu'ici comme reconnue par LFY mettait en jeu 7 nucléotides. Contre toute attente, l'analyse détaillée de la structure du complexe dimère-ADN a révélé que les contacts ADN-protéine s'étendaient en fait sur 19 nucléotides. LFY-C possède en particulier un motif de type Hélice-Tour-Hélice (HTH) qui permet à la protéine de contacter la molécule d'ADN au niveau du petit et du grand sillon. Nous avons donc conclu que le consensus CCANT(G/T) ne permettait plus de décrire de manière réaliste le motif nucléotidique préféré par LFY et nous avons proposé un second modèle, (A/T)NNNNCCANTG(G/T)NNNN(A/T) qui tient compte des données nouvelles apportées par la structure.

Quelles sont les origines de LEAFY et pourquoi un tel degré de conservation ?

LEAFY a été identifié chez toutes les espèces de plantes chez lesquelles il a été recherché, en revanche aucune trace de la protéine n'a été trouvée chez les algues vertes, groupe auquel appartient l'ancêtre aquatique des végétaux terrestres. Connaissant désormais la structure de

LFY-C, nous avons proposé que le domaine C-terminal de LFY dériverait d'une protéine HTH ancestrale de type protéine à homéodomaine. L'absence de similarité de séquence ne permet pas cependant de le démontrer formellement et il pourrait s'agir d'une convergence évolutive. Il est intéressant de noter que les protéines à homéodomaine du règne végétal contrôlent l'homéostasie des méristèmes et la division cellulaire, assurant ainsi une fonction similaire au second rôle proposé pour LFY et ses homologues (Moyroud et al., 2010).
L'étude structurale a identifié 14 acides aminés de LFY-C au contact de l'ADN tandis que trois autres sont impliqués dans l'interaction entre les deux monomères. Ces résidus sont répartis sur l'ensemble de la séquence du domaine C-terminal et seul le repliement de la protéine, reposant sur les structures secondaires constituées des autres acides aminés, permet de positionner les résidus-clés dans une configuration adéquate pour la reconnaissance spécifique des nucléotides. Ainsi, toute modification de séquence du domaine C-terminal a de grandes chances de déstabiliser l'ensemble de la structure et de mener à une protéine non-fonctionnelle. Une telle situation peut expliquer la conservation remarquable du domaine protéique de liaison à l'ADN.

Contexte de l'étude

Cette étude a constitué le projet fondateur de notre équipe et a justifié son implantation sur le site grenoblois : l'expertise biochimique présente au sein du laboratoire de Physiologie Cellulaire Végétale ont largement contribué à l'obtention d'une protéine recombinante purifiée. Par la suite, la proximité d'un laboratoire de l'EMBL (*European Molecular Biology Laboratory*) et du synchrotron nous a permis d'établir une collaboration avec l'équipe de C. Müller, conduisant à la résolution de la structure du complexe LFY-ADN.
Ce projet constitue avant tout le travail de Cécile Hamès qui a mis au point le protocole de production et purification de la protéine et obtenus les cristaux qui ont été utilisés pour l'analyse structurale. La protéine d'*Arabidopsis* était le seul homologue dont nous disposions à mon arrivée au laboratoire. De plus, cette protéine devait servir de référence pour la caractérisation ultérieure des homologues d'autres espèces, j'ai donc débuté mon travail de thèse en participant à cette étude. Ma contribution s'est organisée autour de deux aspects principaux : d'une part, j'ai travaillé à la mise au point du protocole de retard sur gel avec ADN fluorescent et j'ai réalisé plusieurs mutagenèses de la séquence de LFY-C afin d'obtenir des versions mutantes de la protéine recombinante. D'autre part, j'ai tenté de revisiter les données obtenues sous un angle évolutif. Dans un premier temps, j'ai cloné de nouveaux

homologues de *LFY* chez plusieurs espèces occupant des positions clés de la phylogénie. En parallèle, j'ai collecté au sein des bases de données publiques, les séquences de près de deux cents homologues de LFY parmi les 4 grands groupes de plantes terrestres. J'ai ensuite procédé à leur alignement afin d'étudier le devenir des acides aminés jouant un rôle central dans la reconnaissance de l'ADN et identifier d'éventuels homologues divergents.

Certains des résultats obtenus ont été inclus à l'étude publiée dans *The EMBO Journal*, d'autres non. Cer derniers sont présentés en compléments de l'article à la fin de ce chapitre.

The EMBO Journal (2008), 1–10 | © 2008 European Molecular Biology Organization | All Rights Reserved 0261-4189/08
www.embojournal.org

THE EMBO JOURNAL

Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins

Cécile Hamès[1,6], Denis Ptchelkine[2,3,6], Clemens Grimm[2,7], Emmanuel Thevenon[1], Edwige Moyroud[1], Francine Gérard[4], Jean-Louis Martiel[5], Reyes Benlloch[1,8], François Parcy[1,*] and Christoph W Müller[2,3,*]

[1]Laboratoire Physiologie Cellulaire Végétale, UMR5168, Centre National de la Recherche Scientifique, Commissariat à l'énergie atomique, Institut National de la Recherche Agronomique, Université Joseph Fourier, Grenoble, France, [2]European Molecular Biology Laboratory, Grenoble, France, [3]European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany, [4]Unit of Virus Host Cell Interactions, UMR5233 Université Joseph Fourier—European Molecular Biology Laboratory—Centre National de la Recherche Scientifique, Grenoble, France and [5]TIMC-IMAG Laboratory, Université Joseph Fourier, CNRS UMR5525, INSERM, Grenoble, France

The LEAFY (LFY) protein is a key regulator of flower development in angiosperms. Its gradually increased expression governs the sharp floral transition, and LFY subsequently controls the patterning of flower meristems by inducing the expression of floral homeotic genes. Despite a wealth of genetic data, how LFY functions at the molecular level is poorly understood. Here, we report crystal structures for the DNA-binding domain of *Arabidopsis thaliana* LFY bound to two target promoter elements. LFY adopts a novel seven-helix fold that binds DNA as a cooperative dimer, forming base-specific contacts in both the major and minor grooves. Cooperativity is mediated by two basic residues and plausibly accounts for LFY's effectiveness in triggering sharp developmental transitions. Our structure reveals an unexpected similarity between LFY and helix-turn-helix proteins, including homeodomain proteins known to regulate morphogenesis in higher eukaryotes. The appearance of flowering plants has been linked to the molecular evolution of LFY. Our study provides a unique framework to elucidate the molecular mechanisms underlying floral development and the evolutionary history of flowering plants.

The EMBO Journal advance online publication, 11 September 2008; doi:10.1038/emboj.2008.184
Subject Categories: plant biology; structural biology
Keywords: crystal structure; flower development; homeotic genes; LEAFY; transcriptional regulation

*Corresponding authors. F Parcy, Laboratoire Physiologie Cellulaire Végétale, CNRS, UMR5168, CEA, 17 av. des Martyrs, bât. C2, 38054 Grenoble, France. Tel.: +33 438 784 978; Fax: +33 438 784 091; E-mail: francois.parcy@cea.fr or CW Müller, Structural and Computational Biology Unit, EMBL Meyerhofstrasse 1, 69012 Heidelberg, Germany. Tel.: +49 6221 387 8320; Fax: +49 6221 387 519; E-mail: christoph.mueller@embl.de
[6]These authors contributed equally to this work
[7]Present address: Institut für Biochemie, Biozentrum der Universität, Am Hubland, 97074 Würzburg, Germany
[8]Present address: Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Sciences, 90183 Umeå, Sweden

Received: 11 April 2008; accepted: 22 August 2008

Introduction

Homeotic genes control developmental patterns and organ morphogenesis. In animals, they encode transcription factors of the homeodomain family, such as Hox and paired proteins, which contact DNA through one or several helix-turn-helix (HTH) motifs (Gehring *et al*, 1994; Underhill, 2000). In plants, most homeotic genes determining the identity of floral organs encode MADS-box transcription factors, suggesting that plants and animals have adopted distinct types of homeotic regulators (Meyerowitz, 1997; Ng and Yanofsky, 2001). In addition to organ identity genes, plants also use another class of regulators named 'meristem identity genes', which control floral meristem versus shoot/inflorescence fate. In *Arabidopsis thaliana*, the meristem identity genes *LEAFY* (*LFY*) and *APETALA1* (*AP1*) induce flower development, whereas *TERMINAL FLOWER1* (*TFL1*) promotes inflorescence development (Blazquez *et al*, 2006). Mutations or ectopic expression of these genes result in complete or partial interconversions between flower and inflorescence meristems.

The *LFY* gene encodes a plant-specific transcription factor, which has a cardinal function in this process, regulating both the transition to flowering and the subsequent patterning of young floral meristems. During the plant vegetative growth, *LFY* expression increases in newly formed leaves until a certain threshold is reached. LFY then induces the expression of *AP1* and *CAULIFLOWER* (*CAL*) genes and triggers the abrupt floral transition (Blazquez *et al*, 2006). Once the floral meristem is established, LFY governs its spatial patterning by inducing the expression of the floral homeotic ABC genes, such as *AP1*, *AP3* or *AGAMOUS* (*AG*), which control the identity of stereotypically arranged floral organs (Coen and Meyerowitz, 1991; Lohmann and Weigel, 2002).

LFY is found in all terrestrial plants from moss to angiosperms; its sequence shows a high level of conservation throughout the plant kingdom but no apparent similarity to other proteins (Maizel *et al*, 2005). Unlike many plant transcription factors that evolved by gene duplication to form a multigene family (Riechmann and Ratcliffe, 2000; Shiu *et al*, 2005), *LFY* is present in single copy in most angiosperms and *lfy* mutants available from several species such as snapdragon, petunia, tomato or maize show, as in *Arabidopsis*, partial or complete flower-to-shoot conversions (Coen *et al*, 1990; Souer *et al*, 1998; Molinero-Rosales *et al*, 1999; Bomblies *et al*, 2003). In gymnosperms, a paralogous *NEEDLY* (*NLY*) clade of genes exists. No mutant is available in these species, but *LFY* and *NLY* expression patterns are also consistent with a role in reproductive organ development

(reviewed in Frohlich and Chase, 2007). Because of its central role in determining floral meristem identity, and considering that *NLY* disappeared concomitantly with the appearance of flowers, *LFY* has been put at the centre of different evolutionary scenarios that rationalize the appearance of the successful angiosperm group (Albert *et al*, 2002; Frohlich, 2003; Frohlich and Chase, 2007; Theissen and Melzer, 2007).

LFY activates gene expression by recognizing pseudo-palindromic sequence elements (CCANTGT/G) in the promoters of its target genes, including *AP1* (one site) and *AG* (four sites; *AG-I* to *AG-IV*) (Parcy *et al*, 1998; Busch *et al*, 1999; Lohmann *et al*, 2001; Lamb *et al*, 2002; Hong *et al*, 2003). LFY has two domains, a partially conserved N-terminal domain that is thought to contribute to transcriptional activation and a highly conserved C-terminal domain responsible for DNA binding (LFY-C) (Coen *et al*, 1990; Maizel *et al*, 2005). LFY functions synergistically with coregulators such as the WUSCHEL (WUS) homeodomain protein (Lenhard *et al*, 2001; Lohmann *et al*, 2001) or the UFO F-Box protein (Lee *et al*, 1997; Parcy *et al*, 1998; Chae *et al*, 2008).

In this study, we show that LFY binds DNA cooperatively as a dimer, a property shown to be essential to trigger developmental switches. The crystal structure of LFY-C bound to DNA reveals the molecular basis for sequence-specific recognition and cooperative binding as well as an unexpected similarity of LFY with HTH proteins such as homeodomain transcription factors. Our findings enable to formulate new hypotheses on the appearance of angiosperms in evolution.

Results and discussion

LFY-C dimerizes on DNA binding

We produced the recombinant LFY DNA-binding domain (LFY-C, residues 223–424) and showed by size-exclusion chromatography (SEC) that it is monomeric in the absence of DNA (Figure 1A and B). In electrophoretic mobility shift assays (EMSAs), LFY-C recognized a DNA probe bearing an *AP1* site as two distinct species: a major protein–DNA complex and a minor one of higher mobility (Figure 1C). Multi-angle laser light scattering (MALLS) coupled to SEC demonstrated that the major complex contained two LFY-C molecules per DNA duplex (Figure 1B). The homodimeric nature of LFY in this complex was confirmed by mixing untagged and GFP-tagged LFY-C and observing a single new species attributable to the formation of an LFY-C/GFP–LFY-C/DNA complex (Figure 1C). Using probes mutated in one half-site of the palindrome, we confirmed that the minor, high-mobility species corresponds to a single LFY-C monomer bound to DNA (Supplementary Figure 2).

Structure of the LFY DNA-binding domain bound to its DNA recognition site

To understand how LFY specifically recognizes its DNA target sequences, we crystallized LFY-C in complex with DNA. We solved the structure of LFY-C bound to two different LFY-binding sites, *AP1* and *AG-I* at 2.1- and 2.3-Å resolution, respectively (Figures 2, and 3A and B; Table I). The overall structure shows an LFY-C dimer bound to a pseudo-palindromic DNA duplex, where the LFY-C monomers are related by a crystallographic dyad. The DNA duplexes used for co-crystallization deviate from strict two-fold symmetry at the 5′ends and at base pairs (bp) ±9, ±7 and ±0 in the *AP1* site

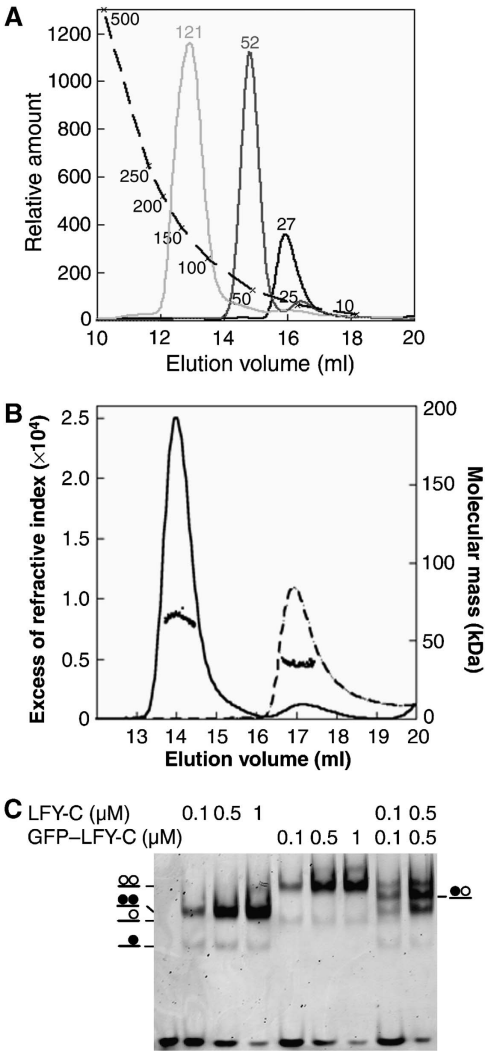


Figure 1 DNA-dependent dimerization of LFY-C. (**A**) Size-exclusion chromatography. LFY-C (40 μM, black curve), *AP1* DNA (10 μM, dark grey curve), LFY-C (40 μM) + *AP1* DNA (10 μM, light grey curve) were analysed. LFY-C elution at a volume corresponding to 28 kDa is consistent with the monomer size (25.7 kDa), the DNA duplex elutes earlier than expected at a volume corresponding to 52 kDa because of its elongated shape. The LFY/DNA complex elutes at a volume corresponding to 121 kDa. Molecular weights estimated from the calibration curve (dashed line) are indicated. (**B**) Molecular mass of LFY-C alone (dashed line) or in combination with *AP1* DNA (solid line) determined by multi-angle laser light scattering and refractometry combined with size-exclusion chromatography. Elution profiles were monitored by excess refractive index (left ordinate axis). Dots show the molecular mass distribution (right ordinate axis). Average molecular mass is 64 ± 2 kDa for the LFY-C/DNA complex (65 kDa theoretical size for a dimeric complex) and 35 ± 1 kDa for LFY-C alone (26 kDa theoretical size for LFY-C monomer). (**C**) Electrophoretic mobility shift assay (EMSA) with 10 nM *AP1* DNA and various LFY-C or GFP–LFY-C concentrations. Schematic complexes with LFY-C (filled circle) and GFP–LFY-C (open circle) are depicted.

(5′ end, bp ±7, ±6, ±4 and ±0 in the *AG-I* site). Nevertheless, the pseudo-dyads of the DNA duplexes coincide with the crystallographic dyad, probably as a result of the random bimodal orientation of the DNA duplex around the dyad (see Materials and methods). The resulting molecular averaging does not impair our interpretation of the protein–DNA interface. In the final $2F_o-F_c$ elec-

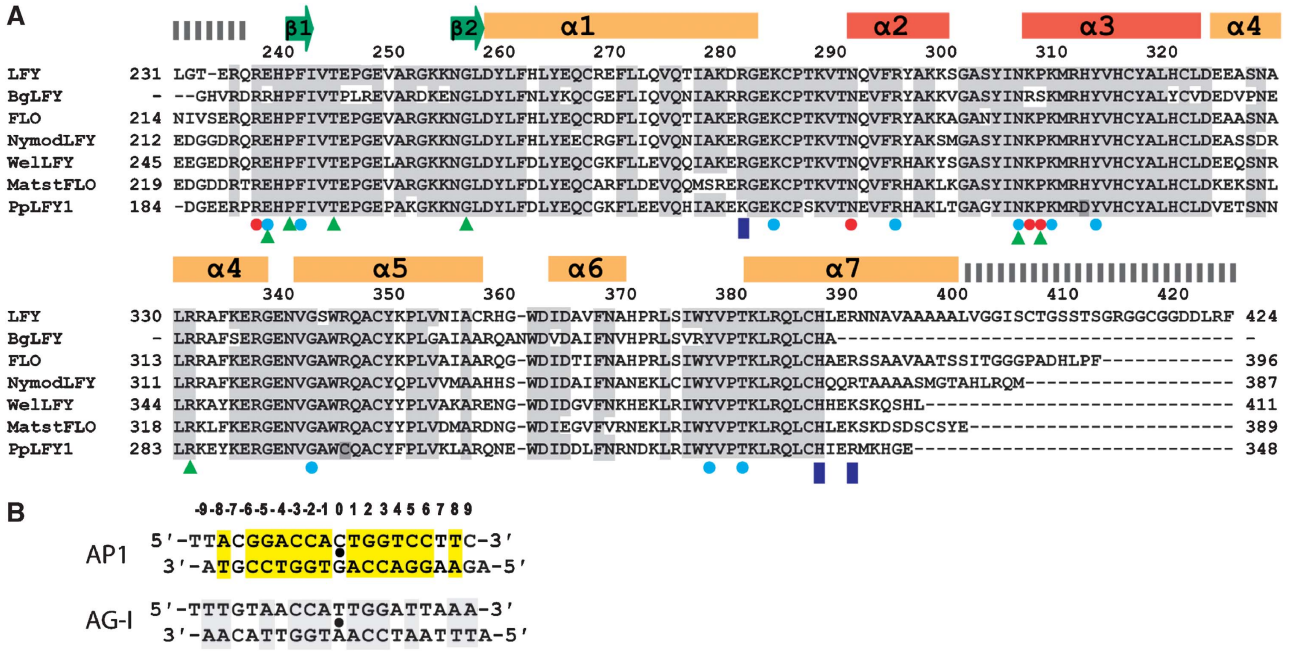


Figure 2 Sequence alignments. (**A**) Aligned C-terminal amino-acid sequences of LFY (*Arabidopsis thaliana*, AAA32826), BgLFY (*Brownea grandiceps*, AAS79888), FLO (*Antirrhinum majus*, P23915), NymodLFY (*Nymphea odorata*, AAF77609), WelLFY (*Welwitschia mirabilis*, AAF23870), MatstLFY (*Matteuccia struthiopteris*, AAF77608) and PpLFY1 (*Physcomitrella patens*, BAD91043). Identical and conservatively substituted residues are depicted on a grey background. Secondary structure elements are indicated. Residues involved in interactions with DNA bases and backbone are labelled with red and blue circles, respectively. Dashed bars indicate disordered regions in the crystal, blue rectangles indicate the residues involved in dimerization. Green triangles indicate the position of *Arabidopsis* mutations and residues divergent in PpLFY1 are highlighted in pink. (**B**) Two DNA duplexes containing the LEAFY-binding sites from *AP1* and *AG* promoters present in the LEAFY–DNA complex crystals are depicted. Base pairs related by a dyad (indicated by a black dot) are highlighted in yellow.

tron density map, but also in the initial solvent-flattened single isomorphous replacement with anomalous scattering (SIRAS) electron density map (Supplementary Figure 1), the sugar–phosphate backbone of the DNA is well defined and shows no evidence of conformational averaging. The density for palindromic DNA bases is clearly defined, whereas the density at non-palindromic positions is consistent with the superposition of two different base pairs. Furthermore, all residues close to the DNA are clearly defined and we do not observe any diffuse density, which suggests that each monomer undergoes only minor changes to adapt to the slightly different half-sites. Despite the differences between the *AP1* and *AG-I* binding sites (Figure 2B), both complex structures are very similar and can be superimposed with an r.m.s. distance of 0.55 Å for 163 C_α and 19 phosphate atoms.

LFY-C (with residues 237–399 ordered in the crystal structure) adopts a compact fold that interacts principally with a single DNA half-site (Figures 3A and B, and 4A and B). The fold is defined by two short β-strands followed by seven helices connected by short loops (Figure 3A and B). The absence of any extended hydrophobic patches at its surface suggests that LFY-C represents an autonomous DNA-binding domain without a large interface to its N-terminal domain. Helices α2 and α3 define a HTH motif (Aravind *et al*, 2005), with helix α3 occupying the major groove and mediating most of the DNA contacts. The DNA in the complex adopts a B-DNA-like conformation exhibiting an overall bend of about 20° (Figure 3C), which can be localized to two kinks of about 10° at base pairs ±2/±3. Both ends of the DNA duplex are AT rich and the minor grooves are narrower compared with classical B-DNA. Narrowing of the minor groove is slightly more pronounced in the *AG-I* duplex than in *AP1*.

DNA recognition in the major and minor grooves

Sequence-specific contacts between LFY and the DNA involve both the minor and major grooves. Base-specific contacts in the major groove are formed by Asn291 and Lys307 in helices α2 and α3, which together specify the two invariant guanines at positions ±2 and ±3 (Figure 4A and B). Mutating either of these residues into alanine resulted in considerably lower DNA-binding affinity (Figure 4C), whereas previous studies showed loss of binding when the corresponding base pairs were mutated (Parcy *et al*, 1998; Busch *et al*, 1999). The *Arabidopsis lfy-20* mutation (N306D) adjacent to Lys307 also leads to a reduced DNA-binding affinity (Supplementary Figure 3) and a weak *lfy* phenotype *in planta* (Weigel *et al*, 1992; Maizel *et al*, 2005), presumably because the negatively charged aspartate interacts unfavourably with the DNA backbone (Figure 4A and B).

Base-specific recognition in the minor groove is mediated by Arg237, which is the first ordered N-terminal residue in the crystal structure. At the *AP1* site, its side chain points towards A:T base pairs ±8 and contacts the exocyclic O2 of thymine 8 and also the O2 of cytosine 7 in one half-site, or the O2 of cytosine-9 in the other half-site (Figure 4A and B; Supplementary Figure 4). In the *AG-I* site, T:A base pair 8 is replaced by A:T, and in the LFY/*AG-I* complex, the Arg237 side chain adopts a different conformation, which allows it to recognize the thymine of the opposite strand (Supplementary Figure 4). The importance of this interaction is underscored by

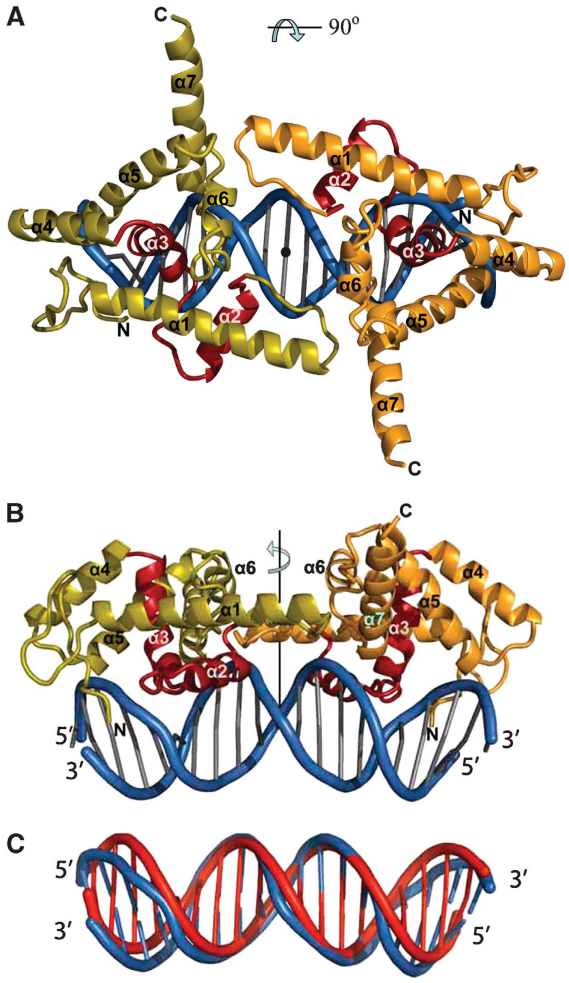


Figure 3 Structure of the LFY-C dimer bound to DNA. (**A**, **B**) Two orthogonal views of the LFY-C dimer (residues 237–399) bound to DNA. Monomers are coloured in olive and orange with the helix-turn-helix (HTH, helices α2 and α3) motif in red. The DNA duplex is depicted in blue. Figures 3, 4B, 5A and 6 were produced with program Pymol (Delano, 2002). (**C**) Superposition of the DNA duplex found in the LEAFY–DNA complex (blue) with regular B-form DNA (red).

the presence of A:T or T:A base pairs at position 8 in all 12 confirmed half-sites (Parcy *et al*, 1998; Busch *et al*, 1999; Lohmann *et al*, 2001; Lamb *et al*, 2002; Hong *et al*, 2003). The consensus LFY-binding site is therefore more accurately defined as T/ANNNNCCANTGT/GNNNNT/A (with the centre of the pseudo-palindrome underlined). The Arg237 side chain is inserted into an AT-rich narrow minor groove (Figure 4A and B), similar to that observed in the Hox homeodomain–Exd–DNA complex, where the narrow minor grove was shown to enhance the electrostatic interaction between DNA backbone and arginine side chain (Joshi *et al*, 2007). The R237A mutation led to a strongly reduced affinity of LFY-C for *AP1* (Figure 4C). In contrast, changing the adenine 8 into a cytosine in *AP1* reduced only moderately the LFY-C-binding affinity (Figure 4C; *AP1* m5), presumably because the arginine side chain can contact the adjacent base (Supplementary Figure 4). Finally, next to Arg237, the two *lfy* mutations (*lfy-4* (E238K) and *lfy-5* (P240L)) result in decreased *in vitro* binding affinities (Supplementary Figure 3) and lead to a mutant phenotype *in planta* (Weigel *et al*, 1992).

An unusual contact with DNA is mediated by Pro308 that points between the guanines in base pairs ±5 and ±6, which results in a pronounced propeller twist for base pair ±5 and local bending of the DNA at this position (Figure 4B). The mutant *lfy-28* (P308L) is impaired in DNA binding and gives rise to an intermediate to strong phenotype *in planta* (Figure 4C–F), as a likely consequence of a steric clash of the leucine side chain with the guanine bases. In contrast, a small side chain such as alanine perfectly fits in this protein–DNA interface and, indeed, the mutant protein P308A showed a wild-type-binding affinity (Figure 4C). Pro308 is not strictly conserved and is substituted by serine in some *Brownea* species (Figure 2A). This substitution probably modifies DNA binding, because serine can form direct hydrogen bonds to DNA bases at positions ±4 and +5 and it replaces P308, which locally distorts DNA. However, the conformational flexibility of serine and its ability to function as hydrogen bond donor or acceptor makes it difficult to predict the preferred binding specificity. Moreover, P308S is systematically associated with the K307R substitution, affecting the base-contacting residue Lys307 (Figure 2A). LFY proteins from *Brownea* species might therefore recognize significantly different DNA target sites.

Similar to most protein–DNA co-crystal structures, not all bases in the consensus LFY site T/ANNNNCCANTGT/GNNNNT/A are specified through direct interactions with the protein. Additional specificity presumably arises from sequence-dependent deformability of the DNA, sometimes referred to as 'indirect readout'. Dinucleotide steps CA/TG at bp ± 1/bp ± 2 are part of the consensus LFY site and are particularly flexible, which might facilitate the observed kink of the DNA at base pairs ±2/±3. However, these particular sequences are not critically required as they are not conserved in the *AP3-I* binding site (Lamb *et al*, 2002).

Not all LFY mutations directly affect DNA contacts. Mutations *lfy-3* (T244M) and *lfy-9* (R331K) (Weigel *et al*, 1992) disturb two interacting amino acids, both of which contribute to a polar network that connects N-terminal residues with helices α1 and α4. Similarly, two other residues (His312 and Arg345) interact in a typical planar stacking. His312 and Arg345 are conserved except in the LFY protein from *Physcomitrella patens* (PpLFY1) where they are substituted by aspartate and cysteine, respectively (Maizel *et al*, 2005). His312 forms part of helix α3 and is located just one helical turn above Pro308 at the N-terminal end of helix α3. In addition, the preceding residue Lys307 directly contacts the guanine in base pairs ±2 (Figure 4A). The loss of the His312/Arg345 stacking interaction in the moss PpLFY1 likely affects the orientation of helix α3, explaining the altered DNA-binding properties of this orthologue, whereas reverting the aspartate into histidine restores the binding activity of PpLFY1 to canonical LFY-binding sites (Maizel *et al*, 2005).

Structural basis for cooperative DNA binding

The structure of the LFY-C/DNA complex also reveals important monomer–monomer interactions governing its DNA-binding mode (Figure 5A). Our EMSA analysis shows that LFY-C binds DNA in a cooperative manner: the monomeric complex is present only in minor amounts as compared with the dimeric complex, even at low LFY-C concentrations (Figure 5B) and binding of the second monomer occurs with a 90-fold higher affinity than binding of the first one

Table I Structure determination of the LEAFY–DNA complex

Data statistics

Data set	Resolution (Å)[a]	Reflections measured/unique	R_{meas} (%)[b]	I/σ	Completeness (%)
AP1/LFY: space group $P6_522$, unit cell dimensions a = b = 98.8 Å, c = 177.4 Å					
Native	20–2.1 (2.2–2.1)	503 443/29 859	4.8 (54.3)	38.5 (5.0)	97.4 (83.9)
EMTS	20–2.4 (2.5–2.4)	283 416/20 463	11.1 (50.2)	15.5 (4.7)	99.6 (99.3)
AG-I/LFY: space group $P6_522$, unit cell dimensions a = b = 98.4 Å, c = 176.4 Å					
Native	20–2.3 (2.4–2.3)	328 932/23 456	7.9 (75.1)	26.7 (3.4)	99.8 (100.0)

Phasing statistics for the EMTS derivative (SIRAS)

Wavelength (Å)	0.934
Phasing power[c]	1.68
Figure of merit	0.468
R_{cullis}[d]	0.674
Number of mercury sites	4

Refinement statistics

	LFY/AP1	LFY/AG-I
Resolution[a]	19.1–2.1 (2.2–2.1)	20.0–2.3 (2.4–2.3)
Total number of non-hydrogen protein atoms	1352	1332
Total number of non-hydrogen DNA atoms	526	567
Number of water molecules	148	105
R-factor (%)[a]	21.0 (25.0) for 26 896 reflections	22.1 (29.0) for 22 559 reflections
R_{free} (%)[a,e]	23.7 (26.4) for 1521 reflections	24.9 (30.7) for 1183 reflections
r.m.s. deviations		
Bond lengths (Å)	0.009	0.008
Bond angles (deg)	1.22	1.21
Average temperature factors (Å^2)		
Protein	41.1	39.8
DNA	48.1	46.4
Solvent	42.8	48.5
r.m.s.d. of covalently linked atoms (Å^2)	2.55	3.45
Residues in Ramachandran plot[f]		
Most favoured regions (%)	93.2 (136)	95.2 (139)
Additionally allowed regions (%)	6.8 (10)	4.8 (7)
Generously allowed regions (%)	0 (0)	0 (0)
Disallowed region (%)	0 (0)	0 (0)

[a]Values for the highest resolution range are given in parenthesis.
[b]R_{meas} is a redundancy independent R-factor as defined in Diederichs and Karplus (1997).
[c]Phasing power is the mean value of the heavy-atom structure factor amplitudes divided by the mean lack of closure.
[d]R_{cullis} is the mean lack of closure divided by the mean isomorphous difference.
[e]R_{free} was calculated from a subset of 5% of the data.
[f]Number of residues are given in parentheses.

(Figure 5C; Supplementary Figure 5). This type of cooperative binding can result either from DNA conformability, where binding of one monomer favours the binding of the second monomer, or from protein–protein interactions between DNA-bound monomers (Senear *et al*, 1998; Schumacher *et al*, 2002; Panne *et al*, 2004). Our structure suggests the latter. The LFY dimer comprises a small interface of 420 Å^2 buried surface area formed by loop α12 and helix α7 in which the two residues His387 and Arg390 form hydrogen bonds with the backbone carbonyl of Asp280 (Figure 5A). We validated the importance of these contacts by mutagenesis: cooperativity of binding is moderately affected in the H387A or R390A single mutants but more strongly reduced in a H387A/R390A double mutant (Figure 5B and C; Supplementary Figure 5). Therefore, the small monomer–monomer interface (with a major contribution of His387 and Arg390) rather than DNA conformability is responsible for the cooperative binding. Whether the N-terminal domain of LFY also participates in dimerization, in the presence or absence of DNA, will require additional experiments.

A better understanding of LFY's DNA-binding mode also provides insight into its molecular switch function. DNA-binding cooperativity, as well as dimerization, allows transcription factors to work at lower concentrations and to enhance the sigmoidality of their response curves. When combined with feedback loops, it has been shown essential for threshold-dependent genetic switches (Burz *et al*, 1998; Cherry and Adler, 2000). LFY is involved in a positive autoregulation loop through activation of the homologous *AP1* and *CAL* genes, that in turn activate *LFY* expression (Bowman *et al*, 1993; Liljegren *et al*, 1999). LFY-binding cooperativity combined with the AP1/CAL feedback loop

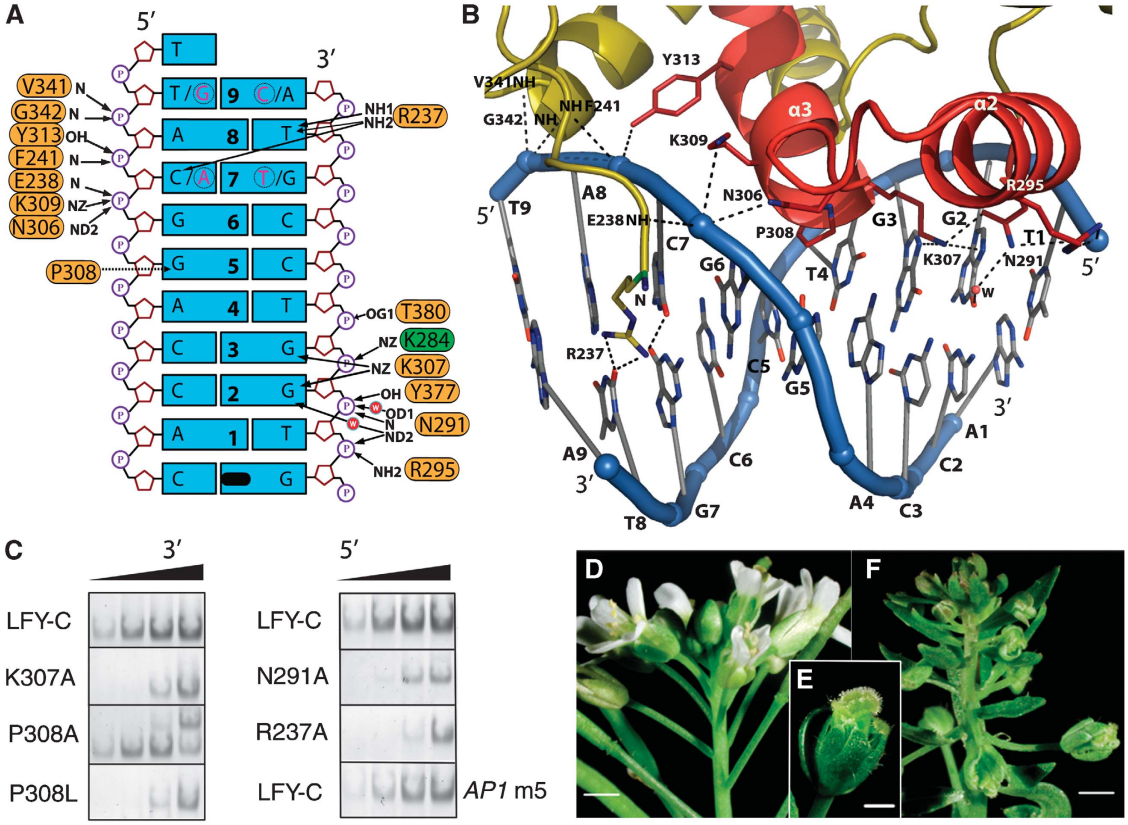


Figure 4 DNA recognition by LEAFY. (**A**) Protein–DNA interactions in one *AP1* half-site. Dyad-related base pairs 7 and 9 from the other half-site are shown in pink and encircled. Polar and hydrophobic interactions are shown with solid and dashed arrows, respectively. K284 belongs to the other monomer and is depicted in green. The pseudo-dyad coinciding with the crystallographic dyad is depicted in black. (**B**) Ribbon diagram of one LEAFY monomer bound to its *AP1* half-site. The protein is coloured in olive except for the HTH motif shown in red. Polar interactions are indicated by dashed lines. For clarity, only side chains in contact with DNA are shown. (**C**) Effect of selected mutations on LFY-C DNA-binding affinity to *AP1* DNA. EMSAs were performed with wild-type and mutant LFY-C (100–250–750–2000 nM from left to right). Only dimeric complexes are shown except for P308A that gave rise to an unknown higher complex. *AP1* m5 mutant DNA contains base pair C:G instead of A:T at position ± 8 (see Supplementary Table 3 for full DNA sequences). Phenotype of the wild-type *Arabidopsis* inflorescence (**D**) and *lfy-28* (P308L) mutant inflorescence (**F**) and flower (**E**). Scale bar is 1 mm on (D, F) and 0.5 mm on (E).

therefore provides a plausible explanation for the threshold-dependent floral switch triggered by LFY.

Many transcription factors bind DNA as homodimers but also form heterodimers, thereby extending their spectrum of recognized DNA target sequences (Klemm *et al*, 1998; Garvie and Wolberger, 2001). LFY has been shown to activate the *AG* organ identity gene synergistically with the homeodomain protein WUS (Lohmann *et al*, 2001). As adjacent WUS- and LFY-binding sites are present on the *AG* regulatory sequence, it has been suggested that LFY and WUS could bind simultaneously (Lohmann *et al*, 2001; Hong *et al*, 2003). Preliminary model building indicates that LFY homodimers cannot be accommodated with WUS at adjacent LFY- and WUS-binding sites. This observation raises the intriguing possibility that they might either compete for the same binding sites or more likely could form LFY–WUS heterodimers.

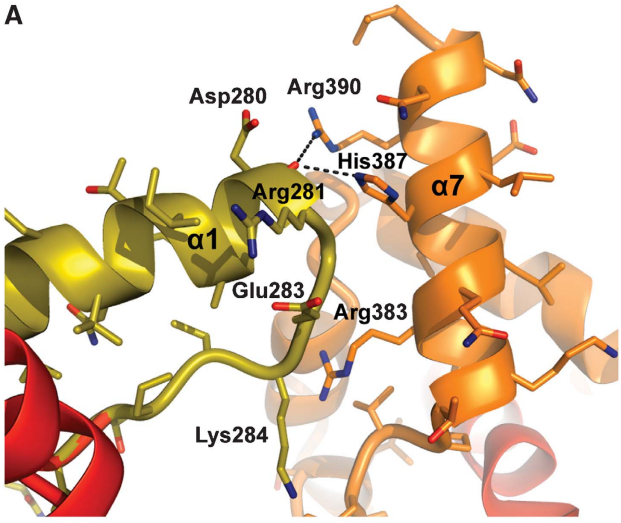
LFY shows similarities with HTH proteins

The nature and origin of LFY had so far remained elusive: LFY-C's primary sequence shows unusually strong sequence conservation within its family but has no apparent similarity to any described transcription factor. The crystal structure of LFY-C bound to DNA reveals a seven-helix domain with many residues involved in protein–DNA interactions, tightly constrained packing interactions in the hydrophobic core and protein–protein interactions with the other monomer. Presumably, these observed tight structural and functional constraints on many residues spread over the entire DNA-binding domain explain the high level of sequence conservation within LFY-C.

The LFY-C structure contains an unpredicted HTH motif formed by helices α2 and α3 as part of the overall fold. HTH motifs are present in a wide variety of DNA-binding proteins throughout the three kingdoms of life. They are typically found in a bundle of 3–6 α-helices or combined with β-sheets (winged HTH/fork head domain), which provide a stabilizing hydrophobic core (Weigel and Jackle, 1990; Aravind *et al*, 2005). Comparison of LFY-C against the Protein Data Bank using program DALI (Holm and Sander, 1993) detects similarity of relative short α-helical segments (~60 amino-acid residues) with different α-helical proteins including HTH proteins (maximal Dali *Z*-score 3.0, pairs with $Z<2.0$ are structurally dissimilar). A search comprising only the first three N-terminal helices, including the HTH motif, mainly showed similarity to different HTH proteins with slightly higher scores (maximal Dali *Z*-score: 4.5). When considering just the three helices α1, α2 and α3, LFY aligns well with other three-helix bundle HTH proteins, including the homeodo-

main protein engrailed (r.m.s.d.$_{40C\alpha}$ = 2.9 Å), the paired domain (r.m.s.d.$_{44C\alpha}$ = 3.5 Å) and the Tc3 transposase (r.m.s.d.$_{30C\alpha}$ = 2.4 Å). LFY and partitioning protein KorB (r.m.s.d.$_{71C\alpha}$ = 3.7 Å) share some similarity beyond the typical DNA/RNA-binding three-helical bundle core (Russell and Barton, 1992; Khare *et al*, 2004), where five of the seven LFY-C helices, including the HTH motif, roughly superimpose with KorB helices. However, LFY cannot be easily assigned to any of the described classes of HTH proteins (Aravind *et al*, 2005) and it therefore represents a new variant of multi-helical bundle proteins.

The DNA recognition mode of LFY is similar to those observed for the paired domain, Tc3A transposase, Hin recombinase and λ repressor (van Pouderoyen *et al*, 1997; Xu *et al*, 1999). The axis of the recognition helix α3 in the HTH of these proteins is oriented parallel to the edges of the nucleotide bases. Only the N terminus of the recognition helix is inserted into the major groove of the DNA, whereas the short helix α2 has a supporting function. In contrast, in homeodomain proteins, the long probe helix α3 runs more parallel to the neighbouring DNA phosphate backbone, and mainly the central part of helix α3 contacts the DNA



	WT LFY-C	R390A	H387A	H387A/R390A
K_{d1}	8603±1365 nM	8603±1365 nM	8603±1365 nM	8603±1365 nM
K_{d2}	95±37 nM	145±50 nM	241±78 nM	1060±464 nM
K_{d1}/K_{d2}	90 [65–145]	59 [44–91]	36 [27–54]	8 [6–14]

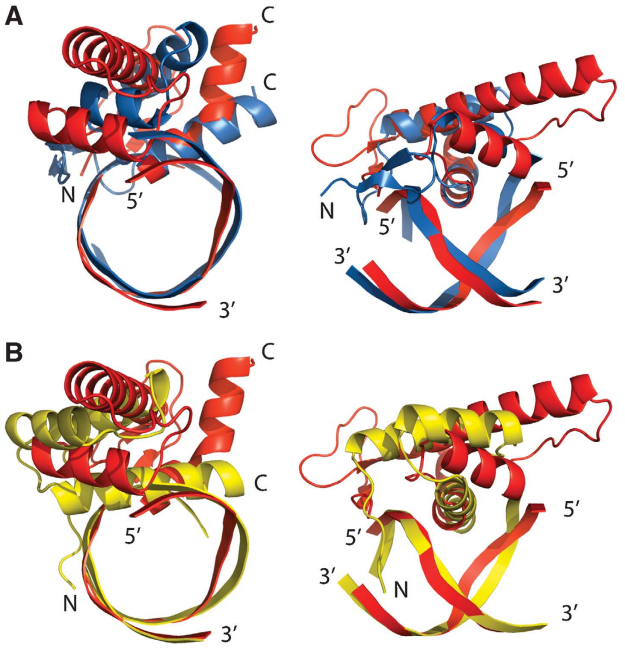


Figure 6 Comparison of LFY-C with paired and homeodomain DNA binding. (**A**) Two orthogonal views of LFY-C helices α1–α3 bound to their DNA target site (red) superimposed with the three-helical bundle core of the N-terminal subdomain of the paired domain of *Drosophila* Prd (blue, PDB-id: 1pdn). (**B**) Superposition with the homeodomain of *Drosophila* engrailed bound to DNA (yellow, PDB-id: 1hdd), where the centre of recognition helix α3 inserts into the major groove.

Figure 5 The LFY-C dimer interface mediates cooperative binding. (**A**) The dimer interface is viewed perpendicular to the DNA axis. Polar contacts between the two monomers (in orange and olive) are shown with dashed lines. (**B**) EMSA with increasing concentrations (0, 10, 20, 50, 100, 200, 500, 1000, 2000 and 3000 nM from left to right) of LFY-C wild-type, R390A mutant, H387A mutant, and H387A/R390A double mutant and 50 μM *AP1* DNA. Free DNA (F), monomeric (M) and dimeric (D) complexes are indicated. (**C**) Estimation of dissociation constants for wild-type LFY-C and three mutant versions (H387A, R390A and H387A/R390A). Binding of LFY-C to *AP1* DNA was modelled as two equilibrium reactions as detailed in Supplementary data: (1) Binding of a first LFY-C monomer to *AP1* DNA, leading to the formation of the monomeric complex (M) and characterized by the K_{d1} dissociation constant; (2) binding of a second LFY-C monomer to M, leading to the formation of the dimeric complex (D) and characterized by K_{d2}. EMSA signals from (B) were quantified and the corresponding experimental values were fitted with theoretical equations describing the two equilibria. The errors and intervals between square brackets indicated correspond to the 95% confidence interval. An elevated K_{d1}/K_{d2} ratio reflects a high level of cooperativity, whereas a ratio of 1 would indicate an absence of binding cooperativity. The single mutations resulted in a weak decrease of cooperativity, whereas the H387A/R390A double mutation strongly decreased the cooperativity.

(Figure 6). Similarity between LFY and the paired domain also includes a small two-stranded β-sheet, which precedes the three-helix bundle, and N-terminal residues, which are inserted into the minor groove. However, the minor groove contacting residues are located at the most N-terminal end of LFY-C, whereas in the paired domain they protrude from the loop connecting the two short N-terminal β strands.

Sequence similarities are too weak to suggest a precise evolutionary origin for LFY, although structural resemblances indicate that it might derive from ancestral HTH proteins, including paired and homeodomain proteins (Rosinski and Atchley, 1999; Breitling and Gerber, 2000; Aravind *et al*, 2005). Until now, most plant homeotic genes were found to encode MADS box transcription factors, whereas plant homeodomain proteins rather control meristem homoeostasis and cell division (Meyerowitz, 1997; Ng and Yanofsky, 2001). Our study reveals that the LFY master regulator, which determines flower meristem fate and controls the expression of floral organ identity genes, shares structural similarity with other HTH proteins, indicating that this universal DNA-binding motif has also been adopted in plants to trigger major developmental switches.

Prospects regarding the appearance of angiosperms

The LFY-C structure combined with more than 200 LFY sequences from all types of terrestrial plants offers a unique opportunity to detect key residues in evolution. Some charged LFY-C surface residues (such as Lys253 or Lys254) are strictly conserved, suggesting that they might participate in interactions with other proteins. Other residues are conserved in all angiosperms but not in the non-flowering plants. For example, R390, identified as one of the residues mediating interaction between monomers and cooperative binding, has been conserved in angiosperm LFY proteins, whereas most LFY from non-flowering plants, such as gymnosperms and ferns, show a lysine at this position. This amino-acid change presumably weakens the interaction between monomers and thereby reduces the DNA-binding affinity. The acquisition of R390 might therefore have been important for flower evolution. Because LFY stands at the very centre of the network regulating flower development, it has been proposed that modifications of the *LFY* gene contributed to the appearance of floral structures in evolution (Albert *et al*, 2002; Frohlich, 2003; Frohlich and Chase, 2007; Theissen and Melzer, 2007). The availability of the LFY-C crystal structure provides a unique framework for generating plausible hypotheses that relate the appearance of angiosperms to specific events during the molecular evolution of LFY. The 'functional synthesis' approach that combines phylogeny, biochemical and structural analyses with functional assays *in vivo* (Dean and Thornton, 2007) can now be applied to LFY to try to solve one of the most puzzling enigmas of plant biology: the origin of flowers.

Materials and methods

Plant material

The *lfy-28* mutant allele of *A. thaliana* (accession Landsberg *erecta*) was kindly provided by D Weigel (Max Planck Institute, Tübingen, Germany) and originally isolated by J Fletcher (PGEC, Albany). *lfy-28* mutant had been back-crossed twice with the wild type, and individuals showing a mutant phenotype were selected from segregating populations. Plants were grown at 25°C in long days (16 h light).

Plasmid constructions

Expression plasmids. LFY-C (residues 223–424 from *A. thaliana LFY* cDNA) was amplified from pIL-8 (obtained from D Weigel) with Pfu Turbo Polymerase (Stratagene, France) and primers oFP1242 (5′CTCTCGAGCCCGGGCTAGAAACGCAAGTCGTCGCC3′) and oFP1244 (5′CTCTCGAGCCCGGGCTATCCGGTACAGCTAATACCGCC3′), subcloned into pCR-TOPO-BluntII (Invitrogen, Cergy Pontoise, France) and shuttled to pETM-11 (Dummler *et al*, 2005) as *Nco*I/*Xho*I fragment to yield the pCH28 expression vector. pETM-11 contains an N-terminal 6× His tag followed by a tobacco etch virus (TEV) cleavage site.

LFY–GFP plasmid. A GFP fragment was amplified from pBS-GLFY plasmid obtained from X Wu (Wu *et al*, 2003) using primers oETH1001 5′CCCACTACTGAGAATCTTTATTTTCAGGGCCAGTTCAG TAAAGGAGAAGAAC3′ and oETH1002 5′CCCCAAACCACTACCTCCG TTGCCGTTATCCTGTTTGTATAGTTCATCCAT3′. The amplified fragment was subsequently used as a megaprimer to amplify plasmid pCH28 and yield pETH8 (6His–TEV–GFP–LFY-C).

Expression plasmids for mutant LFY-C. pCH45 (K307A), pCH46 (N291A), pCH47 (R237A), pCH48 (P308A), pCH49 (D280K), pCH50 (H387A/R390A), pCH54 (H387A), pEDW127 (R390A), pCH55 (*lfy-28*, P308L), pETH21 (*lfy-4*, E238K), pETH23 (*lfy-20*, N306D) and pCH56 (*lfy-5*, P240L) were derived from pCH28 using the megaprimer strategy with appropriate primers (Kirsch and Joly, 1998). All plasmids were verified by sequencing.

Protein expression, purification and crystallization

Wild-type and mutant LFY-C domains were expressed using *Escherichia coli* strain RosettaBlue(DE3)pLysS (Novagen, Strasbourg, France). After induction by 0.5 mM IPTG, cells were grown overnight at 22°C. For cell lysis, the pellet of 1 l culture was sonicated in 30 ml lysis buffer A (500 mM NaCl, 20 mM Tris–HCl pH 8, 5 mM imidazole, 5% glycerol, 5 mM Tris(2-carboxyethyl)phosphine hydrochloride), one protease inhibitor cocktail tablet Complete EDTA-free (Roche, Meylan, France) and centrifuged for 40 min at 30 000 *g*. The clear supernatant was incubated for about 1 h with 1 ml Ni-NTA resin (Qiagen, Courtaboeuf, France). The resin was transferred into a column, washed with 20 column volumes (CVs) of buffer A, buffer A + 50 mM imidazole (10 CV) and eluted with buffer A + 380 mM imidazole. The fractions containing the protein were pooled and applied to a Hi-load Superdex-200 16/60 prep grade column (GE Healthcare, Orsay, France) equilibrated with 200 mM NaCl, 20 mM Tris–HCl pH 8, 5 mM dithiothreitol (DTT) to eliminate aggregated proteins by SEC. Protein concentration was estimated using the Bradford assay (Bradford, 1976).

For crystallographic experiments, after elution on the metal-affinity column, the histidine tag was cleaved at 4°C overnight with TEV protease (0.01% w/w, 16 h, 4°C) during the dialysis step against buffer B (500 mM NaCl, 20 mM Tris pH 7.5, 5 mM DTT). The TEV protease, the histidine tag and the uncleaved protein were removed by repassing the dialysed sample over the Ni-affinity column. The protein was separated from the remaining DNA contamination using the anion-exchange column MonoQ HR10/10 (GE Healthcare) pre-equilibrated in buffer B. Pure protein was recovered in the flow-through, whereas DNA remained bound to the resin. Aggregated protein was removed by SEC with Superdex S75GL column (GE Healthcare) in 200 mM NaCl, 10 mM Tris pH 7.5 and 5 mM DTT. The protein concentration was adjusted to 7.5 mg/ml. DNA oligonucleotides were chemically synthesized and purified by anion-exchange chromatography following established procedures (Cramer and Muller, 1997).

EMSAs

Single-stranded oligonucleotides, 5′-labelled with tetra-methylcarboxy-rhodamine (Sigma, Saint Quentin Fallavier, France), were annealed to non-fluorescent complementary oligonucleotides in annealing buffer (10 mM Tris pH 7.5, 150 mM NaCl and 1 mM EDTA). The sequences of oligonucleotides used are indicated in Supplementary Table 1. Binding reactions were performed in 20 μl binding buffer (150 mM NaCl, 20 mM Tris–HCl pH 7.5, 1% glycerol, 0.25 mM EDTA, 2 mM $MgCl_2$ and 1 mM DTT) supplemented with 28 ng/μl fish sperm DNA (Roche) and 10 nM double-stranded DNA probe or 140 ng/μl fish sperm DNA for 50 nM DNA probe (Figure 5). Binding reactions were loaded onto native 6% polyacrylamide gels

0.5× TBE (45 mM Tris, 45 mM boric acid and 1 mM EDTA pH 8) and electrophoresed at 90 V for 80 min at 4°C. Gels were scanned on a Typhoon 9400 scanner (Molecular Dynamics, Sunnyvale, CA; excitation light 532 nm, emission filter 580 BP 30) and signals were quantified using ImageQuant software (Molecular Dynamics). Estimations of K_{d1} and K_{d2} (Figure 5; Supplementary Figure 5) were based on the quantifications of binding experiments shown in Figure 5B. The binding model equations used to calculate these K_d values are explained in detail in Supplementary data.

SEC

The molecular size of LFY-C/*AP1* complex was determined using a Superdex-200 10/300GL column (GE Healthcare), equilibrated with buffer containing 150 mM NaCl, 16 mM Tris–HCl pH 7.5, 0.6 mM EDTA and 1 mM DTT, and calibrated with low and high molecular weight protein standards (gel filtration calibration kit; GE Healthcare). Our samples (LFY-C 40 μM, *AP1* WT 10 μM and LFY-C 40 μM + *AP1* WT 10 μM) were analysed in the same buffer as protein standards, and molecular size is deduced from the standard curve.

Analytical SEC and MALLS-SEC

Separation by SEC was carried out with a S200 Superdex column (GE Healthcare). The column was equilibrated in 20 mM Tris–HCl, 150 mM NaCl buffer at pH 7.5. Separations were performed at 20°C with a flow rate of 0.6 ml min^{-1}. Protein solution (50 μl) at a concentration of 5 mg ml^{-1} was injected. The elution was monitored by using a DAWN-EOS detector with a laser emitting at 690 nm for online MALLS measurement (Wyatt Technology Corp., Santa Barbara, CA), and with a RI2000 detector for online refractive index measurements (Schambeck SFD). Molecular mass calculation was performed as described using the ASTRA software (Gerard *et al*, 2007).

Crystallization

For co-crystallization using the hanging drop method, protein and DNA duplexes were mixed in a molar ratio of 2:1. The best crystals were obtained at 4°C with 20-mer oligonucleotides bearing complementary A:T overhangs and with 10% PEG 400, 100 mM KCl, 10 mM $CaCl_2$, 50 mM HEPES (NaOH) pH 7.0 as reservoir solution. Single crystals grew to a maximal size of $300 \times 300 \times 500\ \mu m^3$ and were stepwise transferred to reservoir solution containing 30% (v/v) glycerol for cryo-protection. For preparation of the mercury derivative, the crystals were soaked in the reservoir solution supplemented with 0.1 mM ethylmercury thiosalicylate (EMTS) for 2 h.

X-ray structure determination

The crystals of the LFY-C/*AP1*/DNA complex belong to space group $P6_522$ ($a = b = 98.8$ Å, $c = 177.4$ Å), diffracted up to 2.1 Å resolution and contain half a complex per asymmetric unit. Crystals of the LFY-C/*AG-I*/DNA complex are isomorphous but diffracted slightly weaker (Table I). Diffraction data collected at ESRF beamlines ID14-1, ID29 and ID23-2 were processed using program XDS (Kabsch, 1993). The structure of the LFY-C/*AP1*/DNA complex was solved using the SIRAS method with EMTS as derivative. The quality of native and derivative data sets is summarized in Table I. Mercury sites were located using program SOLVE (Terwilliger and Berendzen, 1999) and phases were calculated with program SHARP (de la Fortelle and Bricogne, 1997). The experimental electron density map (Supplementary Figure 1) allowed us to automatically build the initial model using program ARP/wARP (Perrakis *et al*, 2001) followed by manually adjusting some side chain conformations with program COOT (Emsley and Cowtan, 2004) and refinement with program Refmac5 including a TLS refinement with seven groups (Murshudov *et al*, 1997) and later with program Phenix (Adams *et al*, 2002). In space group $P6_522$, the two monomers bound to the pseudo-palindromic DNA duplex are related by a crystallographic dyad. In the crystal, the pseudo-dyad of the DNA coincides with the crystallographic dyad, although the DNA duplexes deviate from strict two-fold symmetry at base pairs 0, ±7, ±9 and the overhanging 5′-end, in the *AP1* site and at base pairs 0, ±4, ±6 and ±7 and the overhanging 5′-end in the *AG-I* site. To confirm our space group assignment and the underlying assumption that the DNA duplexes used for co-crystallization are randomly distributed in two orientations, the data were reprocessed in the lower symmetry space group $P6_5$ lacking the dyad, which did not significantly change the R_{meas} values. Subsequently, models of the LFY-C dimer bound to the 20-mer DNA duplex were built in space group $P6_5$ for the *AP1* and *AG-I* sites and refined in two independent orientations yielding very similar final R_{cryst} and R_{free} values compared with the refinement in space group $P6_522$. In both orientations (and for both target sites), the final F_o-F_c electron density maps showed pairs of difference Fourier peaks ($\sim 7\sigma$) of similar height at the non-palindromic bases, indicating that a unique orientation of the DNA duplexes does not correctly describe the situation in the crystals. Finally, simulated-annealing omit maps in space group $P6_5$ where the non-palindromic bases were omitted showed averaged densities for the omitted bases in both complexes, further confirming the assigned space group $P6_522$.

To account for the two orientations of the DNA in the crystal during the refinement, two nucleotides with 50% occupancy were introduced at the non-palindromic positions. The final model of the LFY-C/*AP1* complex at 2.1-Å resolution ($R_{crystal} = 21.0\%$; $R_{free} = 23.7\%$) comprises residues 237–399 of the LFY DNA-binding domain, whereas the poorly conserved 25 C-terminal residues are disordered. For the refinement of the LFY/*AG-I* complex, the *AP1* DNA sequence in the LFY/*AP1* complex was replaced with the *AG-I* sequence. Multiple rounds of refinement (including TLS refinement with seven groups) using program Refmac5 (Murshudov *et al*, 1997) and Phenix (Adams *et al*, 2002) yielded a model with R_{cryst} of 22.1% and R_{free} of 24.9% using data between 20 and 2.3 Å resolution. The atomic coordinates and structure factors for the LFY/*AP1* and LFY/*AG-I* complexes have been deposited with the Protein Data Bank under accession codes 2vy1, r2vy1sf and 2vy2, r2vy2sf, respectively.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Acknowledgements

We thank D Weigel for providing materials and advice, X Wu for material, L Blanchoin, C Guérin, R Dumas, M Jamin, C Ebel and G Schoehn for help with protein expression and characterization, L Blanchoin, A Maizel, M Blazquez, E Dorcey and C Petosa for critical reading of the paper, R Russell for structure comparisons, the EMBL/ESRF Joint Structural Biology Group for access and support at the ESRF beamlines and the crystallization facility of the Partnership for Structural Biology for support. Funding was provided by ATIP (CNRS) to FP and RB, ATIP+ (CNRS) and ANR BLAN-0211 to FP, Region Rhône-Alpes/Cluster 9 to CH, Programme Emergence of the Region Rhône-Alpes to DP.

References

Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58:** 1948–1954

Albert VA, Oppenheimer DG, Lindqvist C (2002) Pleiotropy, redundancy and the evolution of flowers. *Trends Plant Sci* **7:** 297–301

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* **29:** 231–262

Blazquez MA, Ferrandiz C, Madueno F, Parcy F (2006) How floral meristems are built. *Plant Mol Biol* **60:** 855–870

Bomblies K, Wang RL, Ambrose BA, Schmidt RJ, Meeley RB, Doebley J (2003) Duplicate *FLORICAULA/LEAFY* homologs *zfl1* and *zfl2* control inflorescence architecture and flower patterning in maize. *Development* **130:** 2385–2395

Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR (1993) Control of flower development in *Arabidopsis thaliana* by *APETALA1* and interacting genes. *Development* **119:** 721–743

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein–dye binding. *Anal Biochem* **72:** 248–254

Breitling R, Gerber JK (2000) Origin of the paired domain. *Dev Genes Evol* **210:** 644–650

Burz DS, Rivera-Pomar R, Jackle H, Hanes SD (1998) Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the *Drosophila* embryo. *EMBO J* **17:** 5998–6009

Busch MA, Bomblies K, Weigel D (1999) Activation of a floral homeotic gene in *Arabidopsis*. *Science* **285:** 585–587

Chae E, Tan QK, Hill TA, Irish VF (2008) An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. *Development* **135:** 1235–1245

Cherry JL, Adler FR (2000) How to make a biological switch. *J Theor Biol* **203:** 117–133

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

Coen ES, Romero JM, Doyle S, Elliot R, Murphy G, Carpenter R (1990) *floricaula*: a homeotic gene required for flower development in *Antirrhinum majus*. *Cell* **63:** 1311–1322

Cramer P, Muller CW (1997) Engineering of diffraction-quality crystals of the NF-kappaB P52 homodimer:DNA complex. *FEBS Lett* **405:** 373–377

de la Fortelle E, Bricogne G (1997) Maximum-likelihood heavy-atom parameter refinement for the multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol* **276:** 472–494

Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* **8:** 675–688

Delano WL (2002) *The PyMOL Molecular Graphics System*. Palo Alto, CA: DeLano Scientific

Diederichs K, Karplus P (1997) Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nat Struct Biol* **4:** 269–275

Dummler A, Lawrence AM, de Marco A (2005) Simplified screening for the detection of soluble fusion constructs expressed in *E. coli* using a modular set of vectors. *Microb Cell Fact* **4:** 34

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60:** 2126–2132

Frohlich MW (2003) An evolutionary scenario for the origin of flowers. *Nat Rev Genet* **4:** 559–566

Frohlich MW, Chase MW (2007) After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature* **450:** 1184–1189

Garvie CW, Wolberger C (2001) Recognition of specific DNA sequences. *Mol Cell* **8:** 937–946

Gehring WJ, Affolter M, Burglin T (1994) Homeodomain proteins. *Annu Rev Biochem* **63:** 487–526

Gerard FC, Ribeiro Ede Jr A, Albertini AA, Gutsche I, Zaccai G, Ruigrok RW, Jamin M (2007) Unphosphorylated rhabdoviridae phosphoproteins form elongated dimers in solution. *Biochemistry* **46:** 10328–10338

Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233:** 123–138

Hong RL, Hamaguchi L, Busch MA, Weigel D (2003) Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *Plant Cell* **15:** 1296–1309

Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, Crickmore MA, Jacob V, Aggarwal AK, Honig B, Mann RS (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131:** 530–543

Kabsch WJ (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Cryst* **26:** 795–800

Khare D, Ziegelin G, Lanka E, Heinemann U (2004) Sequence-specific DNA binding determined by contacts outside the helix-turn-helix motif of the ParB homolog KorB. *Nat Struct Mol Biol* **11:** 656–663

Kirsch RD, Joly E (1998) An improved PCR-mutagenesis strategy for two-site mutagenesis or sequence swapping between related genes. *Nucleic Acids Res* **26:** 1848–1850

Klemm JD, Schreiber SL, Crabtree GR (1998) Dimerization as a regulatory mechanism in signal transduction. *Annu Rev Immunol* **16:** 569–592

Lamb RS, Hill TA, Tan QK, Irish VF (2002) Regulation of *APETALA3* floral homeotic gene expression by meristem identity genes. *Development* **129:** 2079–2086

Lee I, Wolfe DS, Nilsson O, Weigel D (1997) A LEAFY co-regulator encoded by UNUSUAL FLORAL ORGANS. *Curr Biol* **7:** 95–104

Lenhard M, Bohnert A, Jurgens G, Laux T (2001) Termination of stem cell maintenance in *Arabidopsis* floral meristems by interactions between *WUSCHEL* and *AGAMOUS*. *Cell* **105:** 805–814

Liljegren SJ, Gustafson-Brown C, Pinyopich A, Ditta GS, Yanofsky MF (1999) Interactions among *APETALA1*, *LEAFY*, and *TERMINAL FLOWER1* specify meristem fate. *Plant Cell* **11:** 1007–1018

Lohmann JU, Hong RL, Hobe M, Busch MA, Parcy F, Simon R, Weigel D (2001) A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. *Cell* **105:** 793–803

Lohmann JU, Weigel D (2002) Building beauty: the genetic control of floral patterning. *Dev Cell* **2:** 135–142

Maizel A, Busch MA, Tanahashi T, Perkovic J, Kato M, Hasebe M, Weigel D (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. *Science* **308:** 260–263

Meyerowitz EM (1997) Plants and the logic of development. *Genetics* **145:** 5–9

Molinero-Rosales N, Jamilena M, Zurita S, Gomez P, Capel J, Lozano R (1999) *FALSIFLORA*, the tomato orthologue of *FLORICAULA* and *LEAFY*, controls flowering time and floral meristem identity. *Plant J* **20:** 685–693

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53:** 240–255

Ng M, Yanofsky MF (2001) Function and evolution of the plant MADS-box gene family. *Nat Rev Genet* **2:** 186–195

Panne D, Maniatis T, Harrison SC (2004) Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *EMBO J* **23:** 4384–4393

Parcy F, Nilsson O, Bush MA, Lee I, Weigel D (1998) A genetic framework for floral patterning. *Nature* **395:** 561–566

Perrakis A, Harkiolaki M, Wilson KS, Lamzin VS (2001) ARP/wARP and molecular replacement. *Acta Crystallogr D Biol Crystallogr* **57:** 1445–1450

Riechmann JL, Ratcliffe OJ (2000) A genomic perspective on plant transcription factors. *Curr Opin Plant Biol* **3:** 423–434

Rosinski JA, Atchley WR (1999) Molecular evolution of helix-turn-helix proteins. *J Mol Evol* **49:** 301–309

Russell RB, Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14:** 309–323

Schumacher MA, Miller MC, Grkovic S, Brown MH, Skurray RA, Brennan RG (2002) Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. *EMBO J* **21:** 1210–1218

Senear DF, Ross JB, Laue TM (1998) Analysis of protein and DNA-mediated contributions to cooperative assembly of protein–DNA complexes. *Methods* **16:** 3–20

Shiu SH, Shih MC, Li WH (2005) Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol* **139:** 18–26

Souer E, van der Krol A, Kloos D, Spelt C, Bliek M, Mol J, Koes R (1998) Genetic control of branching pattern and floral identity during *Petunia* inflorescence development. *Development* **125:** 733–742

Terwilliger TC, Berendzen J (1999) Automated MAD and MIR structure solution. *Acta Crystallogr D Biol Crystallogr* **55:** 849–861

Theissen G, Melzer R (2007) Molecular mechanisms underlying origin and diversification of the angiosperm flower. *Ann Bot (London)* **100:** 603–619

Underhill DA (2000) Genetic and biochemical diversity in the Pax gene family. *Biochem Cell Biol* **78:** 629–638

van Pouderoyen G, Ketting RF, Perrakis A, Plasterk RH, Sixma TK (1997) Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C. elegans* in complex with transposon DNA. *EMBO J* **16:** 6044–6054

Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM (1992) LEAFY controls floral meristem identity in *Arabidopsis*. *Cell* **69:** 843–859

Weigel D, Jackle H (1990) The fork head domain: a novel DNA binding motif of eukaryotic transcription factors? *Cell* **63:** 455–456

Wu X, Dinneny JR, Crawford KM, Rhee Y, Citovsky V, Zambryski PC, Weigel D (2003) Modes of intercellular transcription factor movement in the *Arabidopsis* apex. *Development* **130:** 3735–3745

Xu HE, Rould MA, Xu W, Epstein JA, Maas RL, Pabo CO (1999) Crystal structure of the human Pax6 paired domain–DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. *Genes Dev* **13:** 1263–1275

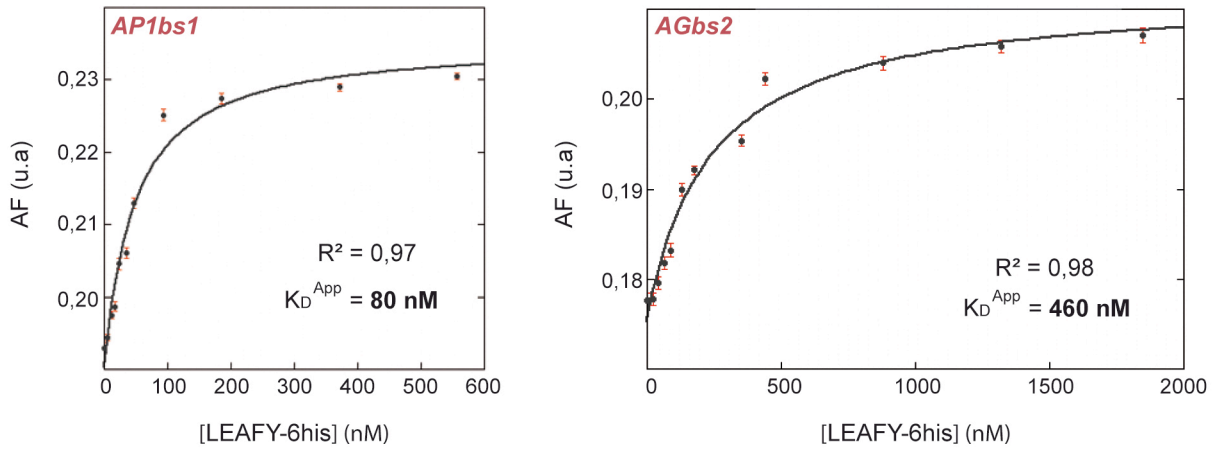
Résultats complémentaires

Spécificité de liaison de LEAFY pour les éléments *cis* connus des gènes homéotiques floraux

Nous avons utilisé l'anisotropie de fluorescence (LeTilly and Royer, 1993) pour comparer la capacité de LFY (protéine entière) à reconnaître deux éléments *cis*, *AP1bs1* (*APETALA1 binding site* 1) et *AGbs2* (*AGAMOUS binding site* 2), respectivement localisés dans le promoteur d'*APETALA1* et le second intron d'*AGAMOUS*. Cette technique de biophysique permet de mesurer rapidement en solution l'affinité globale d'un facteur de transcription pour un ADN cible, *via* une estimation de la constante de dissociation apparente (K_D^{app}) du complexe protéine-ADN. Les séquences des deux éléments *cis* correspondent parfaitement au consensus (A/T)NNNCCANTG(T/G)NNNN(A/T), pourtant le K_D apparent de LFY pour *AP1bs1* est systématiquement plus faible que celui mesuré lorsque *AGbs2* est utilisé, ce qui indique une meilleure reconnaissance de l'élément *cis* d'*APETALA1* par LFY (40 nM pour *AP1bs1* contre 230 nM pour *AGbs2* dans l'exemple présenté Fig.1.1). Ces résultats montrent que deux séquences porteuses du motif (A/T)NNNNCCANTG(G/T)NNNN(A/T) peuvent néanmoins présenter des affinités différentes pour LFY. Ainsi, le second modèle de consensus, établi à partir des données structurales, ne permet pas d'expliquer la différence de comportement observée.

Figure 1.1 | Estimation du K_D^{App} de LFY sur *AP1bs1* et *AGbs2* par anisotropie de fluorescence

Valeur d'anisotropie de fluorescence (AF ; unités arbitraires) en fonction de la concentration de LFY. L'exemple présenté correspond à l'interaction de la protéine avec *AP1bs1* ou *AGbs2*. L'ADN double brin seul couplé à un fluorophore a une forte mobilité rotationnelle: sa valeur d'AF est faible. Lorsque cet ADN fluorescent est lié à une molécule de taille plus importante (ici, la protéine LFY), sa rotation est ralentie donc son AF augmente. La confrontation des équations théoriques aux données expérimentales permet d'estimer la valeur du K_D^{App} (Matériels et Méthodes).



Évolution du complexe LFY-ADN

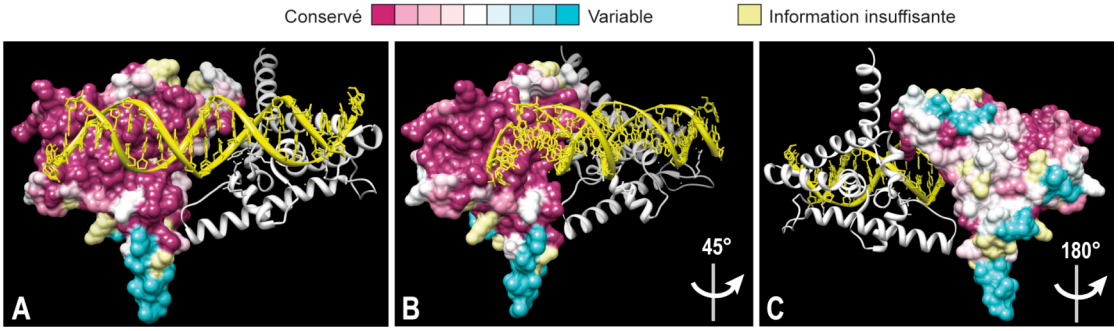
Les bases de données sont particulièrement riches en séquences d'homologues de LFY. En effet, les acteurs de la floraison, en particulier LFY, ont été identifiés chez de nombreuses espèces présentant un fort intérêt agronomique, comme le pommier, la vigne ou le maïs dans le but d'accélérer la mise à fleur et générer de nouvelles variétés capables de produire des fruits plus précocement. De plus, LFY est généralement présent sous forme d'une seule copie au sein d'un génome et l'existence des deux domaines très conservés facilite son identification. Ce gène constitue donc un marqueur moléculaire d'intérêt souvent utilisé en phylogénie pour préciser les relations de parenté entre espèces (Calonje et al., 2009).

Conservation de l'interface LFY-ADN

J'ai aligné les séquences protéiques des homologues de LFY chez plus de 200 espèces, des mousses aux plantes à fleurs. La variabilité existant à chaque position de LFY-C a ensuite été représentée sur la structure tridimensionnelle du facteur de transcription (Fig.1.2). Le modèle obtenu permet de distinguer deux types de comportements. Les résidus dont les chaînes latérales sont tournées vers l'intérieur du monomère et qui constituent le cœur hydrophobe de la protéine ainsi que ceux faisant face à la molécule d'ADN sont extrêmement conservés (Fig.1.2A). Cette conservation remarquable s'étend même au-delà de la région au contact direct des 19 paires de bases utilisées pour la cristallisation (Fig1.2B) suggérant qu'il pourrait exister des interactions supplémentaires entre l'ADN cible et le facteur de transcription. En revanche, la surface de la protéine opposée à l'ADN présente une variabilité importante (Fig.1.2C).

Figure 1.2 | Conservation des acides aminés du domaine de liaison à l'ADN de LFY chez les plantes terrestres.

Un dimère de LFY-C est représenté lié à l'ADN (en jaune). Un monomère apparaît en blanc sous la forme 'ruban' permettant de visualiser les différentes hélices, la surface du second monomère est représentée avec un code couleur indiquant le degré de conservation des acides aminés (programme ConSurf, (Landau et al., 2005)). Le complexe est montré sous 3 angles différents illustrant la conservation de la face en contact avec l'ADN et la variabilité de la face opposée.

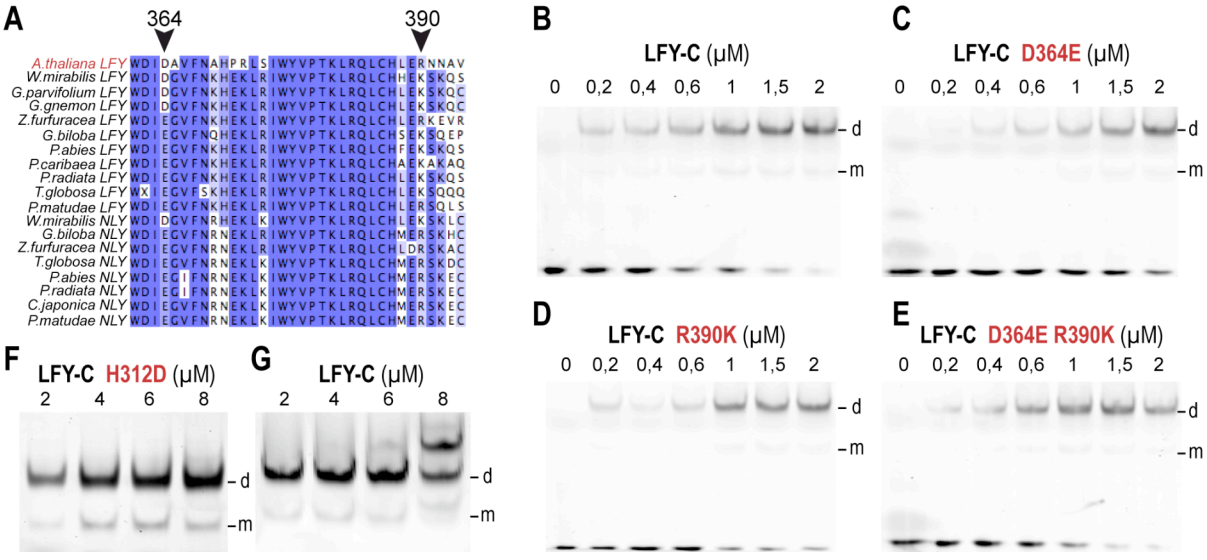


Étude de LFY-C 'ancestralisés'

Les acides aminés impliqués dans l'établissement de la coopérativité ne sont pas toujours conservés chez les groupes ayant divergé avant les plantes à fleurs. Chez les gymnospermes en particulier, l'acide aspartique (D) en position 364 contribuant au positionnement de l'hélice responsable de la coopérativité *via* un pont ionique avec l'arginine 383, est fréquemment remplacé par un acide glutamique (E). De même, une lysine (K) est souvent trouvée en position 390 à la place de l'arginine (R) qui participe normalement à l'une des liaisons hydrogène entre les deux monomères (Fig.1.3A). Les chaînes latérales des résidus E et K n'ayant pas la même longueur que les acides aminés D et R présents chez les angiospermes, la coopérativité pourrait être affectée voire absente chez les gymnospermes. J'ai produit 3 versions mutantes de LFY-C d'*Arabidopsis*, où D364, R390 ou les deux étaient remplacés par un acide glutamique ou une lysine respectivement. Nous avons ensuite testé en retard sur gel la capacité de ces protéines à lier de manière coopérative l'élément *cis AP1bs1*. Les résultats obtenus ne révèlent aucune différence significative de comportement entre la protéine sauvage (Fig.1.3B) et les versions 'ancestralisées' (Fig.1.3C-E) ce qui ne permet pas d'expliquer pourquoi les acides aminés D364 et R390 ont été fixés par l'évolution chez l'ensemble des angiospermes.

Figure 1.3 | Comparaison des capacités de liaison à l'ADN des versions 'ancestralisées' de LFY-C.

(A) Un alignement d'une sous-région du domaine C-terminal de LFY (rouge) et de ses homologues chez plusieurs gymnospermes (noirs) montre la non-conservation des acides D364 et R390. Les positions strictement conservées sont indiquées en bleu. (B-G) Retards sur gel avec 10nM *AP1bs1* et une concentration croissante de LFY-C (B et G), LFY-C D364E (C), LFY-C R390K (D) et LFY-C D364E R390K (E), LFY-C H312D (F). Les différents complexes monomère/ADN (m), dimère/ADN (d) sont indiqués.



PpLFY1, homologue de LFY chez la mousse *Physcomitrella patens*, ne reconnaît pas *AP1bs1* (Maizel et al., 2005). La structure du complexe LFY-C/ADN suggère que l'absence de l'histidine 312 déstabiliserait l'organisation d'ensemble de PpLFY1-C, rendant ainsi la protéine incapable de contacter cette séquence ADN. Afin de tester l'importance fonctionnelle de cette histidine, nous avons produit une version mutante de LFY-C d'*Arabidopsis* présentant un acide aspartique à la place de l'histidine (situation observée chez PpLFY1) et testé sa capacité à lier *AP1bs1*. Contrairement à ce qui est observé avec la protéine de mousse, LFY-C H312D reconnaît parfaitement *AP1bs1*, la perte de l'histidine n'engendrant apparemment aucune perte d'affinité (Fig.1.3F). Il doit donc exister des différences supplémentaires entre la protéine d'Arabidopsis et celle de mousse. Lorsqu'une forte concentration de LFY-C est utilisée (≥8µM), un complexe supplémentaire de taille supérieure est observé (Fig.1.3G) qui contient sans doute plus de deux LFY-C. Ce complexe n'est jamais observé lorsque LFY-C H312D est utilisé, même à forte concentration. Ainsi même si l'histidine 312 n'est pas indispensable à la protéine d'*Arabidopsis* pour contacter un ADN cible, l'absence de cet acide aminé semble perturber sa capacité à former des complexes d'ordre supérieur, ce que la structure du domaine de liaison à l'ADN ne permettait pas d'envisager.

Conclusion

Outre les résultats structuraux et mécanistiques obtenus, cette première étude a généré de nouvelles interrogations. En effet, deux séquences d'ADN porteuses du motif préféré par LFY ne sont pas reconnues avec la même efficacité par la protéine. Le consensus ne rend donc pas compte de manière satisfaisante de la spécificité de liaison du facteur de transcription. Dès lors, comprendre le rôle de LFY nécessite de construire un meilleur modèle de prédiction des sites reconnus par la protéine.

La structure obtenue suggère qu'une modification de la spécificité de LFY au cours de l'évolution semble peu probable puisque l'interface avec l'ADN est extrêmement conservée. Pourtant, les comportements de LFY et PpLFY1 sont extrêmement différents alors que PpLFY1 possède tous les acides aminés impliqués dans les contacts avec l'ADN. La conservation de la séquence protéique ne permet donc pas de prédire les capacités de liaison du facteur, soulignant ainsi les limites de la modélisation.

Dans la continuité du travail de thèse, nous avons cherché à répondre à ces questions. Les résultats obtenus sont présentés dans les deux chapitres suivants.

CHAPITRE 2

Prédiction et évolution de la spécificité de liaison à l'ADN de LEAFY

La structure tridimensionnelle du complexe LFY-C/ADN a permis de visualiser comment LFY interagit avec un ADN cible. Dans un second temps, nous avons cherché à comprendre pourquoi LFY reconnaît une séquence nucléotidique donnée. Cela nous a conduit à étudier les règles de reconnaissance de sites au sein du génome par ce facteur central du développement floral.

Après avoir mis au point les outils permettant de construire un modèle prédictif de liaison de LFY, nous avons utilisé ce modèle pour revisiter les relations établies entre LFY et certains de ses gènes cibles chez Arabidopsis *et autres angiospermes. Les résultats obtenus sont présentés dans l'article (soumis) de ce chapitre. Nous avons ensuite commencé à retracer l'évolution de la spécificité de liaison de LFY en appliquant les méthodologies développées pour la protéine d'*Arabidopsis *à ses homologues chez plusieurs espèces clés de la phylogénie des végétaux. Les résultats préliminaires de cette étude constituent la dernière section du chapitre. Ils serviront de point de départ pour explorer les fonctions originelles de ce facteur unique.*

Introduction

LFY est capable de se lier spécifiquement à certaines séquences d'ADN ce qui participe directement à la sélection de ses gènes cibles. Comprendre les règles qui gouvernent cette liaison est nécessaire pour prédire l'architecture complète des réseaux génétiques qu'il gouverne.

Or, le rôle joué par LFY dépend du répertoire de gènes qu'il contrôle. Retracer son évolution fonctionnelle revient donc à étudier le devenir des régulations orchestrées par LFY au travers de la lignée verte. Plus d'une dizaine de génomes de plantes sont disponibles et les nouvelles techniques de séquençage permettent d'envisager l'accès au génome complet de toute espèce d'intérêt. L'annotation semi-automatique de ces génomes permet d'identifier assez facilement les régions codantes. En revanche, l'analyse des régions régulatrices environnantes reste compliquée. Il s'agit généralement de régions modulaires, constituées d'un ensemble d'éléments *cis* courts souvent décrits comme de petits motifs dégénérés localisés aussi bien dans les régions intergéniques que dans les séquences codantes elles-mêmes et donc difficiles à repérer.

Disposer d'un modèle prédictif de liaison permettrait de localiser ces éléments *cis* de manière fiable et automatisée. Il serait alors possible de comparer les régions régulatrices de plusieurs

espèces entre elles sur la base de la présence ou absence de sites de liaison pour LFY et ses homologues, et ainsi de proposer des hypothèses sur l'état des réseaux.

Contexte

Jusqu'à présent, le motif nucléotidique reconnu par LFY a été représenté sous la forme d'un consensus qui permet uniquement de trier les séquences ADN en deux classes selon que celles-ci présentent ou non les bases préférées aux positions contraintes ('bon' ou 'mauvais' sites respectivement). L'utilisation d'un consensus ne permet pas de prendre en compte des différences d'affinité entre sites de liaison et suppose que les mauvais sites ne sont jamais reconnus. Comme montré dans les systèmes animaux, le consensus est un outil de prédiction pauvre et peu adapté générant beaucoup de faux positifs et de faux négatifs (Wasserman and Sandelin, 2004).

Pour disposer d'un outil plus performant, il faut développer un modèle quantitatif capable de prédire l'affinité de tout motif ADN pour LFY. Ce modèle doit être suffisamment sensible (peu de faux négatifs) et spécifique (peu de faux positifs) pour identifier les régions reconnues par LFY à l'échelle du génome. Afin de construire un tel outil, une approche consiste à déchiffrer le code de reconnaissance ADN-protéine à partir d'un grand nombre de séquences de bonne affinité pour la protéine d'intérêt. Il existe plusieurs techniques permettant d'obtenir ces séquences (Mukherjee et al., 2004; Maerkl and Quake, 2007; Segal and Widom, 2009) et nous avons choisi de développer la méthode du Selex (*Systematic Evolution of Ligand by EXponential enrichment*). Cette méthode, basée sur l'itération de cycles, permet la purification progressive de séquences présentant une haute affinité pour la protéine étudiée (Stoltenburg et al., 2007). Chaque cycle est constitué d'une phase de sélection, au sein d'un mélange d'oligonucléotides aléatoires ou aptamères, des séquences présentant une bonne affinité pour le facteur d'intérêt suivie d'une phase d'amplification des aptamères sélectionnés pour constituer un nouveau mélange de meilleure affinité qui sera réutilisé au tour suivant. Après plusieurs cycles d'enrichissement, les ADN retenus sont séquencés et alignés afin de mettre en évidence le(s) motif(s) reconnu(s) par le facteur de transcription. Une matrice de poids (*Position weight matrix* ou *Position-Specific Scoring Matrix* PSSM) ou profil généralisé, construite à partir de l'alignement, rend compte de la probabilité de présence des 4 nucléotides à chacune des positions du motif (Stormo, 2000; Bulyk, 2003; Wasserman and Sandelin, 2004). Dès lors que le nucléotide d'une position donné ne correspond pas à la base préférée par LFY, la PSSM lui attribue une pénalité sous forme d'une valeur négative

directement reliée à la fréquence du nucléotide à cette position. Un score peut ainsi être obtenu pour tout motif ADN par simple addition des pénalités. D'après la théorie, ce score est directement proportionnel au logarithme du K_D (inverse de l'affinité) LFY pour cette séquence (Berg and von Hippel, 1987). Cette matrice peut finalement être utilisée pour scanner les régions régulatrices d'intérêt et identifier les sites de liaison potentiels qu'elles contiennent (Stormo, 2000; Vingron et al., 2009).

Dans un premier temps, nous avons utilisé la protéine LFY-C d'*Arabidopsis* pour mettre au point les conditions expérimentales permettant la sélection d'aptamères de haute affinité. Le protocole développé a permis l'obtention d'un set d'ADNs spécifiquement reconnus par LFY-C en 5 à 7 cycles de sélection. Après séquençage haut débit, les séquences obtenues ont permis la construction d'un modèle biophysique capable d'estimer la propension de tout motif ADN à être reconnu spécifiquement par LFY-C. Enfin, nous avons montré que ce modèle expliquait bien les résultats d'immunoprécipitation de la chromatine obtenus *in vivo* grâce à un anticorps anti-LFY et pouvait servir à l'identification *in silico* d'homologues fonctionnels. L'ensemble de ce travail est présenté dans l'article qui suit.

Identification of functional homologs based on *cis*-element detection

Edwige Moyroud[1], Eugenio Gómez Minguet[1], Felix Ott[2], Levi Yant[2], David Posé[2], Sandrine Blanchet[1], Marie Monniaux[1], Olivier Bastien[1], Emmanuel Thévenon[1], Detlef Weigel[2], Markus Schmid[2] and Francois Parcy[1]

[1]Laboratoire de Physiologie Cellulaire Végétale, UMR5168, Centre National de la Recherche Scientifique, Commissariat à l'Énergie Atomique, Institut National de la Recherche Agronomique, Université Joseph Fourier, 17 av. des Martyrs, bât. C2, 38054 Grenoble, France.

[2]Max Planck Institute for Developmental Biology, Department of Molecular Biology, 72076 Tübingen, Germany.

Correspondence should be addressed to F.P. (francois.parcy@cea.fr) or M.S. (markus.schmid@tuebingen.mpg.de)

Despite great advances in sequencing capacity, generating functional information for non-model organisms remains a challenge. One solution lies in improved ability to predict genetic circuits based on primary DNA sequence and characterization of regulatory molecules from model species.

Here, we focus on the LEAFY (LFY) transcription factor, a conserved master regulator of early floral development[1,2]. Starting with biochemical and structural information, we built a biophysical model describing LFY DNA binding specificity *in vitro* that accurately predicts *in vivo* LFY binding sites in the *Arabidopsis thaliana* genome. Extending the model to other species, we show that it correctly identifies homologs of known LFY targets in *Arabidopsis thaliana*, even when a functional shift between orthologs and paralogs has occurred. Our strategy should be generalizable to infer regulatory relationships from primary genome sequence.

New technologies rapidly deliver whole genome sequences from a wide variety of organisms at low cost, but functional annotation of these genomes remains a major challenge. While conserved protein sequences are easily identified, transcriptional *cis*-regulatory modules can be evolutionarily fluid[3-5]. Apart from the ability of many transcription factors to bind a wide variety of related sequences[6,7], a further complication is functional shifts among homologs, such that evolutionary orthology does not always indicate functional equivalence[8,9]. Recognizing functional homologs between species therefore requires the identification of transcriptional *cis*-elements in their regulatory regions that determine their spatio-temporal expression patterns. However, these small and degenerated motifs are often difficult to detect based on sequence conservation alone[3,4,10]. To address this problem, we have focused on the genetic circuitry downstream of LFY, a transcription factor with a central role in the evolution and development of flowers[1,2]. LFY directly controls the expression of several homeotic MADS box genes, including *AGAMOUS* (*AG*), *APETALA1* (*AP1*) and *APETALA3*, through a poorly defined consensus sequence, CCANTG[G/T][11,12]. The three-dimensional structure of the LFY DNA binding domain has, however, revealed contacts over 19 base pairs, suggesting greater specificity than previously thought[13]. Here, we describe the development and application of a biophysical model based on *in vitro* binding and optimized using genome-wide *in vivo* data.

We determined the DNA binding preferences of the LFY DNA binding domain (DBD) by SELEX[14]. Alignment of 494 unique sequences revealed a 19 bp motif (Fig. 1), in good agreement with the 3D structure of LFY DBD complexed with DNA[13]. From this motif, which displays the previously established 7 bp consensus as the core, we deduced an asymmetric (ASY) position specific scoring matrix[6] (PSSM) (Fig. 1 and Supplementary Table 1). Using quantitative multifluorescence relative affinity (QuMFRA) assays[15], we found that the ASY matrix predictions correlated well with experimentally measured DNA binding affinities (Pearson correlation, $r^2 = 0.59$) (Fig. 1). Since the LFY DBD binds DNA as a symmetric homodimer[13], we sought to improve the PSSM by imposing symmetry. With the corresponding SYM matrix, r^2 increased to 0.69. Simple PSSMs assume that different positions contribute independently to the overall binding, a condition that is not always satisfied[16]. We observed non-independent triplets at two symmetric positions and in the center of the alignment (Supplementary Fig. 1), and modeled this dependence using the frequency of trinucleotides (Fig. 1). The final SYM-T matrix further increased r^2 to 0.81. Notably, while the SYM-T matrix was well correlated with experimental DNA binding affinities, the simple presence or absence of the 7 bp consensus motif was not always a good predictor of binding, demonstrating the usefulness of the PSSM approach (Fig. 1B-E).

To determine how well the *in vitro* determined DNA binding specificity correlated with *in vivo* binding, we performed chromatin immunoprecipitation (ChIP) experiment with LFY-specific antibodies (Online Methods), followed by Illumina sequencing (ChIP-seq), on seedlings that overexpressed LFY. The regions enriched in overexpressors as compared to wild-type seedlings were ordered using the rank product from both ChIP-seq replicates, and according to their predicted occupancy (POcc), computed using a biophysical model[17] that takes into account all 19 bp sites present on the DNA fragment. There was a good correlation between the predicted and observed rankings, which increased from the ASY (Spearman's rank correlation coefficient, 0.44) and the SYM (0.45) to the SYM-T matrix (0.53).

As further validation, we performed a Receiver Operating Characteristic (ROC) analysis comparing the 1,564 most enriched regions (false discovery rate < 0.1 in both replicates; Supplementary Table 2) with random negative regions. We evaluated the performance of a biophysical model that integrates all sites present on the fragment and of a hit-based model that selects binding sites with a score higher than a cut-off value[17]. The threshold model was best, but both of our models performed very well compared to other studies[17] (Fig. 2).

The most highly ranked ChIP-enriched fragment was in the 3' region of a gene repressed by LFY, *TERMINAL FLOWER1* (*TFL1),* which has important regulatory elements downstream of the transcribed region[18,19]. Another highly ranking region was centered on a previously identified LFY binding site in the promoter of the well-characterized target *AP1*[20,21], as well as a second peak in the first intron (Fig. 3). This strongly suggests that LFY represses *TFL1* both directly, as proposed before based on experiments with an activated form of LFY[22], and indirectly, through *AP1* activation[18]. For both *AP1* and *TFL1*, the similarity between the ChIP-seq profile and the computed binding site landscapes was striking (Fig. 3), underscoring the predictive power of the SYM-T binding model.

The *AG* second intron contains two previously characterized LFY binding sites[11], which were included in a ChIP-enriched fragment. A conserved site in the *AG* intron that conforms to the initially proposed 7 bp consensus[23], had, however, a very low PSSM score, and was bound neither *in vitro* nor *in vivo* (Figs. 1B and 3C). Conversely, one of the most highly scoring sites had not been identified before, as its sequence is not conserved and lacks the original 7 bp consensus motif (Figs. 1B and 3C).

Several other floral regulators, such as *SEPALLATA4* (Fig. 3D), were identified among the ChIP-enriched fragments, along with genes related to hormone signaling, a process known to be important for flower development (Supplementary Table 3). Expression of several of these genes is affected in *lfy* mutants and there is a significant overlap (p-value = 0.025) with genes deregulated in *LFY-GR* overexpressing plants[24,25] (Supplementary Table 4).

A major motivation for developing predictive DNA binding models is the functional annotation of genomes from other plants. For proof of concept, we examined the large intron of *AG* homologs, since it is known to be important for *AG* regulation in various species and contains several conserved motifs[11,23,26-28]. *AG* belongs to a small subfamily of MADS box genes[29,30]. A first duplication led to the formation of the *AG* and *AGL11* lineages at the base of the angiosperms, and a second duplication in ancestral core eudicots to the *euAG* and *PLE* lineages[31] (Fig. 4A). All these proteins have similar DNA binding and protein-protein interaction profiles and it is thought that they evolved specific functions primarily through diversification of their expression patterns[29,30]. As a result of complex evolutionary trajectories

in this MADS box subfamily, paralogs have sometimes usurped the function of orthologs[8]. Sequence similarity and genomic position are therefore not sufficient to predict functional equivalence with *AG* in other species.

In *Arabidopsis, AG*, *SHATTERPROOF1-2* (*SHP)* and *SEEDSTICK* (*STK*) belong to the *euAG*, *PLE* and *AGL11* lineages respectively. Of these, only *AG* is involved in early floral patterning (reproductive organ primordia specification, C-function), while the other genes play later roles in fruits and ovules[32,33]. In both *A. thaliana* and its relative *A. lyrata*, the predicted occupancy of the second intron by LFY is much higher for *AG* than for *SHP1*, *SHP2* and *STK*. As the structural models indicate that the LFY-DNA interface is highly conserved in angiosperms[34], we next applied our threshold based biophysical model of LFY binding to the large intron of *AG* homologs of other species. In the dicot *A. majus*, *PLENA* (*PLE*) has an *AG*-like function and is activated by the LFY ortholog FLORICAULA, even though it is the *SHP* ortholog[8,27]. Consistent with these observations, our model predicted much higher LFY occupancy for the second intron of *PLE* than for the *AG* ortholog *FARINELLI*. In other dicot species, where less functional data is available, the predicted LFY occupancy of introns of eu*AG*, *PLE* and *AGL11* orthologs was in good agreement with whether or not a gene was expressed during early stages of flower development, when *LFY* is active (Fig. 4, Supplementary Table 5).

As in dicots, monocot *AG* orthologs are expressed before *AGL11* orthologs and, in grasses, there are two *AG* orthologs that share the C-function[35]. Consistently, our model correctly predicted much higher DNA occupancy by LFY for both *AG* orthologs compared to *AGL11* genes. These results imply that regulation of *AG* by LFY evolved in the common ancestor of dicots and monocots, as proposed before[26].

While our model correctly predicts global LFY occupancy in the large introns of *AG* homologs, both the relative position of individual binding sites within these introns and their sequence are highly variable (Supplementary Fig. 2). This is in agreement with recent comparative genome-wide analyses of transcription factor binding sites in vertebrate or yeast species[3,10]. Importantly, the strength of our model comes from its ability to identify individual binding sites, where other methods based on sequence conservation or on the presence of the core consensus fail: when we compared recent *euAG* duplicates, such as *VvAG1* and *VvAG2*

(Supplementary Fig. 2), we often observed that good binding sites were present in both copies even if sequence similarity was not obvious.

In conclusion, we propose that the type of biophysical model that we present, built on extensive knowledge established from a few model organisms, enables new ways for the functional annotation of genomes of non-model species. If generalized to multiple transcription factors, it should greatly help to predict the evolution of regulatory networks directly from whole genome sequences. Such knowledge will have direct applications, for example by pinpointing, among a set of paralogs, the functional homologs of important regulators identified in model organisms.

References

1. Moyroud, E., Kusters, E., Monniaux, M., Koes, R. & Parcy, F. LEAFY blossoms. *Trends Plant Sci.* (2010).
2. Liu, C., Thong, Z. & Yu, H. Coming into bloom: the specification of floral meristems. *Development* **136**, 3379-91 (2009).
3. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036-40 (2010).
4. Wilson, M.D. & Odom, D.T. Evolution of transcriptional control in mammals. *Curr Opin. Genet. Dev.* **19**, 579-85 (2009).
5. Weirauch, M.T. & Hughes, T.R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66-74.
6. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-87 (2004).
7. Badis, G. et al. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-3 (2009).
8. Causier, B. et al. Evolution in action: following function in duplicated floral homeotic genes. *Curr. Biol.* **15**, 1508-12 (2005).
9. Wu, H., Mao, F., Olman, V. & Xu, Y. Hierarchical classification of functionally equivalent genes in prokaryotes. *Nucleic Acids Res.* **35**, 2125-40 (2007).
10. Ward, L.D. & Bussemaker, H.J. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* **24**, i165-71 (2008).
11. Busch, M.A., Bomblies, K. & Weigel, D. Activation of a floral homeotic gene in Arabidopsis. *Science* **285**, 585-7. (1999).
12. Lamb, R.S., Hill, T.A., Tan, Q.K. & Irish, V.F. Regulation of *APETALA3* floral homeotic gene expression by meristem identity genes. *Development* **129**, 2079-86 (2002).
13. Hames, C. et al. Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *Embo J.* **27**, 2628-37 (2008).
14. Zhao, Y., Granas, D. & Stormo, G.D. Inferring binding energies from selected binding sites. *PLoS Comput. Biol.* **5**, e1000590 (2009).
15. Man, T.K. & Stormo, G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **29**, 2471-8 (2001).
16. Benos, P.V., Bulyk, M.L. & Stormo, G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* **30**, 4442-51 (2002).
17. Roider, H.G., Kanhere, A., Manke, T. & Vingron, M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23**, 134-41 (2007).
18. Kaufmann, K. et al. Orchestration of floral initiation by *APETALA1*. *Science* **328**, 85-9 (2010).
19. Ratcliffe, O.J., Bradley, D.J. & Coen, E.S. Separation of shoot and floral identity in Arabidopsis. *Development* **126**, 1109-20 (1999).
20. Parcy, F., Nilsson, O., Busch, M.A., Lee, I. & Weigel, D. A genetic framework for floral patterning. *Nature* **395**, 561-566 (1998).
21. Wagner, D., Sablowski, R.W. & Meyerowitz, E.M. Transcriptional activation of *APETALA1* by LEAFY. *Science* **285**, 582-4. (1999).
22. Parcy, F., Bomblies, K. & Weigel, D. Interaction of *LEAFY*, *AGAMOUS* and *TERMINAL FLOWER1* in maintaining floral meristem identity in Arabidopsis. *Development* **129**, 2519-27 (2002).

23. Hong, R.L., Hamaguchi, L., Busch, M.A. & Weigel, D. Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *Plant Cell* **15**, 1296-309 (2003).
24. Schmid, M. et al. Dissection of floral induction pathways using global expression analysis. *Development* **130**, 6001-12 (2003).
25. William, D.A. et al. Genomic identification of direct target genes of LEAFY. *Proc. Natl. Acad. Sci. U S A* **101**, 1775-80 (2004).
26. Causier, B., Bradley, D., Cook, H. & Davies, B. Conserved intragenic elements were critical for the evolution of the floral C-function. *Plant J.* (2008).
27. Davies, B. et al. *PLENA* and *FARINELLI*: redundancy and regulatory interactions between two Antirrhinum MADS-box factors controlling flower development. *EMBO J.* **18**, 4023-34 (1999).
28. Sieburth, L.E. & Meyerowitz, E.M. Molecular dissection of the *AGAMOUS* control region shows that cis elements for spatial regulation are located intragenically. *Plant Cell* **9**, 355-65 (1997).
29. Ferrario, S., Immink, R.G. & Angenent, G.C. Conservation and diversity in flower land. *Curr. Opin. Plant Biol.* **7**, 84-91 (2004).
30. Zahn, L.M. et al. Conservation and divergence in the *AGAMOUS* subfamily of MADS-box genes: evidence of independent sub- and neofunctionalization events. *Evol. Dev.* **8**, 30-45 (2006).
31. Kramer, E.M., Jaramillo, M.A. & Di Stilio, V.S. Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. *Genetics* **166**, 1011-23 (2004).
32. Liljegren, S.J. et al. *SHATTERPROOF* MADS-box genes control seed dispersal in Arabidopsis. *Nature* **404**, 766-70 (2000).
33. Colombo, M. et al. A new role for the *SHATTERPROOF* genes during Arabidopsis gynoecium development. *Dev Biol* **337**, 294-302 (2010).
34. Moyroud, E., Tichtinsky, G. & Parcy, F. The LEAFY floral regulators in Angiosperms: Conserved proteins with diverse roles. *J. Plant. Biol.* **52**, 177-185 (2009).
35. Thompson, B.E. & Hake, S. Translational biology: from Arabidopsis flowers to grass inflorescence architecture. *Plant Physiol.* **149**, 38-45 (2009).

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGEMENTS:

We thank C. Scutt, P. Lemaire, R. Vincentelli, K. Nitta and members of the Parcy and Schmid laboratories for discussion, and A.K. Martin for help with bioinformatic analyses. Supported by funding from the Centre National de la Recherche Scientifique (ATIP+; F.P.), the Agence Nationale de la Recherche (ANR, Plant-TFcode; F.P.), the ANR and the Biotechnology and Biological Sciences Research Council (Flower Model; F.P.), and PhD fellowship from the University J. Fourier, Grenoble (E.M and M.M.), FP7 Collaborative Project AENEAS (contract KBBE-2009-226477; D.W.), ERA-NET Plant Genomics Project BLOOM-NET (SCHM 1560/7-1; MS) and the Max Planck Society (M.S., D.W.).

AUTHORS CONTRIBUTIONS

F.P., E.M., M.S. and L.Y. designed the experiments, E.M., L.Y., D.P., S.B., E.T. performed the experiments. E.G.M and F.O. performed the data analysis with contributions from E.M., O.B. and M.M., the manuscript was written by F.P., M.S and D.W. with contributions from E.M and L.Y.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Figure legends

Figure 1. Optimization of the LFY binding site model. (**a**) Enrichment of sequences bound by LFY over different SELEX cycles. (**b**) Binding of LFY to different sequences, either from *AG* or *AP1* genes, or synthetic (S). (**c-e**) Comparison of experimentally determined and predicted scores (Online Methods) for different DNA sequences with the three PSSM, illustrated below by their logos. Open and closed circles represent sequences with or without the CCANTG[G/T] consensus, respectively.

Figure 2. Comparison of the different models for prediction of *in vivo* LFY binding sites. Receiver operating characteristic (ROC) curves for LFY-bound and unbound sequences, using a biophysical model taking all sites into account or only those with a SYM-T matrix score higher than -23.

Figure 3. Examples of LFY-bound regions identified by ChIP-seq. Non-coding and coding sequences in exons are shown on top as open and closed boxes, respectively. ChIP-seq read coverage combined from both strands are shown in the middle, and comparison of SYM-T model scores to presence of CCANTG[G/T] consensus (arrows) on the bottom.

Figure 4. Prediction of LFY occupancy of the large intron of *AG* homologs using the SYM-T model. (**a**) Schematic phylogeny of *AG* homologs after[31]. (**b-c**) Predicted occupancy (POcc) of *AG* homologs in monocots (**b**) and eudicots (**c**).

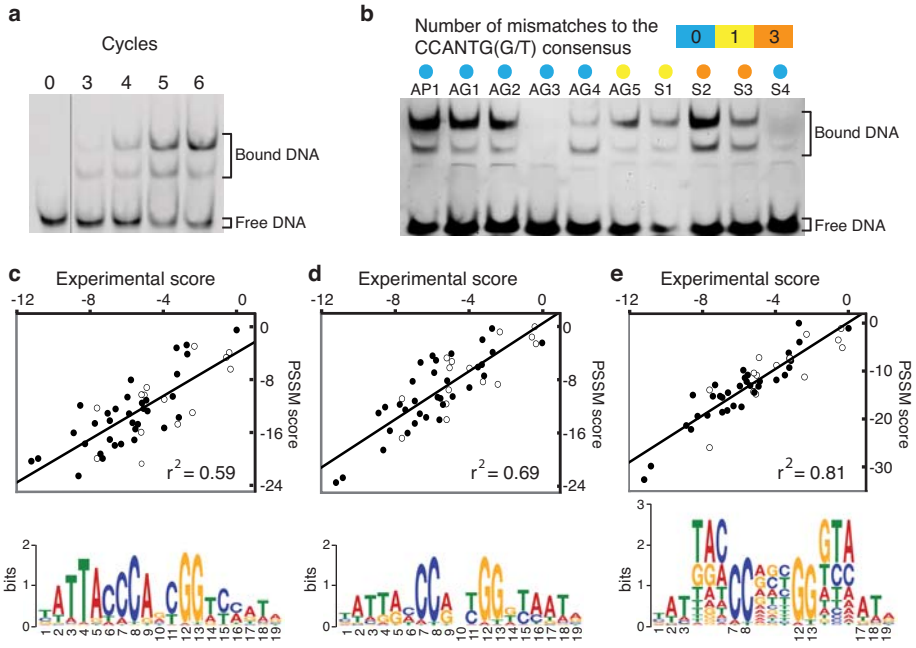

a
Cycles
0 3 4 5 6
Bound DNA
Free DNA
b
Number of mismatches to the CCANTG(G/T) consensus
0 1 3
AP1 AG1 AG2 AG3 AG4 AG5 S1 S2 S3 S4
Bound DNA
Free DNA
c
Experimental score
PSSM score
r2 = 0.59
bits
d
Experimental score
PSSM score
r2 = 0.69
bits
e
Experimental score
PSSM score
r2 = 0.81
bits

Figure 1

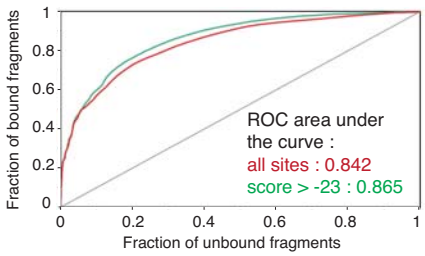

Fraction of bound fragments
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Fraction of unbound fragments
ROC area under
the curve :
all sites : 0.842
score > -23 : 0.865

Figure 2

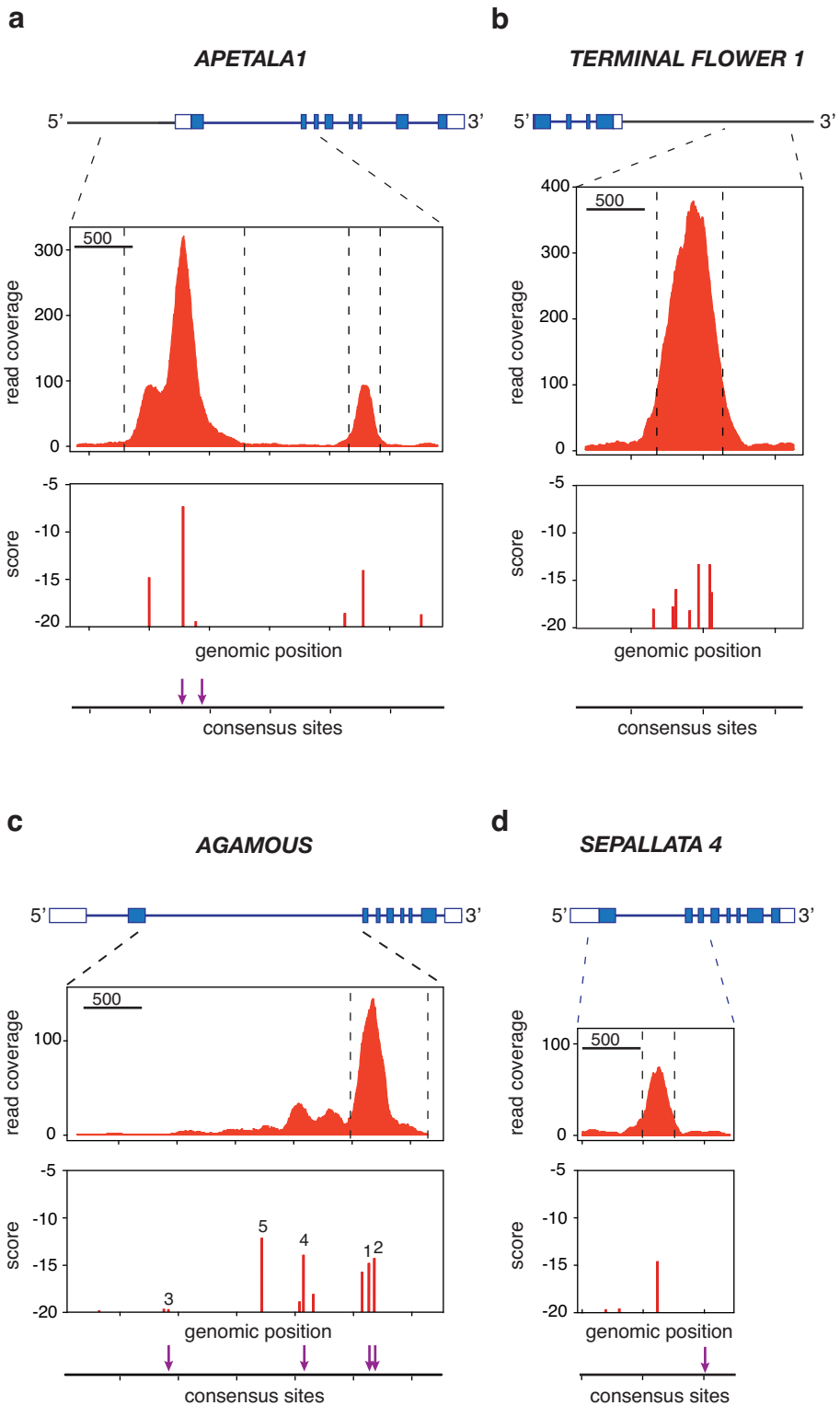

a
APETALA1
5'
3'
500
300
200
100
0
read coverage
-5
-10
-15
-20
score
genomic position
consensus sites
b
TERMINAL FLOWER 1
5'
3'
400
500
300
200
100
0
read coverage
-5
-10
-15
-20
score
genomic position
consensus sites
c
AGAMOUS
5'
3'
500
100
0
read coverage
-5
-10
-15
-20
score
5
4
3
1
2
genomic position
consensus sites
d
SEPALLATA 4
5'
3'
500
100
0
read coverage
-5
-10
-15
-20
score
genomic position
consensus sites

Figure 3

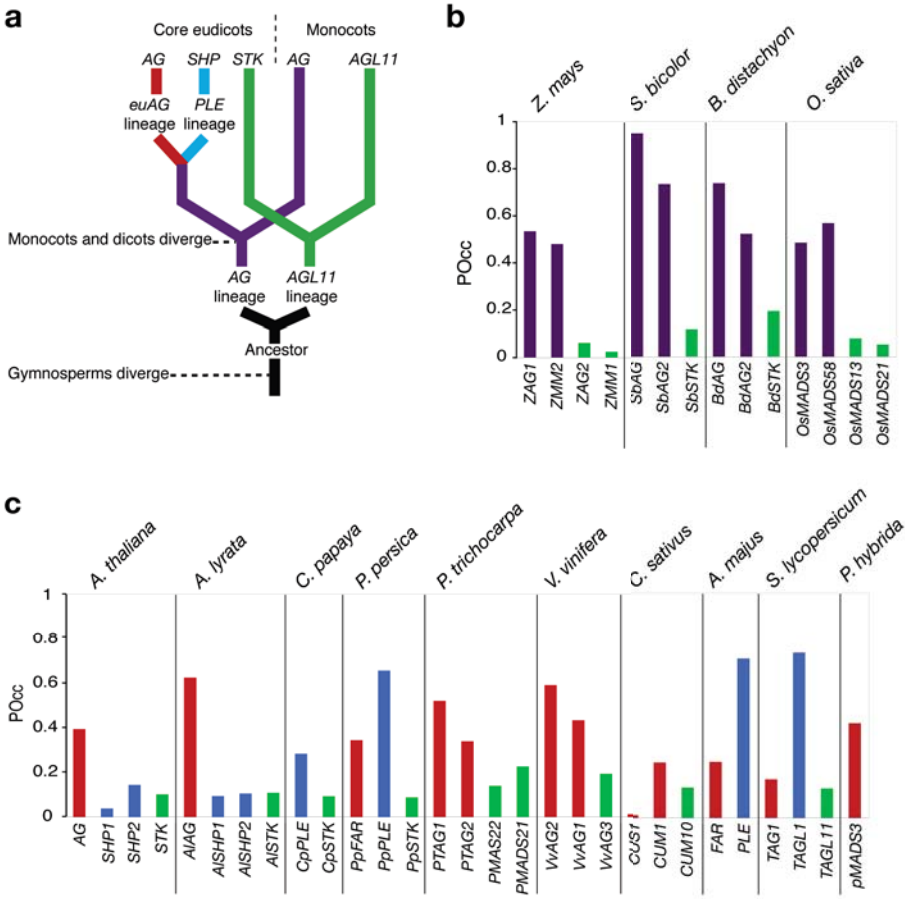

a
Core eudicots
Monocots
AG
SHP
STK
AG
AGL11
euAG lineage
PLE lineage
Monocots and dicots diverge
AG lineage
AGL11 lineage
Ancestor
Gymnosperms diverge
b
Z. mays
S. bicolor
B. distachyon
O. sativa
POcc
ZAG1
ZMM2
ZAG2
ZMM1
SbAG
SbAG2
SbSTK
BdAG
BdAG2
BdSTK
OsMADS3
OsMADS58
OsMADS13
OsMADS21
c
A. thaliana
A. lyrata
C. papaya
P. persica
P. trichocarpa
V. vinifera
C. sativus
A. majus
S. lycopersicum
P. hybrida
POcc
AG
SHP1
SHP2
STK
AlAG
AlSHP1
AlSHP2
AlSTK
CpPLE
CpSTK
PpFAR
PpPLE
PpSTK
PTAG1
PTAG2
PMAS22
PMADS21
VvAG2
VvAG1
VvAG3
CUS1
CUM1
CUM10
FAR
PLE
TAG1
TAGL1
TAGL11
pMADS3

Figure 4

Supplementary Online Materials

This file contains:

Supplementary figures 1 and 2

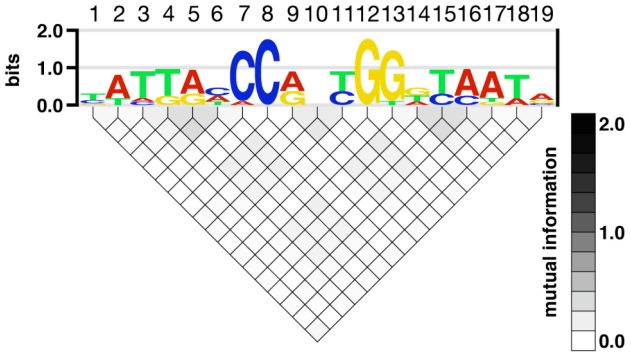
Supplementary tables 1 and 5, legends for Supplementary tables 2-4.

Material and methods

Supplementary Tables 2-4 are provided separately as several sheets in a single .xls document.

Supplementary Figure 1: Detection of dependence between positions of LFY binding sites.

Alignment of the 494 SELEX sequences was analyzed with the enoLOGOS software [1]. The mutual information of each pair of positions of the alignment is displayed below the logo corresponding to the SYM PSSM as a grey-scale-coded matrix plot. Dependence is detected between positions 4,5 and 6 or 14, 15 and 16 (lateral triplets) and, at a lower level between positions 9, 10 and 11 (central triplets).



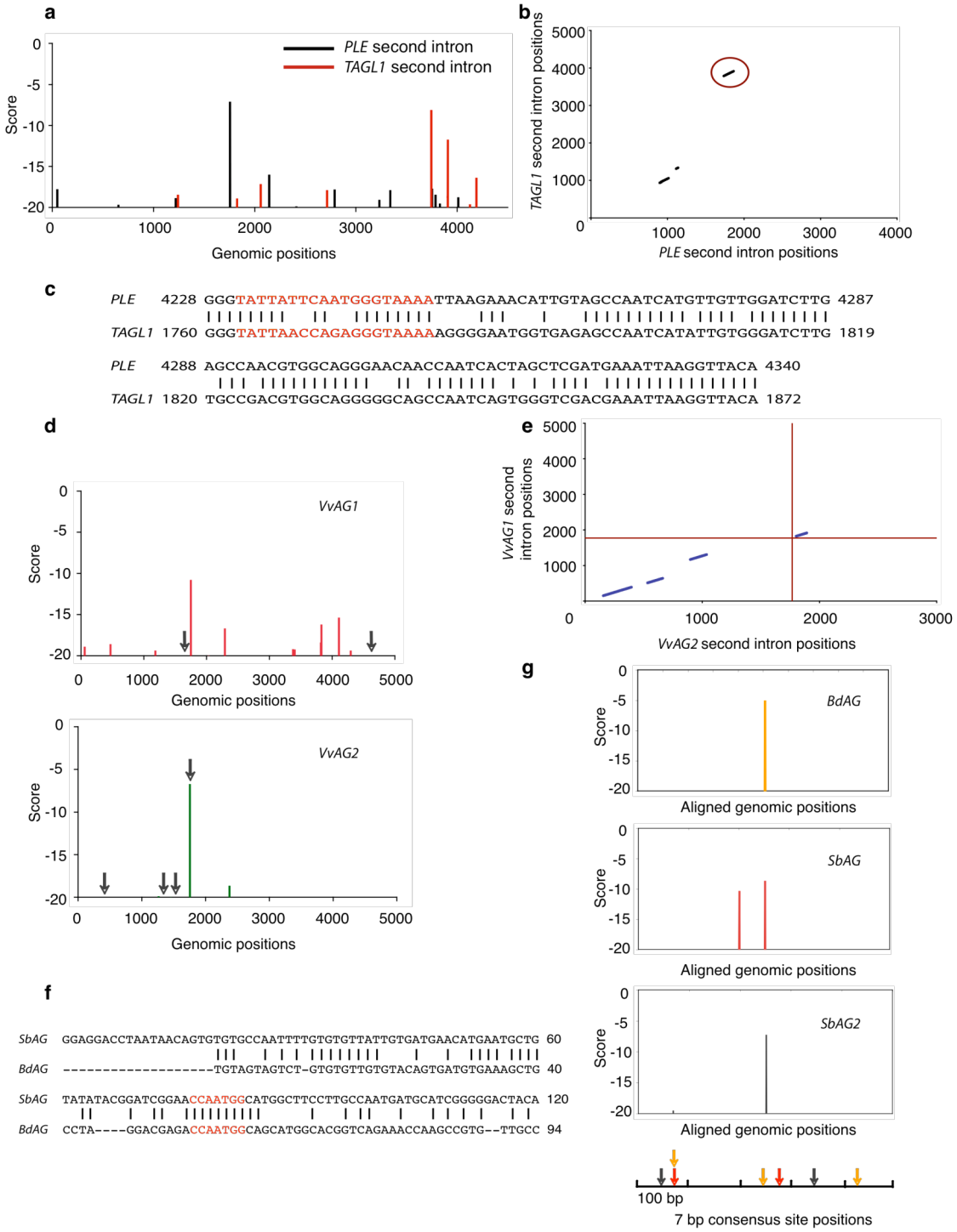
Supplementary Figure 2: LFY binding sites in the large intron of *AG* homologs.

(a-c) LFY binding sites in *PLENA* from *A.majus* and *TAGL1* from *S.lycopersicon*. (a) Scores distribution of LFY binding sites in *PLE* (in black) and *TAGL1* (in red) large introns computed with the SYM-T PSSM. High scores binding sites are detected at different positions in the two introns. (b) Dot matrix view (from the program YASS, available at http://bioinfo.lifl.fr/yass/yass.php, with an E-value of 0.1) of the alignment of *PLE* and *TAGL1* introns showing a short conserved region (circled in red). (c) Alignment of the short conserved region, with the highest score site colored in red. The sequence of this site is conserved but its position is different within the intron.

(d-e) LFY binding sites in the two *AG* paralogs (*VvAG1* and *VvAG2)* from *V.vinifera*. (d) Score distribution for LFY binding sites in the large intron of *VvAG1* and *VvAG2*. Arrows indicate the presence of the CCANTG[G/T] 7 bp consensus. Note the discrepancy between presence of high score sites and consensus sequence (e) Dot matrix view (E-value of 10^{-7}) for the alignment of *VvAG1* and *VvAG2* large introns. The highest score sites (their position is indicated by red lines) do not fall in conserved region.

(f-g) LFY binding sites in the second intron of *AG* in the two grasses, *B.distachyon* and *S.bicolor*. (f) Alignment showing a region (underlined by dots in g) of high sequence conservation between *SbAG* and *BdAG* and containing a consensus site but no high score predicted sites. (g) Score distribution for LFY binding sites in

a 450 bp region centered on the best predicted LFY binding site in the second intron of three *AG* genes. Colored arrows indicate the positions of 7 bp consensus sites.



Supplementary Table 1: Position specific scoring matrices (PSSM)

	ASY				SYM				SYM-T			
	A	**C**	**G**	**T**	**A**	**C**	**G**	**T**	**A**	**C**	**G**	**T**
1	-2.31	-0.78	-1.13	0	-1.84	-0.64	-1.18	0	-1.84	-0.64	-1.18	0
2	0	-3.62	-3.34	-1.31	0	-2.74	-2.67	-1.14	0	-2.74	-2.67	-1.14
3	-2.06	-3.50	-4.22	0	-1.68	-1.82	-3.02	0	-1.68	-1.82	-3.02	0
4	-3.45	-5.38	-4.51	0	-2.71	-3.37	-1.06	0	Lateral triplets 4-5-6			
5	0	-4.10	-2.57	-4.16	0	-4.15	-0.51	-3.49				
6	-2.24	0	-5.15	-1.68	-0.18	0	-4.51	-0.78				
7	-1.95	0	-4.66	-4.04	-2.52	0	-4.51	-4.13	-2.52	0	-4.51	-4.13
8	-6.33	0	-4.62	-3.37	-6.86	0	-5.35	-3.05	-6.86	0	-5.35	-3.05
9	0	-4.88	-2.15	-3.43	0	-4.57	-0.27	-2.89	Central triplets 9-10-11			
10	-0.47	-1.42	0	-1.61	-0.41	0	0	-0.41				
11	-3.52	0	-6.06	-1.16	-2.89	-0.27	-4.57	0				
12	-3.54	-6.32	0	-6.34	-3.05	-5.35	0	-6.86	-3.05	-5.35	0	-6.86
13	-4.52	-5.98	0	-3.70	-4.13	-6.00	0	-2.52	-4.13	-6.00	0	-2.52
14	-0.90	-4.10	-1.87	0	-0.78	-4.51	0	-0.18	Lateral triplets 14-15-16			
15	-3.26	0	-4.41	-0.77	-3.49	-0.51	-4.15	0				
16	-0.10	0	-2.57	-2.01	0	-1.06	-3.37	-2.71				
17	0	-2.36	-1.07	-1.41	0	-3.02	-1.82	-1.68	0	-3.02	-1.82	-1.68
18	-1.06	-2.30	-2.36	0	-1.14	-2.67	-2.74	0	-1.14	-2.67	-2.74	0
19	0	-1.29	-0.54	-1.55	0	-1.18	-0.64	-1.84	0	-1.18	-0.64	-1.84

Lateral triplets 4-5-6

		2nd position					
		A	C	G	T		
1st position	A	-5.8542	-5.8542	-3.1781	-5.8542	A	3rd position
		-3.7377	-4.7185	-5.8171	-5.8542	C	
		-5.8542	-5.8542	-5.8171	-5.8542	G	
		-5.8171	-5.8542	-5.1240	-5.8542	T	
	C	-5.8171	-5.8542	-4.2077	-5.1240	A	
		-5.1240	-5.8542	-5.8542	-5.8542	C	
		-5.8542	-5.8542	-5.8542	-5.8542	G	
		-5.8171	-5.8542	-5.8542	-5.8542	T	
	G	-5.1240	-5.8542	-0.6754	-4.0254	A	
		-3.8712	-5.8542	-5.1240	-5.8542	C	
		-5.8542	-5.8542	-5.8171	-5.8542	G	
		-5.8171	-5.8542	-3.4192	-5.8542	T	
	T	-1.7567	-5.8542	-2.4498	-4.4308	A	
		0	-5.8542	-2.7261	-5.8171	C	
		-5.8171	-5.8542	-5.8171	-5.8542	G	
		-1.1727	-5.8542	-1.9459	-5.8542	T	

Central triplets 9-10-11

1st position		2nd position A	2nd position C	2nd position G	2nd position T	3rd position
A		-3.9318	-5.0304	-3.6441	-3.6441	A
		-0.3670	-1.7346	0	-1.8116	C
		-5.0304	-5.0304	-5.0304	-5.0304	G
		-1.6982	-1.0986	-1.0986	-1.6982	T
C		-5.0304	-5.0304	-5.0304	-5.0304	A
		-5.0304	-3.9318	-5.0304	-5.0304	C
		-5.0304	-5.0304	-5.0304	-5.0304	G
		-5.0304	-5.0304	-5.0304	-5.0304	T
G		-3.2387	-5.0304	-3.2387	-5.0304	A
		-3.6441	-2.3224	-2.3224	-3.6441	C
		-5.0304	-5.0304	-3.9318	-5.0304	G
		-1.8116	0	-1.7346	-0.3670	T
T		-5.0304	-5.0304	-5.0304	-5.0304	A
		-5.0304	-3.2387	-5.0304	-3.2387	C
		-5.0304	-5.0304	-5.0304	-5.0304	G
		-3.6441	-3.6441	-5.0304	-3.9318	T

Lateral triplets 14-15-16

1st position		2nd position A	2nd position C	2nd position G	2nd position T	3rd position
A		-5.8542	-1.9459	-5.8542	-1.1727	A
		-5.8542	-3.4192	-5.8542	-5.8171	C
		-5.8542	-5.8542	-5.8542	-5.8171	G
		-5.8542	-5.1240	-5.8542	-5.8171	T
C		-5.8542	-5.8171	-5.8542	-5.8171	A
		-5.8542	-5.8171	-5.8542	-5.8542	C
		-5.8542	-5.8542	-5.8542	-5.8542	G
		-5.8542	-5.8171	-5.8542	-5.8542	T
G		-5.8171	-2.7261	-5.8542	0	A
		-5.8542	-5.1240	-5.8542	-3.8712	C
		-5.8542	-5.8542	-5.8542	-5.1240	G
		-5.8542	-5.8171	-4.7185	-3.7377	T
T		-4.4308	-2.4498	-5.8542	-1,7567	A
		-4.0254	-0.6754	-5.8542	-5.1240	C
		-5.1240	-4.2077	-5.8542	-5.8171	G
		-5.8542	-3.1781	-5.8542	-5.8542	T

Supplementary Table 2: List of the 1564 regions bound by LFY in ChIP-seq experiments (see material and methods for selection criteria). The order by rank product (orp_rank), the chromosome position of the region's 5' end and the fragment size are indicated.

Supplementary Table 3: **List of genes (upstream and downstream) adjacent to the 1564 bound regions** in ChIP-seq experiments (selection limits: -4kb and +3kb from gene model). The bound region rank (from Table

S2), the distance to the gene model (negative numbers indicate upstream position to gene model; positive numbers indicate downstream position; no number is given when bound region is inside the gene model), the GO annotations and the gene descriptions (from www.arabidopsis.org) are indicated. Data is duplicated when several peaks are detected for the same gene.

Supplementary Table 4: **Overlap between genes bound by LFY in ChIP-seq and genes regulated by LFY.** Genes from Table S3 were compared against microarray data analyzing wild-type and *lfy* mutant inflorescences upon floral transition [2] and 35S::LFY-GR plants induced with dexamethasone + cycloheximide [3]. The expression ratios, the peak information, the gene description and the score values (POcc and best PSSM score site) are given in columns.

Supplementary Table 5: Summary of functional and expression data on the *AG* subfamily. The identification cDNA number (collected in GenBank), expression data, functional data and bibliography are summarized for each gene of the *AG* subfamily from species with sufficient sequence and expression data. Early or late gene expression during reproductive development is represented by a color code (light blue: early expression, dark blue: late expression), and relative values of POcc for each gene are also represented by colours (orange: high POcc value, *i.e.* > 0.22, yellow: low POcc value), highlighting the correlation between early expression and high POcc values, likely reflecting a regulation by LFY.

GenBank	Gene name (clade)	Expression data	Functional data	Ref.
Legend	Early expression	Late expression	High POcc	Low POcc
		Arabidopsis thaliana		
X53579	*AGAMOUS (AG)*	Expressed in stamens and carpels, from primordia to mature organs.	Role in stamen and carpel identity, meristem determinate growth (**C function**) and late role in ovule development.	4,5
DQ446777	*SHATTERPROOF1 (SHP1)*	Expressed in stage 7 in gynoecium and ovule.	Role in fruit development and maturation. Minor role in stigma and style development, and ovule identity.	6-8
M55553	*SHATTERPROOF2 (SHP2)*	Expressed during carpel development (shortly after *AG*).		
AY087201	*SEEDSTICK (STK)*	Expressed in ovules and seeds.	Role in ovule and seed development (**D function**).	9
		Antirrhinum majus		
AB516405	*FARINELLI (FAR)*	Expressed in stamens and carpels, from primordia to mature organs.	The *far* mutant shows male sterility. *FAR*, together with *PLE*, represses the expression of B function genes in the fourth whorl, which is consistent with a **partial C function**.	10
AB516404	*PLENA (PLE)*	Expressed in stamens and carpels, from primordia to mature organs.	Role in stamen and carpel identity, and meristem determination **(C function)**	10
		Carica papaya		
EF645801	*CpPLENA*	Expressed in stamens and carpels of all 3 sex types flowers, from primordia to mature organs.		11

	EU141966	*CpSEEDSTICK*	Expressed in carpels of female and hermaphrodite flowers.		11
			Populus trichocarpa		
	AF052570	*PTAG1*	Expressed in the inner whorl of male and female flowers (giving either stamens or carpels).		12
	AF052571	*PTAG2*			
	XM_002327246	*PtMADS43*			13
	XM_002325495	*PtMADS51*			
			Cucumis sativus		
	X97801	*CUS1 / CAG2*	Expressed late in female flowers, in the pistil and around the ovules. Also expressed in fruit and embryo.		14,15
	AF035438	*CUM1 / CAG3*	Expressed in the third and fourth whorls of male and female flowers.		15
	AF035439	*CUM10 / CAG1*	Expressed in the third and fourth whorls of male and female mature flowers.		15
			Vitis vinifera		
	XM_002263030	*VvAG2*	Expressed in developing flowers, later in developing seeds and berries.	*35S::VvAG1* in Tobacco: nectaries appear at the base of sepals, sepals become carpelloid and petals become filament-like. **Consistent with a partial C-function.**	16
	AF265562	*VvAG1*			17
	AF373604	*VvAG3*	Expressed in the carpels of female flowers, later in berries.		17
			Prunus persica		
	FJ184275	*PpAG / PpFAR*	Expressed in the floral meristem, in the region of future stamen and carpel primordia. Later expressed in ovules, developing fruit and seeds.	*35S::PpAG* in Arabidopsis: sepals turn into carpels, petals develop some stamen features. **Consistent with C-function.**	18,19
	FJ188413	*PpPLENA*	Expressed in stamens and carpels, and throughout fruit development. *PpPLE* is more expressed in flower than *PpFAR*.	*35S::PpPLENA* in Tomato: flowers form carpel-like sepals and fused petals. **Consistent with a partial C-function**. Sepals become "fruit-like" in later developmental stages.	19,20
	EF602037	*PpSTK*	Expressed in the ovary, then in the fruit.		21
			Petunia hybrida		
	X72912	*pMADS3*	Expressed in stamens and carpels, from primordia to mature organs.	*35S::pMADS3* in Petunia; antheroid structures in place of petals. *pMADS3* silencing: homeotic transformation of stamens into petals, reversion of the floral meristem to an inflorescence meristem. **Consistent with C-function.**	22,23
			Solanum lycopersicon		
	L26295	*TAG1*	Expressed early in stamen primordia and in the region of future carpel primordia. Later, it remains expressed in stamens, ovary and ovules.	RNAi against TAG1: loss of flower determinacy, petaloid stamens. **Consistent with a partial C-function.**	24,25
	AY098735	*TAGL1*	Strongly expressed in stamens and carpel, and faintly in petals. Expressed during fruit development.	RNAi against *TAGL1*: no defects in stamen or carpel development, but defects in fruit ripening.	24-26
	AY098736	*TAGL11*	Expressed in mature carpels and fruit.		24

			Oryza sativa		
	L37528	*OsMADS3*	Expressed in the early stamen, carpel and ovule primordia.	T-DNA insertion into *OsMADS3*: homeotic transformation of stamens into lodicules, increase in the number of carpels, ectopic lodicules. **Consistent with a partial C-function (mostly for stamen development).**	27,28
	AB232157	*OsMADS58*	Expressed in stamens and carpels, from primordia to mature organs.	RNAi against *OsMADS58*: transformation of stamens into lodicules, strong defects in carpel development. **Consistent with a partial C-function (mostly for carpel development).**	28
	AF151693	*OsMADS13*	Expressed in ovules and seeds.	Transposon insertion into *OsMADS13*: homeotic transformation of ovules into carpels, loss of floral determinacy.	29
	AY177693	*OsMADS21*	Strong expression in the ovule and parts of the carpel. Slight expression in early anthers.	T-DNA insertion into *OsMADS21*: no phenotype. No additional phenotype for the double mutant *osmads13/osmads21* as compared to the *osmads13* single mutant.	29
			Zea mays		
	NM_001111851	*ZAG1*	Expressed in the ear, in carpel and stamen primordia.	Transposon insertion into *ZAG1*: in the ear, floral determinacy is affected and carpels are not fused. **Consistent with a partial C-function (subfunctionalisation for carpel development).**	30,31
	X81200	*ZMM2*	Mostly expressed in the tassel, slight expression in the ear.	**Partial C-function proposed (subfunctionalisation for stamen development)**	30
	NM_001111908	*ZAG2*	Expressed in the ear, in the carpel primordia, later remains expressed in the ovules.		31
	X81199	*ZMM1*			32

Materials and Methods

Systematic Evolution of Ligands by EXponential enrichment (SELEX)

Selection cycles

In vitro selection of aptamers was performed with fluorescent 81-mers and a recombinant version of the DNA binding domain of *A.thaliana* LEAFY protein (LFY DBD) produced and purified as previously described [33]

Initially, a random sequence library was synthesized by PCR amplification (98°C 1'30s followed by 20 cycles: 98 °C 10 s, 55 °C 25 s, 72 °C 15 s) with Phusion® DNA polymerase (Ozyme) using 81-mers (5'-TGGAGAAGAGGAGAGATCTAGC$(N)_{30}$CTCTAGATCTTGTTCTTCTTCGATTCCGG-3') as template with a fluorescent forward primer (SElex-F: TAMRA 5'-TGGAGAAGAGGAGAGATCTAG-3') and a non-labelled reverse primer (SElex-R: 5'-CCGGAATCGAAGAAGAACAA-3') (Sigma-Aldrich). The size of the PCR products was verified on 3% agarose gels stained with SYBR®Safe (Invitrogen) and dsDNA concentration was measured using SYBR®green (Invitrogen) and a microplate reader (Safire^2, TECAN) according to the manufacturer's instructions.

For each selection cycle, 200 nM LFY-C were mixed to 10 nM fluorescent dsDNA (81-mers) in 225 μl SElex buffer (20 mM Tris pH8, 250 mM NaCl, 2 mM $MgCl_2$, 5 mM TCEP, 10 μg/ml dIdC and 1% glycerol). After a 2-minute incubation on ice, 25 μl Ni Sepharose 6 fast flow (GE Healthcare), previously equilibrated in SElex buffer without TCEP, were added to the reaction mix to immobilize the DNA/protein complexes via the histidine

tag of the protein. After 30 min incubation at 4 °C on a rotating wheel, the reaction mix was loaded on an Ultrafree®-MC centrifugal filter unit (Millipore) and centrifuged for 1 min at 500 x g at 4 °C to eliminate the unbound DNA. Four washes were subsequently made by adding 300 µl of SElex buffer without dIdC on top of the filter unit followed by 1 min centrifugation at 500 x g at 4 °C. Finally, the Ni Sepharose was resuspended in 100 µl water and transferred into a clean tube. Selected 81-mers were amplified by PCR as described above, using 2 µl of the Ni-Sepharose solution as template. PCR products were quantified as described before and the selection cycle was repeated 7 times, using each time the newly synthesized fluorescent DNA as a library.
The whole selection process has been performed twice independently.

Enrichment evaluation

A gel shift assay [33]was used to estimate the enrichment for 81-mers with a good affinity for LFY-C through the successive selection cycles: 10 nM 81-mers library of each cycle was incubated with 200 nM LFY-C in 20 µl binding buffer (20 mM Tris–HCl pH 7.5, 150 mM NaCl, 1% glycerol, 0.25 mM EDTA, 2 mM MgCl2, 28 ng/ml fish sperm DNA (Roche) and 1 mM DTT). After 5 minutes of incubation on ice, the binding reactions were loaded onto native 6% polyacrylamide gels 0.5X TBE (45 mM Tris, 45 mM boric acid and 1 mM EDTA pH 8) and electrophoresed at 90 V for 90 minutes at 4 °C. Gels were scanned on a Typhoon 9400 scanner (Molecular Dynamics; excitation light 532 nm, emission filter 580 BP 30) and libraries that gave a visible shift were selected for sequencing (cycles 3 to 7).
Sequencing of the libraries corresponding to cycle 3 to 7 was performed using the 454 technology (Cogenics) and more than 2500 sequences were obtained.

These sequences yielded 494 unique sequences, which were aligned with the MEME software version 4.3.0 [34] http://meme.sdsc.edu/meme4_3_0/cgi-bin/meme.cgi) using the default parameters with either no constraints or with the symmetry imposed. Frequencies of individual nucleotides and/or triplets were derived from the alignments and used to calculate, at each position i of the motif, the weight (W) associated to each nucleotide (or triplet) n according to:
$W_{n,i} = \ln(f_{n,i}/f_{max,i})$ where $f_{n,i}$ is the frequency of nucleotide n at position i and $f_{max,i}$ is the maximal frequency observed at position i. When $f_{n,i} = 0$, a pseudocount value [35] of 0.001 was applied.

QUantitative Multiple Fluorescence Relative Affinity (QuMFRA) assay [36]

Single-stranded oligonucleotides were annealed to complementary oligonucleotides with an additional dGTP (5'end) in annealing buffer (10 mM Tris pH 7.5, 150 mM NaCl and 1 mM EDTA). The resulting double stranded DNA with a protruding G was fluorescently labeled by end-filling: 4 pmol of dsDNA was incubated with 1 unit of Klenow fragment polymerase (Ozyme and 8 pmol Cy5-dCTP (GE Healthcare) (dsDNA samples) or Cy3-dCTP (dsDNA reference) in 1X Klenow buffer during 2 h at 37 °C in the dark, followed by 10 min enzyme inactivation at 65 °C. Sequences used as references (*) or as samples are listed in the following table.

Max-M8*	GACTTGTTATTACCCAACGGTCCATAGATG
AP1Mat1*	GGCAATACATTACACAGTTGTCAATTGGTT
AP3Mat1*	GGTAGAATATTACCCCTAGGGTTTAAGAAG

W1	GACTTGTCATGACCCGACGGGTCATGGATGT
W2	GACTTGTCATGGCCCGACGGGCCATGGATGT
W3	GACTTGTTAAGAACCGACGGTTCTTAGATGT

AP1	GCTTGGGGAAGGACCAGTGGTCCGTACAAT
AG5	GTTTGGTCAATACCCAATGTTCATATTACA
AG4	GTTCGTAGAGTACCCATCGGATATATAAAA
AG2	GTTGGATTTATACCCAATGTGTTAATGGGT
AG1	GTTTAAATTTAATCCAATGGTTACAATTTT
SYMtest1	GGCAGGCTATTACCCAGTGGGTAATAGCCCT
SYMtest2	GTCTGGTGTTTGACCGTTGGTCCATTTCTGA
SYMtest3	GGATCTATATTAAACACCGGTTAATTAATAT
SYMtest4	GGTGAGGATTGGTCCAACGGTCAAACAGAAA
SYMtest5	GCTCGAGGACAAACCAACATGCAATAACCTT
SYMtest7	GCCTCACTATGGACTAGTAGACAATGGAGGA
S2	GACAGTCTATTGCTCGGCGGATAATAGAGGG
BObs1	GACTTAAGTTAGTACAGTGGAGACTTTCAT
BO2	GAGTATAATTAAAACAATGGTTCAGAACAT
S4	GTAACGAATTAGGCCAATGTTCTTATCCAT
AP1m5	GTTGGGGCAGGACCAGTGGTCCGGACAAT
AP1m6	GTTGGGGAATGACCAGTGGTCAGTACAAT
BEST	GCAGGCTATTGCCCGACGGGCAATATGAA
SYM11	GGCAGGCTGTCGCCCGACGGGCAATATGAA
SYM12	GGCAGGCTATCGCCCCACGGGCGATATGAA
SYM16	GGCAGGCTATTGCCCCCGGGGCAATATGAAT
SYM19	GGCAGGCTATTGCCCGGCGGGCGCTATGAAT
W0	GACTTGTTATGGCCCGACGGGCCATAGATGT

W8	GGCAGGCAGCGATCCTCAGGATCGCTGCCCT
W9	GGCAGGCACCGATCCACTGGATCGGTGCCCT
W10	GGCAGGCTACTGTCCACTGGACAGTAGCCCT
W11	GGCAGGCTATTGTCCACTGGACAATAGCCCT
W12	GGCAGGCTACTATCCACTGGATAGTAGCCCT
W13	GGCAGGCTACTGCCCAGTGGGCAGTAGCCCT
W14	GGCAGGCTAGTGTCCACTGGACACTAGCCCT
W15	GGCAGGCTATTATCCACTGGATAATAGCCCT
W16	GGCAGGCTACTGTCCCCGGGACAGTAGCCCT
W17	GGCAGGCAGCGATCTACTAGATCGCTGCCCT
W21	GTACATTTGTTATCTTTTGGTTTATCGTTTT
W22	GTGGACAGGAGATCCATCGGGCAGAAGAAAT
W23	GCAACTCCATTATCCTTTGAATATTACATAA
W25	GAGCACATATAAACCTTTGGACAAGGTTTTA
W32	GTTAACTTCATAATTAGTGGGTAATAATTAC
W33	GTTTTTATAATATCTTGTGTACCTTAAGAAT
SElex1	GCAGGCTATTACCCAGTGGATAATATGAAT
S1	GCAGGCTATTACCCAACGGATAAACTGAAT
SElex3	GCAGGCCTTTACCCACCGGTCCGTATGAAT
S3	GATGTTTTTTACTTTCTGTGTAAAAGATGG
AG3	GTCATCCATCCTCCATTGTTGTTAATGTCT
EGM01	GCAGGCTATTACCCAGCGGGTAATATGAAT

Binding reactions were performed in 20 μl binding buffer (20 mM Tris–HCl pH 7.5, 150 mM NaCl, 1% glycerol, 0.25 mM EDTA, 2 mM MgCl2, 28 ng/ml fish sperm DNA (Roche) and 1mM DTT) using 10 nM Cy3-dsDNA, 10 nM to 30 nM Cy5-dsDNA and 500 nM or 1 μM LFY DBD. After 10 minutes incubation on ice, the binding reactions were loaded onto native 6% polyacrylamide gels 0.5X TBE (45 mM Tris, 45 mM boric acid and 1 mM EDTA pH 8) and electrophoresed at 90V for 90 minutes at 4°C.

Gels were scanned on a Typhoon 9400 scanner (Molecular Dynamics; excitation light 532 nm (Cy3) and 633 nm (Cy5), emission filter 580 BP 30 (Cy3) and 670 BP 30 (Cy5)) and signals were quantified using ImageQuant software (Molecular Dynamics).

Gel analyses were performed according to [37]: for each gel lane, the fluorescent intensities of the bound and unbound fractions at both emission wavelengths were quantified by using volume analysis of the ImageQuant software (Molecular Dynamics). The resultant fluorescence intensities after background subtraction (FI_{cor}) at both wavelengths were used to calculate the relative dissociation constant (K_D^{Rel}) given by equation (1) :

$$K_D^{Rel} = \frac{\left[FI_{cor}(Bound)/FI_{cor}(Free)\right]_{reference}}{\left[FI_{cor}(Bound)/FI_{cor}(Free)\right]_{sample}} \quad (1)$$

The relative dissociation constant of each dsDNA was measured at least three times independently and the average value was used as K_D^{Rel} for comparison to the scores.

Experimental scores from figure 1C-E are defined as $\ln(K_D^{Rel} / K_D^{Rel}max)$, $K_D^{Rel}max$ corresponding to K_D^{Rel} of the dsDNA with the highest affinity for LFY DBD.

Crosslinking, chromatin isolation, and ChIP-seq

The entire experiment from seed sowing through deep sequencing was performed twice to produce independent biological replicates. ChIP-seq was performed with an antibody raised in rabbit (#4028) against the LFY C-terminal amino acids 223-424 (BioGenes, Berlin, Germany). We harvested 15 day-old seedlings grown under long day photoperiods at 23 °C on MS plates. Briefly, plants were harvested and fixed as described previously [38]. Frozen tissue was ground, filtered three times through Miracloth (Calibrochem), and washed as described previously thorough buffers M1, M2, and M3 [38]. Nuclear pellets were resuspended in sonic buffer as described (1 mM PEFA BLOC SC (Roche Diagnostics) was substituted for PMSF), split into technical duplicate samples, and sonicated with a Branson sonifier at continuous pulse (output level 3) for 8 rounds of 2 x 6 seconds and allowed to cool on ice between rounds. IP reactions were performed by incubating chromatin with 2.5 µl anti-LFY serum overnight at 4 °C as described [38]. The immunoprotein–chromatin complexes were captured by incubating with protein A-agarose beads (Santa Cruz Biotechnology), followed by consecutive washes in IP buffer and then elution as described [38]. Immunoprotein-DNA was then incubated consecutively in RNase A/T1 mix (Fermentas) and Proteinase K (Roche Diagnostics) as described after which DNA was purified using Minelute columns (Qiagen) [38]. ChIP samples were tested for enrichment by QPCR and then deep sequencing libraries were produced by standard Illumina protocols.

ChIP-seq analysis

Standard Illumina base calling software was used to base call the 40-42 nucleotide sequence reads. We used SHORE [39] as a platform for further analysis. The obtained reads were quality filtered and low quality bases at the 3' end were pruned as described [39]. GenomeMapper [40]was used for mapping to the TAIR9 genome, allowing for up to four mismatching nucleotides and no gaps. All ChIP-seq data are freely available from the GEO database (http://www.ncbi.nlm.nih.gov/; accession number: *to be provided*).

To proceed, the mapped data were subjected to a heuristic for removal of duplicate sequence reads which were assumed to be uninformative for the detection of enriched loci. A threshold was applied limiting the number of 5' ends mapping to the same position on the same strand. To retain the power to discriminate between multiple strongly enriched regions, the threshold for any particular position was varied depending on the coverage in close vicinity, such that the variance of the number of reads per position would roughly equal its mean in a 30 bp sliding window.

We further applied a two-step procedure to identify regions significantly enriched in the positive sample when compared to the control. First, potentially enriched regions were identified based on the positive samples only. These sites were then directly compared to the corresponding control sample regions to assess statistical significance.

For estimation of the depth of coverage for each position in the genome, all positive sample reads mapping to unique positions were extended in 3' direction to 130 bp, corresponding to half the experimentally observed approximate DNA fragment size, while discarding all other reads. To detect possible peak sites, a 2 kb wide sliding window was applied to the coverage graph in single base steps. In each step a p-value was assigned to the

coverage value at the central base using a one-sided Poisson test, with the distribution parameter set to the average coverage within the sliding window. Only positions with coverage greater than zero were included in the calculation of the average, assuming all other positions to be inaccessible to the experiment. Finally, any consecutive stretch of positions with p-value <0.05 and length >130 bp was retained as a potentially enriched site. To further reduce the number of regions to be considered, each was checked for unwarranted high average coverage in the control sample. A potential peak in the positive sample was discarded if the coverage mean in the control sample in the corresponding region was larger than the median average control coverage plus a tolerance of three standard deviations in all peak regions.

For assignment of final p-values to each candidate region, in each replicate a one-sided binomial test was applied to the number of reads mapping to the region in the positive sample, with the distribution parameter *N* set to the joint read count for the site for the positive and the corresponding control sample. To estimate the probability parameter for the test, from now on called *r*, we computed a scaling factor *s* for the control sample and the chromosome containing the considered region. The complete chromosome sequence was subdivided into 400 base pair bins, and for each bin the positive sample as well as the control sample read counts were recorded. Then *s* was chosen such that the median ChIP sample read count for all bins equaled the median control sample read count multiplied by *s*. From this the binomial test parameter *r* was calculated as *r=s/(s+1)*.
Finally, false discovery rates were obtained through the Benjamini-Hochberg correction method. To establish a ranking of peak regions across replicates, the rank product over the per-replicate fdr ranks was used.

Model for LFY-DNA binding

We use the Predicted occupancy (POcc) [41], defined as the expected number of bound transcription factor (TF) molecules for a given TF matrix of length W and a DNA sequence of length L, as given by equation (2), where $K_{A,s}$ is the relative equilibrium association constant for site s.

$$POcc = \sum_{s=1}^{L-W} p_s = \sum_{s=1}^{L-W} \frac{K_{A,s} \cdot [TF]}{1 + K_{A,s} \cdot [TF]} \quad (2)$$

$K_{A,s}$ can also be expressed as the inverse of the relative equilibrium dissociation constant ($1/K_{D,s}$), which can be calculated thanks to the correlation curve in Fig1, as given by equation (3):

$$score_s = -\ln(K_{D,s}) \cdot a + b \rightarrow K_{D,s} = e^{\frac{(b - score_s)}{a}} \quad (3)$$

Model	a	b
ASY	1.6349	-3.9647
SYM	1.8031	0.4133
SYM-T	2.5663	0.3598

We used [TF] equal to the K_D for the optimal site (score = 0), resulting in $p_{s\text{-}opt} = 0.5$ [41,42].
In the analyses presented in figures 2 and 4, we used a variant of POcc in which only binding sites with a score higher than a threshold t= -23 are considered [41].

Spearmann's rank correlation factor

POcc was calculated for all peaks in ChIP experiment (about 20 000). The correlation between ChIP and POcc ranking while using different PSSM was measured with the Spearman's rank correlation coefficient. This is a non-parametric measure of statistical dependence between the two variables ChIP and POcc. First, the n raw values (ChIPi and Pocci) are converted to ranks (x_i, y_i). Secondly, the differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated. The Spearman's rho (i.e. the correlation coefficient) is then given by equation (4):

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \quad (4)$$

Selection of bound peaks set and unbound genomic set

To define the bound DNA set, we used all peaks with fdr > 0.1 in both ChIP experiments, resulting in a set of 1564 peaks. The peaks were ranked using the rank product from both ChIP-seq replicates. The unbound set was generated by randomly selecting 1564 sequences from Arabidopsis genome that did not overlap with bound fragments and with the same size distribution as the bound set.

Bioinformatic data processing

Python programming language (www.python.org v2.6.4) was used to generate various scripts for automatic data processing: PSSM score calculation, POcc determination and ROC-AUC estimation.

Microarray data source

Microarray data was retrieved from GEO datasets (www.ncbi.nlm.nih.gov/geo): record GDS515 [3] and record GDS453 [2]. From GDS515, we use wt plants versus *lfy12* floral transition microarrays at 0, 3, 5 and 7 days. From GDS543, we use dexamethasone versus mock treatment and dexamethasone+cycloheximide versus cycloheximide treatment in 35S::LFY-GR plants to select for potential direct targets of LFY. We use a factor 2 as threshold for ratio selection.

The significance of the overlap between deregulated genes in the GDS543 microarray and the bound genes from the LEAFY ChIP-seq experiment was computed by using a Hypergeometric distribution.

Genomic sequences retrieval and analysis

The genomic sequences of *PLENA* (AY935269) and *FARINELLI* (AY935268) from *A.majus* and *AGAMOUS* (AT4G18960), *SHATTERPROOF1* (AT3G58780), *SHATTERPROOF2* (AT2G42830) and *SEEDSTICK* (AT4G09960) from *A.thaliana* were retrieved from GenBanK (http://www.ncbi.nlm.nih.gov). For all other species (except for *B.distachyon* and *S bicolor*) used in this study, the coding regions of previously identified members of the *AG* subfamily were retrieved from GenBank (http://www.ncbi.nlm.nih.gov) and used as BLAST queries against their respective species genome assembly to identify the corresponding genomic sequences. Coding sequences of members of the AG subfamily in *O. sativa* or *Z. mays* were blasted against the genomes of *S.bicolor* or *B distachyon* to find the orthologs in these species. Plant genomes assemblies of *A.thaliana*, *A.lyrata*, *P.trichocarpa*, *C.papaya*, *V.vinifera*, *M.esculenta*, *P.persica*, *C.sativus*, *B.distachyon*, *O.sativa*, *S.bicolor* and *Z.mays* were browsed and queried at Phytozome v5.0 (http://www.phytozome.net). The

S.lycopersion genome assembly (v1.50) was browsed and queried at the Sol genomic network (http://solgenomics.net). The POcc values (t=-23) were then calculated on the longest intron of each gene, which corresponds to the first or the second intron depending on the gene.

Supporting references

1. Workman, C.T. et al. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res*. **33**, W389-92 (2005).
2. Schmid, M. et al. Dissection of floral induction pathways using global expression analysis. *Development* **130**, 6001-12 (2003).
3. William, D.A. et al. Genomic identification of direct target genes of LEAFY. *Proc. Natl. Acad. Sci. U S A* **101**, 1775-80 (2004).
4. Bowman, J.L., Smyth, D.R. & Meyerowitz, E.M. Genes directing flower development in *Arabidopsis*. *Plant Cell* **1**, 37-52 (1989).
5. Yanofsky, M.F. et al. The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. *Nature* **346**, 35-9 (1990).
6. Colombo, M. et al. A new role for the *SHATTERPROOF* genes during Arabidopsis gynoecium development. *Dev. Biol.* **337**, 294-302 (2010).
7. Flanagan, C.A., Hu, Y. & Ma, H. Specific expression of the *AGL1* MADS-box gene suggests regulatory functions in Arabidopsis gynoecium and ovule development. *Plant J.* **10**, 343-53 (1996).
8. Liljegren, S.J. et al. *SHATTERPROOF* MADS-box genes control seed dispersal in Arabidopsis. *Nature* **404**, 766-70 (2000).
9. Rounsley, S.D., Ditta, G.S. & Yanofsky, M.F. Diverse roles for MADS box genes in Arabidopsis development. *Plant Cell* **7**, 1259-69 (1995).
10. Davies, B. et al. *PLENA* and *FARINELLI*: redundancy and regulatory interactions between two Antirrhinum MADS-box factors controlling flower development. *EMBO J.***18**, 4023-34 (1999).
11. Yu, Q., Steiger, D., Kramer, E.M., Moore, P.H. & Ming, R. Floral MADS-box genes in triecious papaya: characterization of *AG* and *AP1* subfamily genes revealed a sex-type-specific gene. *Tropical Plant Biol.* **1**, 97-107 (2008).
12. Brunner, A.M. et al. Structure and expression of duplicate *AGAMOUS* orthologues in poplar. *Plant Mol. Biol.* **44**, 619-34 (2000).
13. Leseberg, C.H., Li, A., Kang, H., Duvall, M. & Mao, L. Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* **378**, 84-94 (2006).
14. Filipecki, M.K., Sommer, H. & Malepszy, S. The MADS-box gene *CUS1* is expressed during cucumber somatic embryogenesis. *Plant Science* **125**, 63-74 (1997).
15. Perl-Treves, R., Kahana, A., Rosenman, N., Xiang, Y. & Silberstein, L. Expression of multiple *AGAMOUS*-like genes in male and female flowers of cucumber (Cucumis sativus L.). *Plant Cell Physiol.* **39**, 701-10 (1998).
16. Diaz-Riquelme, J., Lijavetzky, D., Martinez-Zapater, J.M. & Carmona, M.J. Genome-wide analysis of MIKC C-type MADS box genes in grapevine. *Plant Physiol.* **149**, 354-69 (2009).
17. Boss, P.K., Vivier, M., Matsumoto, S., Dry, I.B. & Thomas, M.R. A cDNA from grapevine (*Vitis vinifera* L.), which shows homology to *AGAMOUS* and *SHATTERPROOF*, is not only expressed in flowers but also throughout berry development. *Plant Mol Biol.* **45**, 541-53 (2001).
18. Martin, T. et al. *PpAG1*, a homolog of *AGAMOUS*, expressed in developing peach flowers and fruit. *Can. J. Bot.* **84**, 767-776 (2006).
19. Tadiello, A. et al. A *PLENA*-like gene of peach is involved in carpel formation and subsequent transformation into a fleshy fruit. *J. Exp. Bot.* **60**, 651-61 (2009).
20. Tani, E., Polidoros, A.N. & Tsaftaris, A.S. Characterization and expression analysis of *FRUITFULL*- and *SHATTERPROOF*-like genes from peach (Prunus persica) and their role in split-pit formation. *Tree Physiol.* **27**, 649-59 (2007).
21. Tani, E. et al. Characterization and expression analysis of *AGAMOUS*-like, *SEEDSTICK*-like, and *SEPALLATA*-like MADS-box genes in peach (Prunus persica) fruit. *Plant Physiol Biochem.* **47**, 690-700 (2009).
22. Tsuchimoto, S., van der Krol, A.R. & Chua, N.H. Ectopic expression of *pMADS3* in transgenic petunia phenocopies the petunia blind mutant. *Plant Cell* **5**, 843-53 (1993).
23. Kapoor, M. et al. Role of petunia pMADS3 in determination of floral organ and meristem identity, as revealed by its loss of function. *Plant J.* **32**, 115-27 (2002).
24. Hileman, L.C. et al. Molecular and phylogenetic analyses of the MADS-box gene family in tomato. *Mol. Biol. Evol.* **23**, 2245-58 (2006).

25. Pan, I.L., McQuinn, R., Giovannoni, J.J. & Irish, V.F. Functional diversification of *AGAMOUS* lineage genes in regulating tomato flower and fruit development. *J. Exp. Bot.* **61**, 1795-806 (2010).
26. Vrebalov, J. et al. Fleshy fruit expansion and ripening are regulated by the Tomato *SHATTERPROOF* gene *TAGL1*. *Plant Cell* **21**, 3041-62 (2009).
27. Kang, H.G., Jeon, J.S., Lee, S. & An, G. Identification of class B and class C floral organ identity genes from rice plants. *Plant Mol Biol* **38**, 1021-9 (1998).
28. Yamaguchi, T. & Hirano, H.Y. Function and diversification of MADS-box genes in rice. *ScientificWorldJournal* **6**, 1923-32 (2006).
29. Dreni, L. et al. The D-lineage MADS-box gene *OsMADS13* controls ovule identity in rice. *Plant J.* **52**, 690-9 (2007).
30. Mena, M. et al. Diversification of C-function activity in maize flower development. *Science* **274**, 1537-40 (1996).
31. Schmidt, R.J. et al. Identification and molecular characterization of *ZAG1*, the maize homolog of the Arabidopsis floral homeotic gene *AGAMOUS*. *Plant Cell* **5**, 729-37 (1993).
32. Theissen, G., Strater, T., Fischer, A. & Saedler, H. Structural characterization, chromosomal localization and phylogenetic evaluation of two pairs of *AGAMOUS*-like MADS-box genes from maize. *Gene* **156**, 155-66 (1995).
33. Hames, C. et al. Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *Embo J.* **27**, 2628-37 (2008).
34. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28-36 (1994).
35. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276-87 (2004).
36. Liu, J. & Stormo, G.D. Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic Acids Res.* **33**, e141 (2005).
37. Man, T.K. & Stormo, G.D. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **29**, 2471-8 (2001).
38. Gomez-Mena, C., de Folter, S., Costa, M.M., Angenent, G.C. & Sablowski, R. Transcriptional program controlled by the floral homeotic gene *AGAMOUS* during early organogenesis. *Development* **132**, 429-38 (2005).
39. Ossowski, S. et al. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024-33 (2008).
40. Schneeberger, K. et al. Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**, R98 (2009).
41. Roider, H.G., Kanhere, A., Manke, T. & Vingron, M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23**, 134-41 (2007).
42. Granek, J.A. & Clarke, N.D. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.* **6**, R87 (2005).

Résultats complémentaires

*L'étude réalisée avec la protéine d'*Arabidopsis *souligne que la pertinence de la prédiction génomique dépend de la qualité du modèle utilisé. Nous avons obtenu un bon modèle de liaison de LFY-C qui n'est cependant pas optimal pour expliquer les résultats obtenus* in vivo. *Un moyen potentiel d'amélioration serait d'utiliser un modèle de liaison établi pour la protéine LFY entière.*

Influence du domaine N-terminal

LFY possède un domaine N-terminal très conservé de fonction inconnue. Ce domaine pourrait influencer la liaison à l'ADN, soit directement en établissant des interactions supplémentaires avec les nucléotides extérieurs au motif de 19 paires de bases, soit indirectement en influant sur la structure tridimensionnelle du domaine C-terminal, modifiant ainsi sa spécificité de liaison. Nous avons mis en place deux stratégies successives pour évaluer la contribution de ce domaine à l'interaction LFY/ADN.

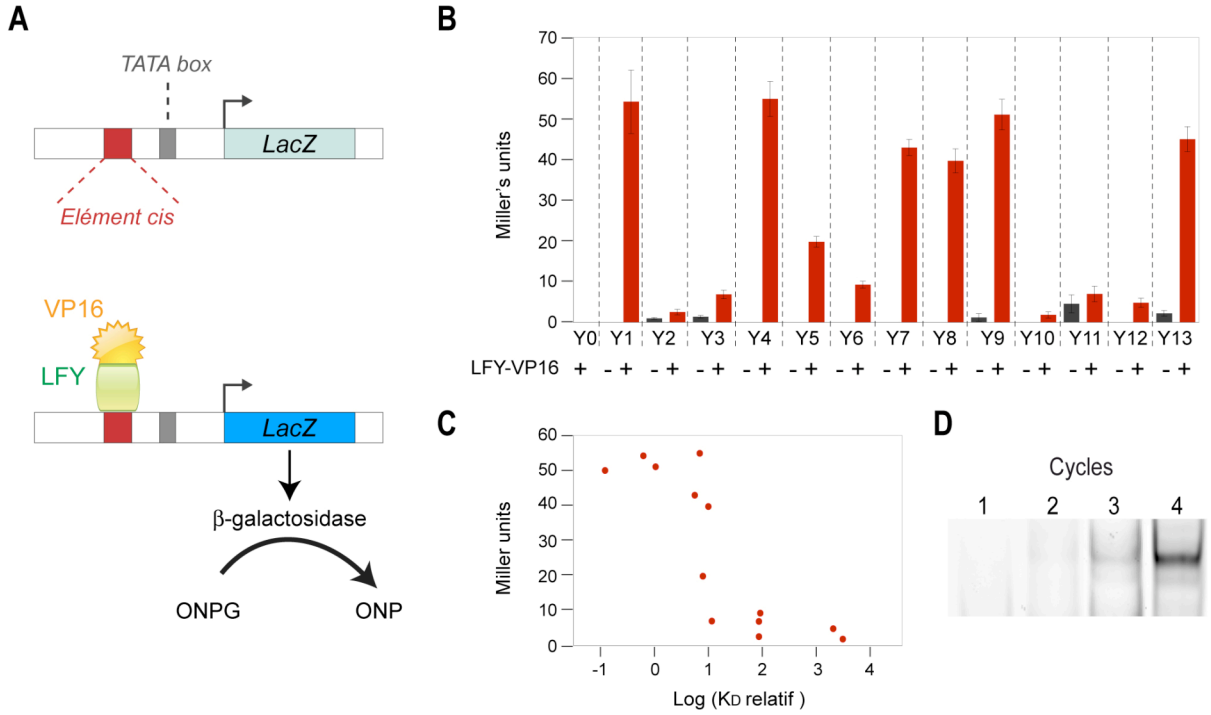
Relation entre affinité et activation de la transcription en système hétérologue

Ne disposant pas initialement d'une version recombinante de la protéine entière suffisamment pure et concentrée pour réaliser un Selex, nous avons effectué un test simple hybride en levure afin d'évaluer la spécificité de liaison de LFY. Le gène *lacZ* encodant l'enzyme β-galactosidase a été placé sous le contrôle d'un promoteur minimum auquel nous avons rajouté un élément *cis* dont l'affinité pour LFY-C a été préalablement mesurée. Treize constructions rapporteurs ont ainsi été construites avec des éléments *cis* dont le K_D^{App} varie entre 10 nM et 300μM (Fig.2.1A). Des levures ont ensuite été cotransformées avec une des constructions rapporteur et un vecteur exprimant la protéine LFY entière fusionnée à un domaine d'activation constitutif de la transcription (domaine activateur de transcription VP16 du virus de l'herpès ; (Triezenberg et al., 1988)) puisque LFY ne possède pas de pouvoir activateur intrinsèque. Après une période d'induction, l'activité de la β-galactosidase est mesurée et permet d'estimer directement le niveau de transcription du gène rapporteur (Fig.2.1B). Comme l'élément *cis* est la seule variable entre les différentes constructions, il est possible de tester l'existence d'une relation entre l'affinité de l'élément *cis* et sa capacité à stimuler l'expression du gène en aval. Les résultats obtenus montrent que plus l'affinité de LFY-C pour l'élément *cis* est élevée, plus la transcription est activée par la protéine entière ce qui suggère fortement que la présence du domaine N-terminal ne modifie pas la spécificité de liaison à l'ADN (Fig.2.1C). Par ailleurs, aucune activité β-gal n'est détectée lorsque le motif ADN testé présente une affinité en deçà d'une valeur seuil. En revanche, dès qu'un élément

cis d'affinité supérieure au seuil est présent dans le promoteur minimal, l'expression du gène rapporteur est enclenchée. Un élément *cis* peut donc conduire à une activation optimale de la transcription sans pour autant présenter une affinité maximale pour LFY-C. La valeur d'affinité seuil dépend probablement de la concentration de LFY.

Figure 2.1 | Etude de la liaison de LFY/ADN dans la levure *Saccharomyces cerevisiae*.

(A) Représentation schématique des constructions rapporteur utilisées et principe du test simple hybride effectué. Le niveau d'activation du gène *lacZ* est évalué par mesure de l'activité β-galactosidase (B) Mesures d'activité β-galactosidase dans des levures transformées avec une construction rapporteur d'affinité variable pour LFY-C (Y0 à Y13) et un plasmide exprimant (+) ou non (-) LFY-VP16. L'activité moyenne de 10 clones indépendants (exprimée en *Miller's Units*) est indiquée dans l'histogramme et (C) reportée en fonction du logarithme du K_D pour LFY-C de l'élément *cis* testé par rapport à celui d'*AP1bs1*. (D) Selex avec AtLFYmin40: 10nM d'ADN de chaque cycle de sélection sont incubés avec 200nM de protéine AtLFYmin40. L'intensité croissante du retard indique un enrichissement en aptamères de haute affinité pour AtLFYmin40.



Modélisation de la spécificité de liaison de la protéine entière LFY par Selex

En parallèle, des efforts ont été entrepris par Renaud Dumas, chercheur dans notre équipe, pour tenter d'obtenir une version complète de LFY utilisable *in vitro*. Le protocole mis au point a abouti à la purification d'AtLFYmin40, une version quasi-complète de la protéine (les 40 premiers acides aminés sont manquants) incluant le domaine N-terminal. Nous avons utilisé AtLFYmin40 pour réaliser un Selex. Un enrichissement en séquences de bonne affinité a été obtenu en 4 cycles de sélection (Fig.2.1D). Les ADN retenus après chacun des cycles

sont actuellement en cours d'analyse et permettront d'établir rapidement si la matrice de liaison de la protéine entière présente ou non des variations par rapport à celle de LFY-C.

L'histoire évolutive de LEAFY a précédé de 300 millions d'années la naissance de la première fleur. Nous avons donc cherché à adapter les méthodes développées précédemment pour étudier les propriétés de ce facteur chez les autres groupes de plantes terrestres. La dernière partie de ce chapitre présente les travaux réalisés dans ce sens et les nouvelles interrogations apportées par les résultats préliminaires.

Évolution de la spécificité de liaison à l'ADN

Sans modification drastique de sa séquence protéique, LFY a acquis a un moment donné une nouvelle fonction de régulateur du développement floral. Quel a été le moteur de cette évolution ? Plusieurs raisons, telles qu'une modification de son patron d'expression, l'acquisition de nouveaux cofacteurs ou encore un changement de gènes cibles, peuvent être invoquées. Dans un premier temps, nous avons choisi d'explorer la piste pour laquelle nous disposions d'outils d'étude, à savoir la spécificité de liaison à l'ADN des homologues de LFY.

Chaque partenaire de l'interaction facteurs de transcription/ADN peut influencer le devenir de cette relation. En effet, la séquence nucléotidique peut perdre ou gagner des éléments *cis* et ainsi échapper ou se soumettre au contrôle d'un facteur de transcription donné. Ce phénomène a participé activement à l'histoire évolutive de LFY et ses homologues chez les plantes à fleurs comme le révèle l'étude de la sous-famille *AG* (cf. Article 2). Réciproquement, la protéine peut voir ses préférences de liaison modifiées et ne plus reconnaître les mêmes motifs ADN. En suggérant un changement de spécificité graduel entre l'homologue de mousse et la protéine d'*Arabidopsis*, Maizel et al. ont proposé que ce second mécanisme s'applique également à LFY (Maizel et al., 2005). Nous avons donc abordé l'analyse des spécificités de liaison des protéines LFY-like en caractérisant tout d'abord PpLFY1, l'homologue de LFY chez la mousse *Physcomitrella patens*.

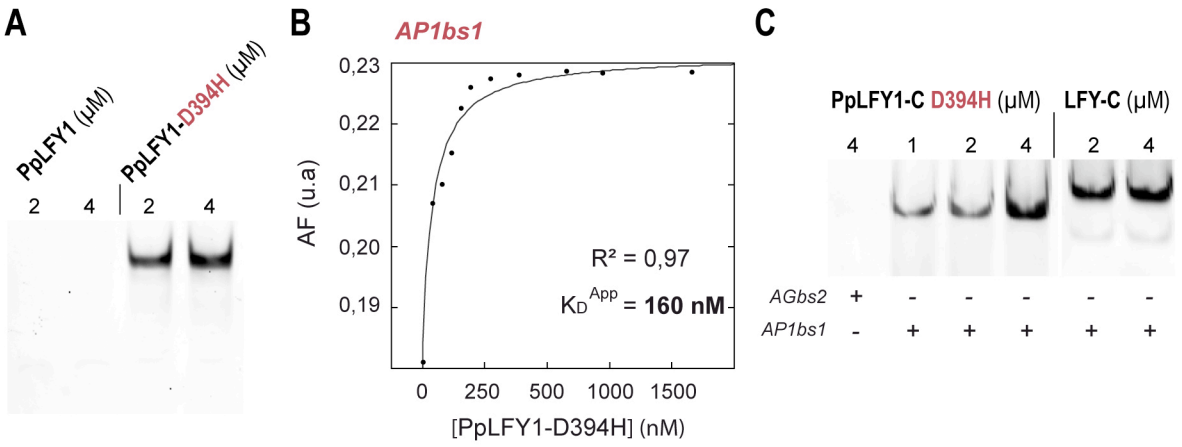
LEAFY chez les premières plantes terrestres : le cas des mousses

Nous avons produit plusieurs versions recombinantes (complète ou domaine C-terminal, sauvage ou mutantes) de PpLFY1 et testé leur capacité à lier deux éléments *cis* reconnus par LFY-C: *AP1bs1* et *AGbs2*. Les expériences de retard sur gel réalisées montrent que PpLFY1 est incapable de contacter *AP1bs1*, contrairement à la protéine d'*Arabidopsis* (Fig.2.2A). PpLFY1 possède un acide aspartique en position 394, à la place de l'histidine présente chez

les homologues d'autres espèces. Lorsque cette histidine est restaurée, PpLFY1-D394H fixe *AP1bs1* avec une affinité comparable à LFY (Fig.2.2B). À la différence de LFY-C, un seul complexe PpLFY1-C D394H/*AP1bs1* est détecté lorsque le domaine C-terminal de la protéine mutée de mousse est utilisé (Fig.2.2C) et la liaison à *AGbs2* n'est jamais observée. Ces premiers résultats sont cohérents avec les données publiées indiquant que la présence de l'histidine rétablit une activité partielle de la protéine (Maizel et al., 2005) mais indiquent également que cet acide aminé n'explique pas à lui seul les différences de comportement observées.

Figure 2.2 | Caractérisation de l'interaction PpLFY1/ADN.

(A) Retards sur gel avec 10nM *AP1bs1* et deux concentrations différentes de PpLFY1 ou PpLFY1 D394H. (B) Mesure de l'affinité apparente de PpLFY1 D394H pour *AP1bs1* par anisotropie de fluorescence. (C) Retards sur gel avec 10nM *AGbs2* et 4µM PpLFY1-C D394H ou 10nM *AP1bs1* et une concentration croissante de PpLFY-C D394H ou LFY-C.

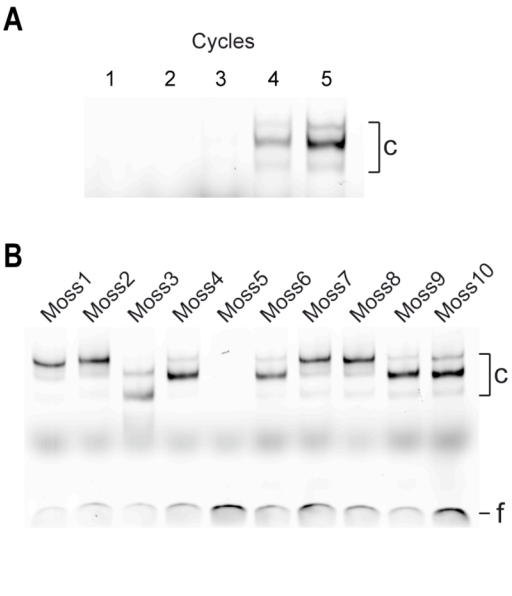


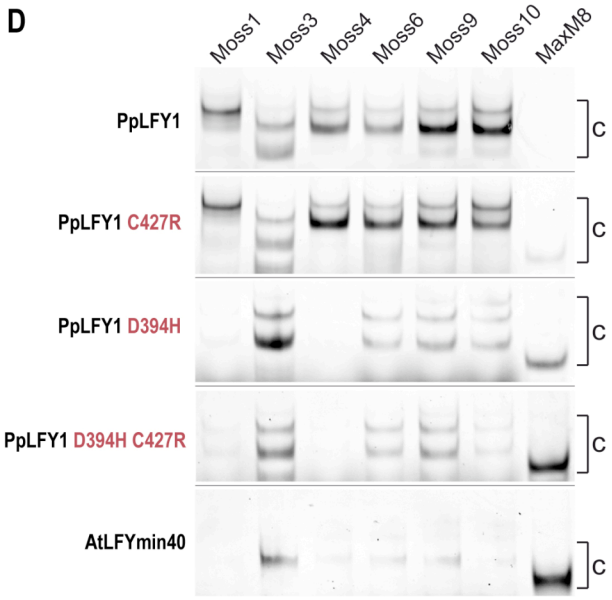
PpLFY1 présente un degré de conservation très élevé avec AtLFY, notamment au niveau de l'interface avec l'ADN, suggérant que cet homologue agit également comme un facteur de transcription. Dès lors, deux scénarios sont possibles : soit la protéine de mousse présente une spécificité de liaison très différente de son homologue angiosperme, soit PpLFY1 possède les mêmes préférences que LFY mais il existe un facteur *in vivo* (*i.e.* un partenaire) qui lui permet de contacter l'ADN du génome. Nous avons appliqué le protocole de Selex à la protéine PpLFY1 et obtenu un retard visible après 4 tours d'enrichissement, ce qui démontre qu'il existe des séquences spécifiquement reconnues par l'homologue de mousse seul (Fig.2.3A). Afin de déterminer la proportion de séquences du mélange spécifiquement reconnus par PpLFY1, nous avons cloné les oligonucléotides du cycle 5 puis amplifié séparément les aptamères pour tester leur liaison à PpLFY1. Parmi les 10 clones testés, neuf interagissent efficacement avec l'homologue de mousse (Fig.2.3B). Après séquençage, il est apparu que

ces 9 clones correspondaient à 6 séquences uniques, nommées *Moss1*, *Moss3*, *Moss4*, *Moss6*, *Moss9* et *Moss10* (Fig.2.3C), qui ne présentent aucune similarité apparente avec le motif préféré par la protéine d'*Arabidopsis*. AtLFYmin40 s'avère effectivement incapable de lier la majorité de ces séquences (Fig.2.3D). Seul un retard de faible intensité est observé avec la séquence *Moss3* ce qui est cohérent avec les prédictions de la matrice SYM-T (Fig.2.3E).

Figure 2.3 | Spécificité de liaison à l'ADN de PpLFY1.

(A) Selex avec PpLFY1 : 10nM d'ADN de chaque cycle de sélection sont incubés avec 200nM de protéine PpLFY1. L'intensité croissante du retard indique un enrichissement en séquences de bonne affinité pour PpLFY1. (B) Test de l'interaction entre PpLFY1 et les séquences individuelles Moss1 à Moss10 dont les séquences ne possèdent pas le motif reconnu par AtLFY (C) Séquences Moss des aptamères clonés après 5 cycles de selex. (D) Spécificité des différentes protéines PpLFY1 sauvage et mutantes comparées à AtLFYmin40 : 10nM d'ADN correspondant à chaque séquence testée sont incubés en présence de 200nM de protéine. Conformément aux prédictions de la matrice SYM-T (E), les séquences Moss sont peu ou pas reconnues par AtLFYmin40. ADN libre (f) et complexes ADN/protéine (c) sont indiqués.





C

TACCGCGGGGCGCCGTCTTACAGACGCAGG	Moss1,2,7,8
CACGGGCGACCAGCGGACGGTGAGCACCAA	Moss3
GTCCGGTAGGCGCCGTCCGGTTGGCAACAT	Moss4
GACCGGAGGGCGGCGTCCGGTTGGCCAATG	Moss6
GGTCGGCGTCCACCGGTCGGGTCGTCAGCA	Moss9
CAGCGGACGCCGCCCAGCAGTCGGGAACCA	Moss10

E

Oligonucléotide	Score SYM-T
Moss1	-30,1
Moss3	-19,7
Moss4	-24,9
Moss6	-23,8
Moss9	-23,1
Moss10	-23,6
AP1bs1	-7,33
MaxM8	-1,07

En plus de l'histidine remplacée par un acide aspartique en position 394, PpLFY1 possède une cystéine en position 427, à la place d'une arginine (Maizel et al., 2005). Le rétablissement de l'un ou l'autre de ces acides aminés permet la reconnaissance du motif préféré par LFY-C (oligonucléotide MaxM8). Cependant, les protéines PpLFY1-C427R et PpLFY1-D394H ont

conservé leur capacité à interagir avec toutes ou une sous-partie des séquences *Moss* respectivement (Fig.2.3D) et les séquences *Moss3, 6, 9* et *10* sont toujours reconnues lorsque qu'une protéine combinant les deux mutations, PpLFY1 D394H C427R, est utilisée (construction réalisée par E.Thévenon, technicien de notre équipe). Ainsi, même si les résidus histidine et arginine identifiés participent à l'établissement de la spécificité de liaison de la protéine d'*Arabidopsis*, leur rétablissement chez PpLFY1 D394H C427R ne suffit pas à mimer parfaitement le comportement de LFY.
Bien que PpLFY1 possède tous les acides aminés impliqués dans les contacts spécifiques entre LFY-C et l'ADN, les spécificités de liaison des deux homologues semblent très dissemblables. Le gradient de liaison observé suggère cependant qu'il est possible de passer progressivement d'un comportement à un autre en modifiant quelques résidus clés.

Afin de représenter de manière exhaustive les préférences de PpLFY1, nous avons réalisé un séquençage massif des ADNs issus du Selex (cf. fin de ce chapitre) ce qui devrait permettre de construire un modèle prédictif de liaison à l'ADN de PpLFY1. Ce modèle sera ensuite utilisé pour explorer le rôle de LFY chez cette espèce : l'inactivation de l'activité PpLFY engendre une létalité embryonnaire chez *Physcomitrella patens*, il n'est donc pas possible d'accéder directement aux gènes dérégulés lorsque PpLFY est absent et d'identifier ainsi ses cibles. L'utilisation d'un modèle de liaison PpLFY-spécifique pour repérer les sites de haute affinité au sein du génome séquencé de *P. patens* permettra de donner des pistes de travail en suggérant de potentiels gènes cibles.

Retour aux origines de LFY

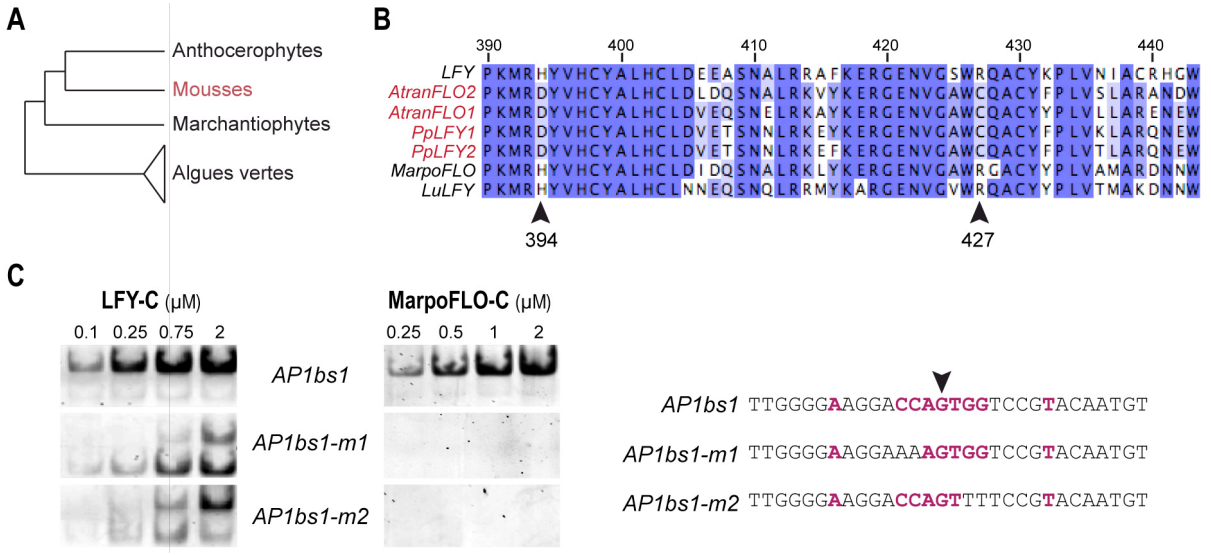
PpLFY1 est une protéine de mousse, c'est-à-dire une espèce appartenant aux bryophytes, le premier groupe de plantes terrestres. Pour autant, le comportement de PpLFY1 reflète-il vraiment les propriétés ancestrales du facteur?

Trois phyla (marchantiophytes, mousses et anthocerophytes) sont rassemblés au sein des bryophytes (Fig.2.4A). En plus de PpLFY1 et PpLFY2, trois autres homologues de LFY chez les bryophytes sont répertoriés dans les bases de données. AtranFLO1 et son paralogue AtranFLO2, homologues de LFY chez une autre espèce de mousse (*Atrichum angustatum*), possèdent également les résidus D394 et C427. En revanche, MarpoFLO, l'homologue de LFY de *Marchantia polymorpha*, une marchantiophyte, possède l'histidine et l'arginine retrouvées chez les plantes à fleurs (Fig.2.4B). Afin de vérifier qu'il ne s'agissait pas d'une

particularité de *M. polymorpha*, nous avons cloné l'homologue de *LFY* chez *Lunularia cruciata,* une seconde espèce de marchantiophyte révelant ainsi la présence des acides aminés de type 'Arabidopsis' chez cette deuxième espèce (Fig.2.4B). Les marchantiophytes ayant divergé avant le clade des mousses, les acides aminés 'mousses' relèveraient plus d'un état dérivé que de l'état ancestral.

Figure 2.4 | Diversité des comportement des homologues de LFY chez les bryophytes.

(A) Représentation simplifiée des 3 lignées de bryophytes. (B) Les acides aminés D394 et C427 sont retrouvés chez les espèces de mousse (rouge) alors que les deux espèces de marchantiophytes possèdent à ces positions les mêmes acides aminés qu'AtLFY (H394 et R427). (C) Capacité de LFY-C ou MarpoFLO-C à reconnaître *AP1bs1* et ses versions mutées *AP1bs1-m1* et *AP1bs1-m2*. Dans chaque cas, 10nM d'ADN fluorescent sont incubés en présence d'une concentration croissante de protéine. Les bases principales du motif reconnu par LFY-C sont indiquées en rouge et le centre du motif est indiqué par une flèche.



Nous avons montré que H394 et C427 ne garantissaient pas une spécificité de liaison similaire à la protéine d'*Arabidopsis*. Afin d'évaluer les capacités de liaison des protéines de marchantiophytes, nous avons produit une version recombinante du domaine C-terminal de MarpoFLO. Cette protéine s'est révélée parfaitement capable d'interagir avec *AP1bs1*. En revanche, aucune interaction ADN/protéine n'est observée en présence des motifs dégénérés *AP1bs1-m1* et *AP1bs1-m2* toujours reconnus par LFY-C (Fig.2.4C). Ces résultats suggèrent que MarpoFLO pourrait présenter une spécificité intermédiaire entre PpLFY1 et LFY.

Il existe donc une diversité de comportements des protéines LFY vis-à-vis de l'ADN chez les premières plantes terrestres. La question est désormais de déterminer dans quelle mesure cela témoigne d'une tendance évolutive générale.

Reconstituer 400 millions d'années d'évolution

Afin d'établir le chemin évolutif emprunté par la spécificité de liaison des homologues de LFY, nous avons réalisé un Selex pour plusieurs espèces emblématiques des grands groupes de plantes terrestres (Table 2.5). Pour cette étude à grande échelle, nous avons transformé le protocole utilisé pour la protéine LFY-C en méthode haut débit. Tout d'abord, nous avons amélioré la performance de l'étape de sélection (cf. Matériel et Méthodes) : désormais, un enrichissement maximal est obtenu en 3 à 4 tours contre 7 initialement (Table 2.5). Nous avons aussi mis au point un protocole de séquençage innovant : afin de séquencer en une seule fois l'ensemble des Selex réalisés, nous avons ajouté un code barre de 6 nucléotides aux différents mélanges d'aptamères. Chaque code barre correspond à un cycle de Selex donné ce qui permet de restituer à chaque homologue de LFY les séquences qu'il a sélectionnées. Enfin, nous avons ajouté aux oligonucléotides codés les amorces nécessaires au séquençage par Illumina (cf. Matériels et Méthodes). En générant plusieurs millions de séquences, la technologie Illumina améliore l'estimation des fréquences de chaque nucléotide (et triplets) à chaque position, pour révéler des différences même subtiles entre homologues proches.
Le séquençage Illumina de nos échantillons, effectué par une équipe de collaborateurs (*Max Planck Institute*, Tübingen, Allemagne) a généré plus de 4 millions de séquences. Marie Monniaux, étudiante en thèse dans notre équipe, a trié ces séquences d'après leur code barre et éliminé la redondance présente dans chacun des groupes. Entre 3000 et 150000 séquences uniques ont été obtenues pour chaque homologue (Table 2.5).

Ces résultats valident la stratégie que nous avons conçue puisque plusieurs millions de séquences ont été obtenues et que le code barre a permis une détermination rapide des différents sets. Au cours des prochaines semaines, nous construirons le modèle de liaison de chaque homologue. Un Selex sera également réalisé sur des homologues de ptéridophytes (fougères *sensus lato*) afin de compléter l'analyse. Une comparaison des matrices générées permettra finalement de préciser la nature des changements de spécificité et le moment où ces changements sont survenus. Le modèle de liaison peut être utilisé comme outil de lecture d'un génome (cf. Article 4), nous nous sommes donc attachés à travailler sur des espèces dont le génome est séquencé ou en cours de séquençage afin que les matrices construites soient utiles par la suite pour explorer le rôle de LFY chez ces espèces.

Table 2.5 | Evolution de la spécificité de liaison de LFY : Selex à grande échelle.

Pour chaque homologue de LFY pour lequel la méthode Selex a été utilisée, le numéro du cycle séquencé ainsi que le nombre de séquences totales et séquences uniques (après élimination des séquences redondantes) sont indiqués. L'espèce et le groupe de plante auxquels correspond la protéine sont rappelés.

Groupe	Espèce	Protéine	Cycle	Nombre de séquences	
				Total	Unique
BRYOPHYTES	*Physcomitrella patens*	PpLFY1	3	138443	131355
			4	98740	51401
			5	103539	8317
		PpLFY1 D394H	4	55367	8516
GYMNOSPERMES	*Ginkgo biloba*	GinLFYmin40	3	95498	15648
			4	119926	7151
		GinLFY-C	4	102026	9637
	Welwitschia mirabilis	WelLFYmin58	3	121517	7767
			4	111400	4646
		WelLFY-C	4	121323	8544
		WelNLYmin56	2	179563	92636
			3	55790	7472
			2bis	24348	10987
			3bis	44140	8005
		WelNLY-C	4	85069	57501
ANGIOSPERMES	*Amborella trichopoda*	AmboLFYmin40	3	151738	97091
			4	78204	34343
		AmboLFY-C	4	258282	28006
	Oryza sativa	RFLmin40	2	117819	81887
			3	120449	12154
		RFL-C	3	167188	14460
	Vitis vinifera	VFLmin40	2	89930	42362
			3	36107	3005
	Arabidopsis thaliana	AtLFYmin40	1	153491	153346
			2	141680	135335
			3	152259	31824
			4	146783	11063
		LFY-C	1	189261	189017
			2	73201	71481
			3	137800	56342

En plus d'isoler AmboLFY, l'homologue de LFY chez *Amborella trichopoda*, nous avons également cloné et séquencé le 2nd intron (6kb) de l'homologue d'*AGAMOUS* chez cette espèce. Nous utiliserons la matrice d'AmboLFY pour calculer la probabilité d'occupation de cet intron par AmboLFY et tester si le lien LFY/'fonction C' est déjà établi chez l'espèce sœur de toutes les autres plantes à fleurs.

Conclusion

La dynamique des éléments *cis* des gènes cibles de LFY et la modification des propriétés de liaison de ce facteur de transcription semblent tous deux participer à l'évolution de LFY. Il reste maintenant à déterminer dans quelle mesure ces mécanismes ont induit un changement de fonction de ce facteur et quelles ont pu être les conséquences à l'échelle morphologique. Ces aspects sont abordés dans le dernier chapitre.

Le dialogue entre ADN et facteur de transcription est au cœur des phénomènes biologiques, en particulier des processus développementaux. Nous avons construit un modèle biophysique expliquant globalement les sites du génome auxquels se lie LFY. Or, seul un sous-ensemble des sites reconnus par un facteur de transcription conduit à une régulation (Wasserman and Sandelin, 2004). L'étude de la sous-famille *AG* indique que la conservation entre espèces d'un site de liaison de haute affinité est un critère utilisable pour repérer les motifs fonctionnels (cf. Article 2). Le challenge des études à venir sera de comprendre l'ensemble des règles autorisant l'identification des éléments modulant l'expression des gènes parmi les sites potentiellement liés.

Les particularités de LFY (absence de famille multigénique, domaine de liaison à l'ADN étendu) font de ce facteur un candidat idéal pour aborder ces questions expérimentalement.

CHAPITRE 3

La résonance plasmonique de surface, un nouvel outil pour l'étude des interactions entre facteur de transcription et promoteur

Le chapitre précédent a montré qu'un modèle biophysique pouvait être utilisé pour prédire la capacité d'un facteur de transcription à reconnaître efficacement une séquence ADN donnée. Dans de nombreux cas cependant, on ignore tout des éléments cis *reconnus par la protéine d'intérêt. Chez les espèces non-modèles dont le génome n'a pas été séquencé, les séquences régulatrices potentielles sont elles aussi inconnues. En parallèle du Selex, nous avons donc cherché à développer une méthode permettant de tester rapidement et sans connaissance préalable sur la séquence d'ADN et la protéine étudiée, la capacité d'un facteur de transcription à se fixer efficacement à un promoteur entier.*

Introduction

Une fois que les sites reconnus par un facteur de transcription sont identifiés, il devient nécessaire de déterminer comment ces éléments *cis* travaillent ensemble pour influencer l'expression génique. Des progrès ont été réalisés récemment pour comprendre comment les séquences génomiques sont converties en réponse transcriptionnelle (Segal and Widom, 2009). Plusieurs modèles thermodynamiques en particulier ont démontré qu'il était possible d'expliquer la majorité des variations d'expression observées en utilisant les constantes d'équilibre des interactions entre facteur de transcription et promoteurs entiers (Segal et al., 2008; Gertz et al., 2009).

Ces paramètres restent cependant difficiles à modéliser car la façon dont un facteur de transcription contacte un promoteur complet résulte souvent d'une intégration complexe. En effet, il a été montré que différents arrangements (position, orientation) d'un ensemble donné d'éléments *cis* pouvaient conduire à une fixation variable de la protéine. De même, les phénomènes de 'looping' de l'ADN permettent de rapprocher des motifs éloignés sans qu'il soit possible de prédire à l'avance quels motifs vont être co-recrutés et à quelle condition. Enfin, une coopérativité entre sites est fréquemment observée, autorisant une liaison efficace là où aucun site individuel de haute affinité n'est détecté. Le contexte promoteur joue donc un rôle-clé pour autoriser ou non la fixation du facteur de transcription. Or, ce 'code promoteur' est en grande partie incompris en raison du peu de moyens expérimentaux mis à disposition

pour l'aborder : les approches dites 'classiques' (*i.e.*, retard sur gel, *footprinting*) sont limitées par la taille de l'ADN utilisé (quelques dizaines voire centaines de paires de bases) et il n'existe pas de méthode permettant d'étudier directement *in vitro* l'interaction entre une protéine et le promoteur entier d'un gène (plusieurs kilobases).

Choix de la technique et mise au point du protocole

Pour développer un protocole adapté à l'utilisation de longs fragments d'ADN, nous avons recherché une technique rapide, permettant une analyse quantitative de l'interaction étudiée même lorsque les caractéristiques du facteur de transcription d'intérêt ne sont pas établies ou lorsque l'enchaînement des nucléotides de la séquence utilisée est inconnu.

La résonance plasmonique de surface ou SPR est une méthode biophysique largement répandue pour examiner comment deux partenaires s'assemblent en complexes puis se dissocient: cette technique a été utilisée avec succès sur une variété de molécules, aussi bien pour caractériser les interactions protéine-protéine que pour évaluer la capacité d'un anticorps à reconnaître un épitope à la surface de cellules entières (Quinn et al., 2000; McDonnell, 2001). En pratique, la SPR repose sur l'utilisation d'un senseur optique capable de détecter des évènements de liaison se produisant à proximité d'une surface métallique en mesurant les variations de l'indice local de réfraction (Majka and Speck, 2007). Cette méthode très sensible utilise de faible quantité d'échantillon et ne nécessite aucun marquage (fluorescent ou radioactif) des molécules d'intérêt. Enfin, comme les mesures sont réalisées en temps réel, il est possible d'accéder directement aux paramètres cinétiques et thermodynamiques de l'interaction étudiée.

Plusieurs études ont employé la SPR pour analyser le comportement d'un facteur de transcription vis-à-vis d'un motif nucléotidique court (oligonucléotide) et nous avons cherché à savoir si l'utilisation du phénomène de résonance plasmonique de surface pouvait s'avérer possible sur de longues séquences ADN. En pratique, un ADN cible couplé à la biotine est fixé sur une puce constituée d'une lame de verre recouverte d'un film d'or et d'une surface dextran streptavidine. Cette puce est placée dans une cellule à flux permettant l'injection en solution du facteur de transcription étudié. La fixation éventuelle de la protéine à l'ADN provoque une modification de l'indice de réfraction local à l'interface métal/solution. Cette variation est instantanément détectée *via* un système optique (microréfractomètre de précision) qui permet de suivre en temps réel la formation des complexes ADN/protéine puis leur dissociation lorsque le facteur de transcription cesse d'être injecté.

Afin de tester la faisabilité de la méthode et de mettre au point sa mise en oeuvre expérimentale, nous avons tout d'abord réalisé une série de tests préliminaires en utilisant le second intron du gène homéotique *AGAMOUS* et une version recombinante de la protéine LFY-C d'*Arabidopsis thaliana*. Il est établi de manière solide que LFY induit l'expression d'*AG* en se liant spécifiquement à plusieurs éléments *cis* (au moins 3) présents dans le second intron du gène (Busch et al., 1999; Hong et al., 2003). De plus, cet intron de 3 kb est particulièrement long ce qui en fait un candidat de premier choix pour développer le protocole. La démarche que nous avons retenue consiste dans un premier temps à amplifier l'ADN d'intérêt par PCR en utilisant une paire d'amorces dont l'un des oligonucléotides est couplé à une biotine. L'ADN double brin biotinylé ainsi obtenu est immobilisé sur la puce grâce à l'interaction très forte qui s'établit entre biotine et streptavidine. La réponse SPR est ensuite enregistrée pour une gamme de concentration de LFY-C. Les premiers tests réalisés nous ont permis d'établir que la présence d'un compétiteur ADN (ici ADN de poisson) était nécessaire afin de prévenir toute interaction aspécifique entre LFY-C et la molécule d'ADN.
Nos résultats démontrent pour la première fois qu'il est possible d'étudier directement par SPR l'interaction entre facteur de transcription et un long fragment d'ADN. La méthode développée est quantitative puisque l'obtention d'un K_D apparent rend compte de la spécificité de l'interaction étudiée et permet de distinguer cibles (interaction spécifique) et non-cibles (interaction non-spécifique). Il s'agit également d'une approche expérimentale sensible qui permet de disséquer les différents constituants d'un promoteur d'intérêt. En effet, nous avons employé des versions mutées du second intron d'*AG* pour montrer que la perte d'un ou plusieurs éléments *cis* affectait le K_D apparent de façon mesurable et reproductible. Enfin, nous avons éprouvé avec succès cette méthode sur un facteur de transcription humain, HSF1, et le promoteur de son gène cible *hsp70* ce qui suggère que notre protocole est potentiellement applicable à n'importe quel autre couple facteur de transcription/région régulatrice de la transcription.

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2009) **59**, 851–858 doi: 10.1111/j.1365-313X.2009.03903.x

TECHNICAL ADVANCE

The analysis of entire gene promoters by surface plasmon resonance

Edwige Moyroud[1,2,†], Mathieu C. A. Reymond[1,†], Cécile Hamès[2], François Parcy[2,*] and Charles P. Scutt[1,*]

[1]*Laboratoire de Reproduction et Développement des Plantes, (UMR CNRS 5667 – INRA – Université de Lyon), Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France, and*

[2]*Laboratoire de Physiologie Cellulaire Végétale, (UMR CNRS 5168 – CEA – INRA – UJF), Bâtiment C2, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France*

Received 23 February 2009; revised 6 April 2009; accepted 23 April 2009; published online 27 May 2009.

*For correspondence (fax +33 472728600; e-mail Charlie.Scutt@ens-lyon.fr; fax +33 438785091; e-mail francois.parcy@cea.fr).

†These two authors contributed equally to this work.

SUMMARY

We demonstrate that the biophysical technique of surface plasmon resonance (SPR) analysis, which has previously been used to measure transcription factor binding to short DNA molecules, can also be used to characterize interactions involving entire gene promoters. This discovery has two main implications that relate, respectively, to novel qualitative and quantitative uses of the SPR technique. Firstly, SPR analysis can be used qualitatively to test the capacity of any transcription factor to interact physically with its putative target genes. This application should prove particularly useful for the confirmation of predicted transcriptional interactions in model species, and for comparative studies of non-model species in which transcriptional interactions are not amenable to study by other methods. Secondly, SPR may be used quantitatively to characterize interactions between transcription factors and gene promoters containing multiple *cis*-acting sites. This application should prove useful for the detailed dissection of promoter function in known target genes. The qualitative and quantitative applications of the SPR analysis of whole promoters combine to make this a uniquely powerful technique, which should prove particularly useful in systems biology, evolutionary developmental biology and various branches of applied biology.

Keywords: surface plasmon resonance, transcription factor, transcriptional regulation, DNA–protein interaction, promoter.

INTRODUCTION

Transcriptional regulation lies at the heart of most biological processes in plants and other organisms. This form of regulation depends on physical interactions that take place between transcription factors and DNA binding sites present in the *cis*-regulatory regions (promoters etc.) of their target genes. However, the binding site preferences of transcription factors are often poorly defined or unknown, rendering difficult or impossible the direct characterization of their DNA binding interactions using existing *in vitro* methods. Furthermore, transcriptional regulation often depends on the positions and binding affinities of multiple sites present in gene promoters. Even in cases where binding site preferences are known, the existing *in vitro* techniques are largely incapable of characterizing interactions involving multiple binding sites.

Surface plasmon resonance (SPR) has been used to quantify many types of molecular interaction, including transcription factor binding to individual target sites. In this method, short DNA molecules are typically attached to the surface of a gold-coated chip, in contact with a solution containing a transcription factor of interest. Molecular interactions between these two components are then measured, using a dedicated SPR analyser, from the angular deflection of a band of extinction that occurs within a beam of plane-polarized light reflected from the surface of the chip. Under typical experimental conditions, the coupling of this optical deflection to molecular interactions extends for approximately 200 nm into the solution (Lukosz, 1997; Kunz and Cottier, 2006). We reasoned that this range should be sufficient to measure interactions involving immobilized

DNA molecules of considerable length, given that the non-linear average conformation of long DNA molecules would tend to bring interaction sites closer to the surface of the chip.

We present here a proof of concept for the use of SPR analysis to measure transcription factor binding to long DNA molecules, such as entire gene promoters. We show this technique to be capable of both qualitative use, to discriminate between target and non-target genes, and quantitative use, to integrate the effects of binding to multiple sites within long DNA molecules. SPR analysis can be performed using recombinant proteins and promoters from any species, and can thus be used to study transcriptional interactions in any organism, including non-models. We discuss the potential qualitative and quantitative uses of this novel application of SPR in various fields of plant biology in which transcriptional regulation is of key importance.

RESULTS

SPR analysis distinguishes qualitatively between target and non-target genes

In many cases, it would be useful to determine whether a gene of interest may be directly regulated by a given transcription factor. It is frequently not possible to predict such transcriptional interactions from the presence of consensus binding sites in promoter regions, firstly because binding preferences for the factor of interest may be unknown, and secondly because not all consensus sites occur in a DNA context that will permit binding. Conversely, cryptic sites may also occur that bind transcription factors with high affinity, but do not conform to the known consensus.

To test the use of SPR to discriminate qualitatively between target and non-target DNA, we studied DNA–protein interactions involving the well-characterized transcription factor LEAFY (LFY) from *Arabidopsis thaliana*. LFY controls floral patterning by inducing the expression of genes encoding several members of the MADS-box transcription factor family (Parcy *et al.*, 1998; Busch *et al.*, 1999; Lamb *et al.*, 2002). We compared the interactions of the C-terminal DNA binding domain of LFY (LFY-C) (Hamès *et al.*, 2008) with the promoters of two of its MADS-box target genes, *APETALA1* (*AP1*) and *APETALA3* (*AP3*), and with three negative control DNAs derived from the Epstein bar virus gene *BHLF1*, and from the *A. thaliana* genes *GAPA-2*, encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and *LONG HYPOCOTYL IN FAR-RED* (*HFR1*) (Fairchild *et al.*, 2000). The *AP1* and *AP3* DNA molecules used in these assays contained two and one consensus binding motifs for LFY, CCANTG(G/T), respectively. One of these sites in *AP1*, and the unique site in *AP3*, has been experimentally verified to bind LFY (Busch *et al.*, 1999). Among the negative control DNAs used, *GAPA-2* and *HFR1* contained four and one LFY-consensus motifs (as above), respectively, although these genes are not known or suspected to be regulated by LFY. Equal quantities, in arbitrary SPR response units (RU), of biotinylated DNA samples were immobilized in separate channels on streptavidivin-coated SPR chips. Interactions between LFY-C and immobilized DNA molecules were then monitored by SPR over a range of protein concentrations (Figure 1a–e). These analyses were performed in the constant presence of non-specific competitor DNA in solution, so as to avoid signal saturation by non-specific binding.

Differences between the target and non-target genes of LFY were visually apparent in the SPR interaction curves obtained: low concentrations of LFY-C (lower traces in Figure 1a–e) gave higher SPR responses to target (Figure 1a,b) than to non-target (Figure 1c–e) DNA. Moreover, at the end of protein injection, when LFY-C was removed from DNA by washing (shown by the decreasing traces), this dissociation occurred more rapidly from target than non-target DNA. The apparent equilibrium constant for dissociation (K_D^{App}) values were calculated from real-time interaction data using dedicated SPR analysis software (Table 1), assuming a 1:1 (Langmuir) binding model. Chi-squared test values of less than two for specific interactions (Table 1) indicated a good fit of this model to the observed data. As an independent verification, K_D^{App} values were also calculated from the linear gradients of graphs based on the equilibrium condition (Figure 1f). The K_D^{App} values obtained using both methods of calculation were in good agreement, and fell within the 10-nM range for both the *AP1* and *AP3* target genes (Figure 1f and Table 1). The measurement of interactions with non-target DNA yielded K_D^{App} values of between 1.3 and 40 μM, some two to three orders of magnitude higher than those for target genes (Table 1). In the case of LFY, SPR analysis thus proved capable of easily distinguishing between target and non-target DNA strands of up to 3.1 kb in length (Table 1). We conclude that the consensus motifs for LFY present in the *GAPA-2* and *HFR1* non-target DNAs used in this study do not occur in a DNA context that allows high-affinity binding to LFY. SPR analysis was thus able to discriminate between target genes containing consensus sites of biological significance from non-target genes containing very similar sites, also corresponding to the LFY consensus motif.

To test the ability of SPR to discriminate between target and non-target genes using a second type of transcription factor, we analysed the interactions of the human heat shock factor, HSF1, with the promoter of its predicted target gene *hsp70* (Morgan *et al.*, 1987), and with a negative control DNA derived from the *BHLF1* gene. The *hsp70* DNA fragment used in these assays measured

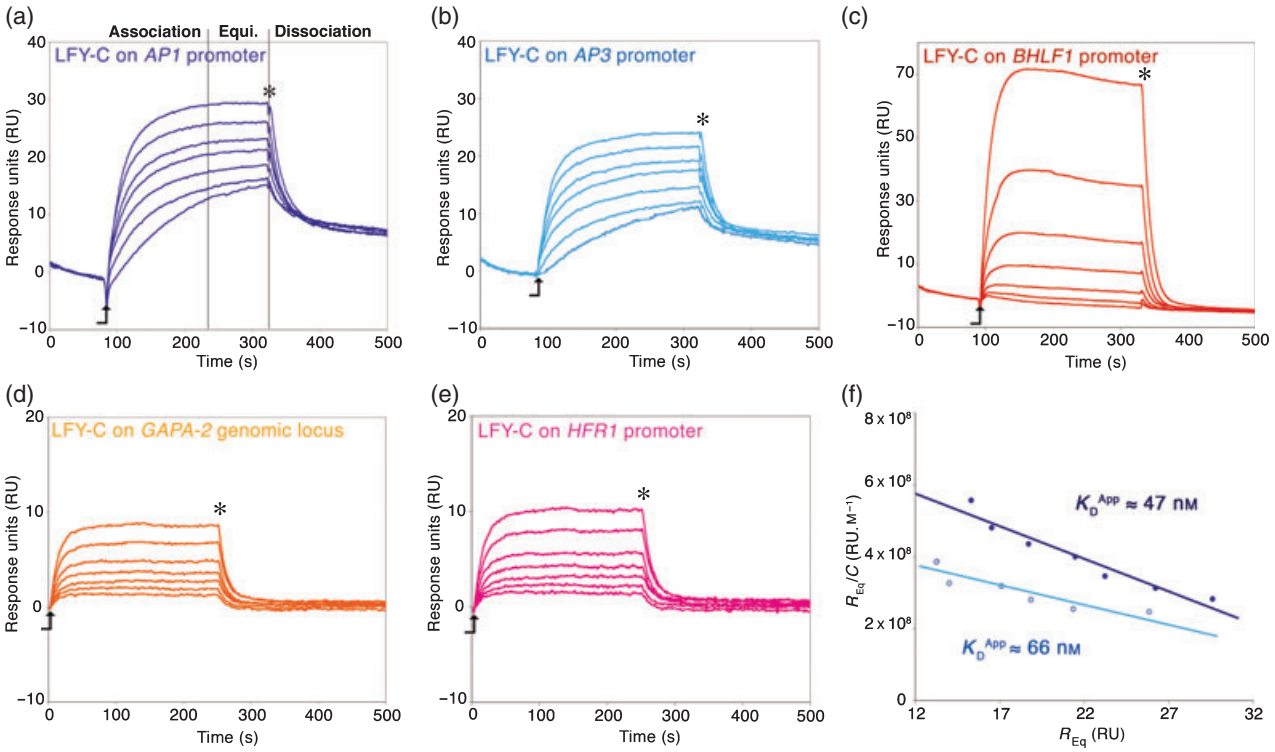


Figure 1. Simple discrimination between target and non-target genes of LFY.
(a–e) Surface plasmon resonance (SPR) curves for interactions of LFY-C with the regulatory regions of two target (a and b) and three non-target genes (c–e), showing specific interactions with target genes. The protein concentrations used, in descending order of response units (RU), were: 105, 84, 67, 54, 43, 34 and 28 nM (or 104, 69, 46, 31, 21, 14 and 9.1 nM, in the case of BHLF). The start and end of protein injections are marked with an arrow and an asterisk, respectively.
(f) Linear relationship of R_{Eq}/C against R_{Eq} for interactions of LFY-C with target promoter (a and b), where R_{Eq} corresponds to the SPR response (RU) at equilibrium for a given protein concentration (C). The linearity of the plots provides an independent validation of the 1:1 binding model. The apparent equilibrium constant for dissociation (K_D^{App}) values (shown), derived from the inverse negative of the linear gradients (Majka and Speck, 2007), are in agreement with the calculated values (Table 1).

Table 1 Kinetic constants for interactions of LFY-C with target and non-target genes

DNA tested	Length (bp)	Number of LFY consensus sites [CCANTG(G/T)]	LFY target	Apparent rate constant for dissociation k_{off}^{App} (sec^{-1})	Apparent rate constant for association k_{on}^{App} (M^{-1} sec^{-1})	Apparent equilibrium constant for dissociation K_D^{App} (nM)	χ^2
AP1	2386	2	Yes	$2.52 \times 10^{-3} \pm 0.5 \times 10^{-4}$	$1.59 \times 10^{5} \pm 0.5 \times 10^{4}$	15.8 ± 0.9	0.249
AP3	3050	1	Yes	$2.38 \times 10^{-3} \pm 0.2 \times 10^{-5}$	$1.11 \times 10^{5} \pm 0.7 \times 10^{4}$	21.4 ± 2.6	0.200
BLHF1	917	0	No	$2.96 \times 10^{-1} \pm 0.1 \times 10^{-5}$	$6.76 \times 10^{3} \pm 2 \times 10^{3}$	>40 000	6.2
GAPA-2	2444	4	No	$5.99 \times 10^{-2} \pm 0.5 \times 10^{-3}$	$2.55 \times 10^{3} \pm 0.2 \times 10^{3}$	~24 000	0.763
HFR1	2150	1	No	$5.12 \times 10^{-2} \pm 0.4 \times 10^{-3}$	$3.56 \times 10^{3} \pm 0.2 \times 10^{3}$	~15 000	1.05

2.4 kb and contained one HSF1 binding site [GAA(C/T)NTTC; Kroeger and Morimoto, 1994]. The K_D^{App} for the interaction between HSF1 and *hsp70* (Figure 2a), either calculated using dedicated SPR software (Table 2) or calculated from plots of equilibrium data (Figure 2b), fell within the nanomolar range. No interaction was observed between HSF1 and non-target *BHLF1* DNA (Figure 2a). Hence, in the case of HSF1, SPR was able to completely discriminate between target and non-target DNA, even without quantitative analysis. Experiments using both LFY-C and HSF1 have thus shown SPR analysis to provide a simple test of whether a given promoter may be bound by a transcription factor of interest.

SPR quantifies binding to multiple sites in *cis*-regulatory regions

In many cases, the level of transcription of a given gene will depend quantitatively on the binding affinity of

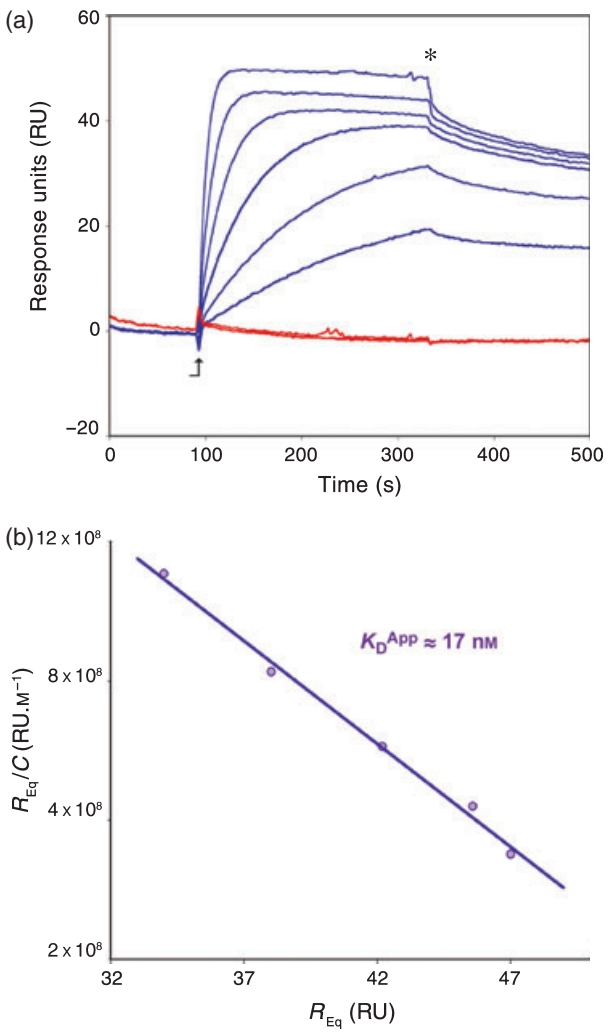


Figure 2. Simple discrimination between target and non-target genes of HSF1.
(a) Surface plasmon resonance (SPR) curves for the interaction of HSF1 with the *hsp70* target gene (blue traces) and *BHLF1* (red traces) non-target gene, showing the specific interaction with *hsp70*. The HSF1 concentrations used, in descending order of SPR response, were: 155, 104, 69, 46, 31 and 21 nM. The start and end of protein injections are marked with an arrow and an asterisk, respectively.
(b) Linear relationship of R_{Eq}/C against R_{Eq}, where R_{Eq} corresponds to the SPR response (RU) at equilibrium for a given protein concentration (C), for the interaction between HSF1 and *hsp70* provides an independent validation of the 1:1 binding model and an estimate of apparent equilibrium constant for dissociation (K_D^{App}, shown), in good agreement with the calculated value (Table 2).

transcription factors to multiple sites in its promoter region. SPR analysis is known to represent one of the most quantitative techniques available for the measurement of physical interactions involving individual binding sites (Majka and Speck, 2007). To test whether these quantitative characteristics are conserved when long DNA molecules containing multiple sites are analysed, we used SPR to investigate interactions involving the second intron of *AGAMOUS* (*AG*), which is necessary for the transcriptional regulation of this gene by LFY. The *AG* second intron measures 3.0 kb and contains four consensus LFY binding sites (Hong *et al.*, 2003). We used the classical technique of electrophoretic mobility shift assay (EMSA) to compare the individual binding of LFY-C to these four sites, BS1–4 (Figure 3a). This study identified BS1 as the site with the highest affinity for LFY, followed by BS2, BS4 and finally BS3 (Figure 3b).

Following the analysis of individual LFY binding sites, we made mutant versions of the *AG* second intron in which either the two highest affinity sites, BS1 and BS2, or all four sites, BS1–4, were mutated to eliminate their capacity to bind LFY (Figure 3a). We then performed SPR analyses to derive K_D^{App} values for interactions of LFY-C with the wild-type and two mutated versions of the *AG* second intron produced (Table 3). A K_D^{App} value of 0.6 nM was obtained for the wild-type intron, compared with values of 42 and 270 nM for the doubly and quadruply mutated introns, respectively. K_D^{App} values for the overall interaction with the *AG* second intron thus increased progressively with the elimination of LFY binding sites, showing a 70-fold proportional increase on mutation of the two highest affinity binding sites, and a further sixfold increase on mutation of the two remaining, lower affinity sites. These results clearly demonstrate SPR analysis to be capable of the quantitative comparison of transcription factor binding to long DNA molecules possessing multiple binding sites, indicating the usefulness of the SPR technique for the *in vitro* dissection of promoter function. Interestingly, the K_D^{App} value of the quadruply mutated intron (Table 3) remained somewhat lower than the values measured for interactions of LFY-C with non-target DNAs (Table 1), which may indicate the presence of a cryptic binding site in the *AG* second intron, which does not conform to the known consensus.

Table 2 Kinetic constants for interactions of HSF1 with target and non-target genes

DNA tested	Length (bp)	Number of HSF1 consensus sites [GAA(C/T)NTTC]	HSF1 target	Apparent rate constant for dissociation k_{off}^{App} (sec^{-1})	Apparent rate constant for association k_{on}^{App} (M^{-1} sec^{-1})	Apparent equilibrium constant for dissociation K_D^{App} (nM)	χ^2
hsp70	2532	1	Yes	$6.48 \times 10^{-4} \pm 0.4 \times 10^{-5}$	$3.03 \times 10^{5} \pm 0.3 \times 10^{4}$	3.5 ± 0.033	0.255
BHLF1	917	0	No	No interaction	No interaction	No interaction	No interaction

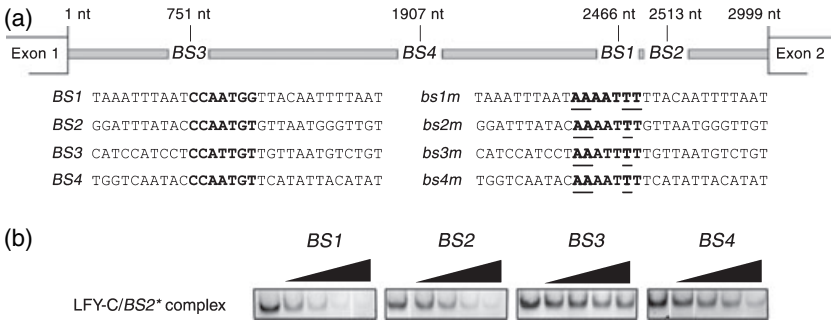


Figure 3. Quantitative measurement of interactions with multiple binding sites in *cis*-acting regions.
(a) Positions and sequences of wild-type (BS1–4) and mutant (bs1–4m) LFY binding sites in the *AG* second intron, as used in electrophoretic mobility shift assay (EMSA) and surface plasmon resonance (SPR) analyses. LFY consensus sites are indicated in bold, and mutated bases are underlined.
(b) EMSA assays showing the relative affinities of LFY consensus motifs BS1–4, by the competition of different quantities of unlabelled target oligonucleotides (0-, 5-, 25-, 100- and 500-fold excess, from left to right) with a fluorescently labelled BS2 complex with LFY-C.

Table 3 Kinetic constants for interactions of LFY-C with wild-type and mutated versions of the *AG* second intron

DNA tested	Length (bp)	Number of LFY consensus sites [CCANTG(G/T)]	Apparent rate constant fordissociation k_{off}^{App} (sec^{-1})	Apparentrate constant for association k_{on}^{App} (M^{-1} sec^{-1})	Apparent equilibrium constant for dissociation K_D^{App} (nM)	χ^2
AG WT	3176	4	$7.04 \times 10^{-5} \pm 0.2 \times 10^{-5}$	$1.18 \times 10^{5} \pm 2 \times 10^{3}$	0.6 ± 0.02	1.20
AG bs12m	3176	2	$2.38 \times 10^{-3} \pm 0.4 \times 10^{-5}$	$5.68 \times 10^{4} \pm 0.9 \times 10^{3}$	41.9 ± 0.7	0.735
AG bs1234m	3176	0	$1.49 \times 10^{-2} \pm 0.5 \times 10^{-3}$	$5.57 \times 10^{4} \pm 0.5 \times 10^{4}$	267.5 ± 30.5	0.654

DISCUSSION

Why length is so important

Despite the availability of commercial SPR analysers for around 20 years, SPR analysis has not been widely taken up by researchers working on transcription factors. This lack of widespread interest has probably been because of the perception that SPR, as applied to transcriptional interactions, was limited to the analysis of short DNA fragments. Accordingly, this technique has mostly been used for detailed, quantitative studies of transcription factor binding to individual *cis*-acting sites, rather than to relate such biophysical events to their higher level biological effects. In the present work, we demonstrate two things: firstly that SPR can be used to detect transcription factor binding to much longer DNA fragments than was previously believed possible, and secondly that this technique retains its quantitative value when applied to such long DNA molecules. The first of these findings means that SPR can be used as a simple and rapid, qualitative test of whether a gene of interest may be the direct target of a given transcription factor. As a general rule of thumb, any interaction yielding a K_D^{App} of below 100 nM is probably worthy of further investigation. The second finding opens the possibility of more subtle uses of SPR in the analysis of complex transcriptional interactions involving multiple binding sites, and/or multiple transcription factors. Both the qualitative and quantitative uses of SPR should find many applications to biological problems, as described below.

To bind or not to bind: novel qualitative applications of SPR analysis

A major objective of systems biology is to describe how networks of transcriptional regulators control complex biological processes. A first requirement of such studies is the identification of the direct target genes of transcription factors participating in the networks of interest, which can be achieved using such *in vivo* techniques as microarray analyses (Gomez-Mena *et al.*, 2005) and chromatin immunoprecipitation (ChIP) (Weinmann, 2004). These procedures yield lists of putative target genes, which must then be verified by independent techniques including the *in vitro* analysis of DNA binding. Previously, such *in vitro* analyses have only been possible for transcription factors for which binding sites could be found within *cis*-regulatory regions. The SPR analysis of entire gene promoters, as demonstrated in the present work, will provide a rapid means of verifying such predicted transcriptional interactions, even in cases in which binding sites cannot be predicted from the gene sequences under analysis. This feature is important not only for transcription factors for which binding site preferences are currently unknown, but

also for the many cases in which transcription factors bind to cryptic sites that show little similarity to their known consensus sequences. SPR analysers of the latest generation are capable of measuring binding to several hundred DNA samples simultaneously, and should therefore prove ideal for the rapid verification of lists of direct target genes identified using large-scale approaches such as microarray and ChIP-seq (Robertson *et al*., 2007) analyses. Indeed, with the possibility of analysing large numbers of bound DNA molecules simultaneously, SPR could also be used for the *in vitro* characterization of entire transcriptional networks by sequentially passing all transcription factors in a network over the complete set of the promoters of that network. In this way, the transcriptional relationships linking all the components of a system could be identified.

Differences in transcriptional control relationships between organisms account for much of the biodiversity of the natural world, as demonstrated on a micro-evolutionary scale by many of the molecular causes underlying the domestication of crop plants (Doebley *et al*., 2006). However, our current knowledge of transcriptional interactions derives exclusively from the study of a few model organisms that are amenable to genetic analysis. To identify the differences in transcriptional interactions that are responsible for biodiversity in plants and other organisms, a means is needed to test whether interactions are conserved between the well-studied model species and others chosen for their key phylogenetic positions or strategic importance. The SPR analysis of entire gene promoters should prove ideal for this purpose as it can be performed using recombinant transcription factors and putative target genes from any species. The use of SPR to characterize transcriptional relationship in non-model species will thus be of great importance to both evolutionary developmental (evo-devo) biology (Frohlich and Chase, 2007) and to agricultural science and other branches of applied biology.

Putting a figure on it: novel quantitative applications of SPR analysis

In the present work, we have shown that the quantitative value of SPR analysis is conserved for interactions involving long DNA molecules. Accordingly, the SPR analysis of transcription factor binding to a given target promoter may yield a single K_D^{App} value that quantifies the overall binding interaction between those two components. By repeating SPR analyses using mutated versions of a promoter of interest, it should be possible to determine the importance to the overall binding interaction of the positions and affinities of all the individual binding sites present. Such an approach would rapidly indicate the presence of such phenomena as cooperative binding, where several sites of individually low affinity may, for example, combine to produce a strong overall binding interaction.

Transcription factors frequently bind to DNA as complexes, and these may also interact simultaneously with several *cis*-acting sites positioned at considerable distances along the target gene. For example, the MADS-box transcription factors encoded by three of the target genes analysed in the present work, *AP1*, *AP3* and *AG*, are hypothesized, on the basis of genetic evidence, to form various combinations of tetramers, known as floral quartets, (Theissen and Saedler, 2001) with a further class of MADS-box protein, SEPALLATA (Pelaz *et al*., 2000; Honma and Goto, 2001). According to this hypothesis, four different tetrameric complexes of MADS-box proteins would interact with multiple *cis*-acting sites in four different sets of target genes to specify the four types of floral organ: sepals, stamens, petals and carpels. Several studies have identified putative MADS-box targets in specific floral organs (Sablowski and Meyerowitz, 1998; Ito *et al*., 2004; Gomez-Mena *et al*., 2005; Sundstrom *et al*., 2006). Furthermore, a recent study employing the classical EMSA method has shown SEP proteins to be capable of interacting as tetramers with pairs of precisely spaced consensus MADS-box transcription factor binding sites in a short, artificial DNA molecule (Melzer *et al*., 2009). However, classical techniques such as EMSA cannot be used to characterize the binding of MADS-box complexes to sets of native target promoters measuring up to several kilobases in length. SPR analysis may thus prove the ideal technique to finally demonstrate or refute the floral quartet hypothesis, and to investigate the many other cases of transcription factor binding to multiple target sites in plants and other organisms.

EXPERIMENTAL PROCEDURES

Expression and purification of recombinant protein

The *A. thaliana* LFY DNA binding domain (LFY-C), corresponding to amino acids 223–424 of the full-length LFY protein, was inserted into the *pETM11* expression vector (Dummler *et al*., 2005) to yield the *pCH28* plasmid (Hamès *et al*., 2008), thereby permitting the production of LFY-C fused to an N-terminal 6x histidine tag. Cell cultures were generated in the *Escherichia coli* Rosetta Blue DE3-pLysS strain (Novagen, Merck, http://www.merckbiosciences.co.uk/html/NVG/home.html), and the production of recombinant protein was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM. Following incubation overnight at 22°C, cell cultures were lysed by sonication. Recombinant protein was then purified by affinity chromatography on Ni-NTA resin (Qiagen, http://www.qiagen.com). Protein-containing fractions were pooled and subjected to size-exclusion chromatography using a Hi-load Superdex-200 16/60 preparation grade column (GE Healthcare, http://www.gehealthcare.com) to eliminate protein aggregates.

Preparation of biotinylated DNA molecules for SPR analysis

DNA fragments containing *cis*-acting regulatory regions were obtained by PCR amplification using the primers shown in Table S1,

on templates of genomic DNA from *A. thaliana* plants of the Columbia ecotype or from human, except for *AP1*, which was amplified from a plasmid supplied by Dr R. Benlloch (Umeå Plant Science Centre, Sweden). The DNA fragments obtained were ligated into the PCR cloning site of *pGEM-T-Easy* (Promega, http://www.promega.com). A *pBluescriptI*-derived plasmid containing part of *BHLF1*, corresponding to bases 52401–53092 of the Epstein–Barr virus genome (Genbank accession V01555), was obtained from Dr Henri Gruffat (École normale supérieure de Lyon, http://www.ens-lyon.eu). Mutant versions of the second intron of *AG*, containing mutations to disrupt the function of LFY binding sites, as shown in Figure 3(a), were constructed by sequential site-directed mutagenesis of this sequence in *pGEM-T-Easy* (Kirsch and Joly, 1998). Double-stranded DNA molecules for SPR analysis were synthesized by PCR amplification from recombinant *pGEM-T-Easy* plasmids using T7 and SP6 primers, or from recombinant *pBluescriptI* plasmids using M13 forward and reverse sequencing primers, one primer of each pair being biotinylated in each case, and subsequently purified using a NucleoSpin® Extract II kit (Macherey-Nagel, http://www.macherey-nagel.com).

Immobilization of DNA samples for SPR analysis

CM5 SPR chips (Biacore, http://www.biacore.com) were activated in a Biacore T100 SPR Analyser to accept a streptavidin coating, using an amine coupling kit (Biacore) according to the manufacturer's instructions. Amine-coupling reagents were injected in a continuous flow of 5 μl min^{-1} of HBS-P+ buffer (0.01 M HEPES, pH 7.4, 0.15 M NaCl, 0.05% w/v P20 detergent). Aliquots of 50 μl of streptavidin (0.1 mg ml^{-1} in 10 mM sodium acetate, pH 4.2) were then injected, followed by 50-μl aliquots of ethanolamine (1 M, pH 8.5), to inactivate the residual carboxyl groups. The chips were then washed by two injections of 5 μl of HBS-P+ buffer. Double-stranded DNA molecules (5 ng μl^{-1} in HBS-P+ buffer, degassed before use) each carrying a single biotin moiety at its 3′ terminus, relative to the direction of transcription of the encoded gene, were immobilized in separate channels on SPR chips by injection at 10 μl min^{-1}, until 200 RU (arbitrary SPR units) of DNA had been added, corresponding to approximately 0.15 ng mm^{-2}. One channel was left empty on each chip as a reference. Unbound DNA was finally removed by injections of 10-μl aliquots of NaCl (1 M).

SPR analysis of DNA–protein interactions

SPR chips containing immobilized DNA samples were equilibrated in a Biacore T100 Analyser by injection of HBS-P+ buffer containing non-homologous DNA (40 ng μl^{-1}, molecular size >120 bp) from salmon testis (Roche, http://www.roche.com) until SPR responses were stable. Transcription factors, dissolved in the above buffer, were then injected into all channels of these chips for 250 sec at a flow rate of 50 μl min^{-1}, followed by an injection of further buffer for 300 sec to monitor protein dissociation. Transcription factor solutions were used in decreasing order of concentration, and SPR chips were regenerated between analyses by the sequential injection of guanidinium hydrochloride (3 M) and sodium dodecyl sulfate (0.03%, w/v), each for 60 sec at a flow rate of 50 μl min^{-1}.

SPR data analysis

Real-time SPR interaction curves were analysed using BIOEVAL T100 software (Biacore). For each analysis, the response of the reference channel was subtracted from the interaction curves obtained from the three experimental channels, and the resulting curves were then adjusted to zero at the start of transcription factor injection. These normalized curves were then fitted globally to a 1:1 (Langmuir) interaction model, permitting the determination of forward and reverse apparent rate constants (k_{on}^{App} and k_{off}^{App}). The validity of the interaction model was verified by data fitting, with a good fit indicated by $\chi^2 < 2$ (or <10 for low-affinity interactions). Equilibrium binding constants (K_D^{App}) were calculated from the ratio $k_{off}^{App}/k_{on}^{App}$, and also from SPR values (RU), at equilibrium over a range of protein concentrations.

EMSA assays

EMSA assays were performed as described in (Hamès *et al.*, 2008) using oligonucleotides (10 nM) labelled with 5-carboxytetramethylrhodamine, varying concentrations of unlabelled competitors and LFY-C protein at a concentration of 300 nM.

ACKNOWLEDGEMENTS

We thank Mike Frohlich for helpful discussions, Sylvie Ricard-Blum and Clément Faye for advice on SPR analysis, Reyes Benlloch and Henri Gruffat for supplying plasmid DNA and Hüseyin Besir (EMBL, Heidelberg) for providing the *pETM-11* vector. The SPR analysis was performed using equipment provided by the Institut Féderatif de Recherche-128 (Biosciences-Gerland, Lyon-Sud). We acknowledge grant funding from the French National Research Agency (Plant TF-Code/ANR-07-BLAN-0211-01) and doctoral studentships to EM (French Ministry of Research), CH and MR (Rhone-Alpes Region, Cluster 9).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. Oligonucleotide sequences used in the PCR amplifications.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

Busch, M.A., Bomblies, K. and Weigel, D. (1999) Activation of a floral homeotic gene in Arabidopsis. *Science* **285**, 585–587.

Doebley, J.F., Gaut, B.S. and Smith, B.D. (2006) The molecular genetics of crop domestication. *Cell* **127**, 1309–1321.

Dummler, A., Lawrence, A.M. and de Marco, A. (2005) Simplified screening for the detection of soluble fusion constructs expressed in *E-coli* using a modular set of vectors. *Microb. Cell Fact.* **4**.

Fairchild, C.D., Schumaker, M.A. and Quail, P.H. (2000) HFR1 encodes an atypical bHLH protein that acts in phytochrome A signal transduction. *Genes Dev.* **14**, 2377–2391.

Frohlich, M.W. and Chase, M.W. (2007) After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature* **450**, 1184–1189.

Gomez-Mena, C., de Folter, S., Costa, M.M.R., Angenent, G.C. and Sablowski, R. (2005) Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. *Development* **132**, 429–438.

Hamès, C., Ptchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gerard, F., Martiel, J.L., Benlloch, R., Parcy, F. and Muller, C.W. (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *EMBO J.* **27**, 2628–2637.

Hong, R.L., Hamaguchi, L., Busch, M.A. and Weigel, D. (2003) Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. *Plant Cell* **15**, 1296–1309.

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

Ito, T., Wellmer, F., Yu, H., Das, P., Ito, N., Alves-Ferreira, M., Riechmann, J.L. and Meyerowitz, E.M. (2004) The homeotic protein AGAMOUS controls microsporogenesis by regulation of SPOROCYTELESS. *Nature* **430**, 356–360.

Kirsch, R.D. and Joly, E. (1998) An improved PCR-mutagenesis strategy for two-site mutagenesis or sequence swapping between related genes. *Nucleic Acids Res.* **26**, 1848–1850.

Kroeger, P.E. and Morimoto, R.I. (1994) Selection of new Hsf1 and Hsf2 dna-binding sites reveals differences in trimer cooperativity. *Mol. Cell. Biol.* **14**, 7592–7603.

Kunz, R.E. and Cottier, K. (2006) Optimizing integrated optical chips for label-free (bio-)chemical sensing. *Anal. Bioanal. Chem.* **384**, 180–190.

Lamb, R.S., Hill, T.A., Tan, Q.K.G. and Irish, V.F. (2002) Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. *Development* **129**, 2079–2086.

Lukosz, W. (1997) Integrated-optical and surface-plasmon sensors for direct affinity sensing. 2. Anisotropy of adsorbed or bound protein adlayers. *Biosens. Bioelectron.* **12**, 175–184.

Majka, J. and Speck, C. (2007) Analysis of protein-DNA interactions using surface plasmon resonance. *Adv. Biochem. Eng. Biotechnol.* **104**, 13–36.

Melzer, R., Verelst, W. and Theissen, G. (2009) The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in floral quartet-like complexes in vitro. *Nucleic Acids Res.* **37**, 144–157.

Morgan, W.D., Williams, G.T., Morimoto, R.I., Greene, J., Kingston, R.E. and Tjian, R. (1987) 2 Transcriptional Activators, Ccaat-Box-Binding Transcription Factor and Heat-Shock Transcription Factor, Interact with a Human Hsp70-Gene Promoter. *Mol. Cell. Biol.* **7**, 1129–1138.

Parcy, F., Nilsson, O., Busch, M.A., Lee, I. and Weigel, D. (1998) A genetic framework for floral patterning. *Nature* **395**, 561–566.

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

Robertson, G., Hirst, M., Bainbridge, M. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657.

Sablowski, R.W.M. and Meyerowitz, E.M. (1998) A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA. *Cell* **92**, 93–103.

Sundstrom, J.F., Nakayama, N., Glimelius, K. and Irish, V.F. (2006) Direct regulation of the floral homeotic APETALA1 gene by APETALA3 and PISTILLATA in Arabidopsis. *Plant J.* **46**, 593–600.

Theissen, G. and Saedler, H. (2001) Plant biology – floral quartets. *Nature* **409**, 469–471.

Weinmann, A.S. (2004) Innovation – Novel ChIP-based strategies to uncover transcription factor target genes in the immune system. *Nat. Rev. Immunol.* **4**, 381–386.

Conclusion

Prédire exactement le comportement des facteurs de transcription vis-à-vis des séquences promotrices des génomes est un challenge ambitieux de la bioinformatique. Nous avons développé ici une méthode qui teste expérimentalement la capacité d'une séquence promotrice à fixer efficacement un facteur d'intérêt en caractérisant de manière quantitative l'interaction observée quel que soit le nombre d'éléments *cis* présents dans la séquence génomique. Ce type d'analyse peut fournir les données expérimentales qui font souvent défaut pour valider les modèles bioinformatiques proposés et améliorer ainsi leur pouvoir de prédiction.

Nous avons proposé que cette méthode pourrait servir à vérifier les cibles directes d'un facteur de transcription chez un organisme modèle. Mathieu Reymond (Equipe C. Scutt, RDP, Lyon) a récemment utilisé la SPR pour tester si la régulation d'une liste de gènes cibles du facteur de transcription d'*Arabidopsis* SPATULA établie par expérience de puce à ADN passait par une interaction directe entre la protéine et le promoteur des gènes en question. Cette méthode a aussi été développée pour examiner l'existence d'interactions facteur de transcription/ADN chez les espèces non-modèles pour lesquelles il n'existe pas d'autre technique facilement utilisable. Ainsi, nous avons pu tester l'existence d'une reconnaissance par LFY et ses homologues des promoteurs des gènes B chez une gymnosperme. Les données obtenues sont présentées dans le dernier chapitre de résultats de cette thèse.

CHAPITRE 4

LEAFY et l'origine de la fleur : existence d'un réseau pré-floral chez Welwitschia mirabilis

Après avoir examiné l'évolution de la spécificité de liaison de LFY et ses homologues, nous avons exploré le rôle méconnu de ce facteur de transcription chez une espèce non-angiosperme. Nous avons combiné les techniques classiques de la démarche « évo-dévo » (clonage d'homologues chez une espèce occupant une position clé de la phylogénie, analyse des patrons d'expression) avec les méthodologies développées au cours de cette thèse et présentées dans les chapitres précédents. Les résultats obtenus devraient participer au développement de modèles gymnospermes afin d'élucider les mécanismes à l'origine de la naissance de la fleur.

Introduction

Nous avons cherché à établir les propriétés de la protéine LEAFY (LFY) et son paralogue NEEDLY d'une gymnosperme, *Welwitschia mirabilis*, afin de tester dans quelle mesure la seule duplication 'réussie' de LFY au cours de l'histoire des plantes terrestres s'est accompagnée d'une spécialisation des rôles au sein de la famille. Les résultats obtenus au cours de cette étude apportent des données sur une espèce jamais caractérisée d'un point de vue moléculaire et contribuent à esquisser une vue générale du rôle joué par LFY et son paralogue chez le groupe frère des angiospermes, pré-requis nécessaire pour imaginer les fonctions de ces protéines chez l'ancêtre commun disparu. Notre approche vise aussi à tester dans quelle mesure l'utilisation des nouveaux outils développés (Selex, Biacore) permet de progresser en l'absence d'organisme modèle chez un groupe clé de la phylogénie des plantes. Les résultats générés ne constituent pas une démonstration en tant que tel, mais permettent de réviser les idées en vigueur dans les derniers scénarios évolutifs proposés en apportant des évidences expérimentales d'un genre nouveau.

Les premières données obtenues ont servi à la rédaction d'une ébauche d'article qui constitue le cœur de ce chapitre. Des résultats supplémentaires sont exposés dans la seconde partie du chapitre.

Control of B genes expression by *LEAFY* and *NEEDLY* orthologs in *Welwitschia*: evidence of a pre-floral network in gymnosperms

Edwige Moyroud, Emmanuel Thévenon, Florence Louis, Mike Frohlich and François Parcy

Abstract

Most of today's plant species are flowering plants or angiosperms as the flower represents unique advantages for plant reproduction. However, the origin of the flower in evolution remains mysterious due to the absence of intermediate morphologies between reproductive structures of angiosperms and gymnosperms (their sister group) in extant plants or in the fossil records. The gene regulatory network responsible for the development of flowers in angiosperms is well established: it involves the activation of ABC floral homeotic genes by the LEAFY (LFY) transcription factor. In order to gain insight into the molecular events that could have led to the creation of flowers, we analyzed the structure of this network in the gymnosperm *Welwitschia mirabilis*. We found that expression patterns were consistent with a role of *LFY* and its paralog *NEEDLY* (*NLY*) in the regulation of B and C genes; more precisely LFY could regulate B and NLY the C genes expression. Biochemical analysis of LFY and NLY revealed that both proteins exhibit different properties, consistently with this hypothesis. This work provides the first evidence for the existence of a pre-floral network in gymnosperms and allows us to build hypotheses of flower origin.

Introduction

Morphologically unlike and long-diverged taxa often share similar toolkits of genes to build and pattern their bodies. Such genes stand at the heart of crucial genetic regulatory networks (GRNs) frequently recruited by evolution and domestication to govern the formation and differentiation of various anatomies, thus participating in diversity (Carroll, 2008). In flowering plants, the *LEAFY/FLORICAULA* (*LFY/FLO*) gene encodes a central regulator of development, controlling both the transition from vegetative to reproductive phase and the patterning of floral meristems (Moyroud et al., 2009). Genetics studies in *Arabidopsis* and other angiosperms showed that this unique plant transcription factor acts as a direct activator of at least three MIKC-type MADS-box genes of the ABC model: *APETALA1* (*AP1*, A-class gene), *APETALA3* (*AP3*, B-class gene) and *AGAMOUS* (*AG*, C-class gene), which confer

their identity to the four typical organs of a flower (Parcy et al., 1998; Busch et al., 1999; Lohmann et al., 2001; Lamb et al., 2002). As a result, LFY controls a minimal network controlling the first critical stages of floral development (here referred as 'floral network'). *LFY* homologs have been identified in the other groups of extant land plants that do not form flowers but when *LFY-like* genes started to be involved in reproductive development is unknown (Moyroud et al., 2010).

Homeotic selectors of the ABC model arose from an expansion of the MADS-box genes family that predates the divergence of gymnosperms, the sister-group to flowering plants. As a result, homologs of B and C genes are also present in gymnosperms genomes (Theissen and Becker, 2004) and LFY homologs could possibly regulate their expression before the invention of the flower. Orthologs of *AP1* are missing in gymnosperms (Becker et al., 2000; Litt and Irish, 2003) but *AP1* belongs to a larger clade, which also include *AGL6* (Theissen et al., 2000). Homologs of *AGL6* have been isolated in several conifers and it has been been proposed that they could also be involved in reproductive development (Tandre et al., 1995; Mouradov et al., 1998b; Carlsbecker et al., 2004). If a 'pre-floral network' was already active in the last common ancestor of gymnosperms and angiosperms (seed plants), then the floral network was not created *de novo* but rather comes from the recycling of a former genetic circuit. Understanding the nature of this recycling would help elucidating the still mysterious origins of flowering plants.

The last common ancestor of seed plant was probably a gymnosperm species today extinct (Chaw et al., 2000; Frohlich and Chase, 2007). As gymnosperms do not form flowers, evolution mechanisms may have omitted to recycle their reproductive GRN. Thus, relationships between *LFY*, B and C homologs in gymnosperms are likely to have remained more similar to the ancestral behaviour. Flowers bear their male and female organs together, i.e., stamens and carpels, on contiguous whorls; gymnosperms however, generally display their male and female reproductive structures on separate cones. One of the many events that led to the creation of the first flower was to gain the ability to form a bisexual structure (Baum and Hileman, 2006). Therefore, a reorganization of the molecular mechanisms that sustain reproductive development of seed plant certainly occurred on the lineage leading to angiosperms.

B and C genes are in charge of sexual organs identity in flowering plants: C-class genes confer a reproductive fate to a floral primordium while a more restricted expression of B-class genes specifies male versus female identity: stamens develop where B and C genes are both

expressed whereas carpels derive from primordia that express the C regulator only. A similar genetic system could specify the reproductive organs of gymnosperms, despite their morphological dissemblance to the flower (Theissen and Becker, 2004): previous studies indicated that C genes in gymnosperms are active in cones independently of their sex whereas B genes expression is only detected in male cone. Moreover, expressing B or C-class protein from gymnosperm in flowering plants deficient for B or C activities is generally sufficient to restore a wild type phenotype suggesting that B and C homologs in seed plants share most of their biochemical properties (Sundstrom and Engstrom, 2002; Winter et al., 2002; Zhang et al., 2004). To create a bisexual structure, one needs to generate a C-class domain next to a B+C class territory on a single reproductive axe. Therefore, a change in the regulators of B and C-class genes expression patterns and a modification of the *cis*-elements found within their promoters probably occurred on the lineage leading to flowering plants, once extant gymnosperms had diverged. As LFY is the best-known regulator of B and C-class genes in angiosperms, testing the ability of *LFY-like* genes to regulate B and C-class homologs in gymnosperms would help understanding the pre-floral network possibly at work in the last common ancestor of seed plants.

Gymnosperms (except the *Gnetum* genus) possess two *LFY-like* genes, first described in a pine species as *PRFLL* (*LFY* homolog from Monterey pine) and its paralog *NEEDLY* (*NLY*, (Mellerowicz et al., 1998; Mouradov et al., 1998a)). The expression patterns of *LFY* and *NLY* homologs have been analysed in several gymnosperms but are often incomplete and different between species (Mellerowicz et al., 1998; Mouradov et al., 1998a; Shindo et al., 2001; Carlsbecker et al., 2004; Guo et al., 2005; Dornelas and Rodriguez, 2006; Shiokawa et al., 2008)). A recent study however, provided an exhaustive description of *LFY*/*NLY* expression pattern in three conifers species (Vazquez-Lobo et al., 2007) and general tendencies were observed: the two paralogs are active in male and female cones but their expression patterns diverge as development progresses. In early stages, transcripts corresponding to both *LFY-like* genes are detected in overlapping domains such as reproductive meristems and organ primordia but the expression of *LFY* orthologs tends to decrease as organs grow, while *NLY* often remains highly transcribed. As development progresses, their expression domains become mutually exclusive: *LFY* orthologs are expressed in the pollen-producing tissues unlike *NLY*, whose expression is limited to the sterile appendage of the cone. Similarly, in female cones of two genera, *Picea* and *Podocarpus*, transcripts of *LFY* orthologs are detected in the ovule and the subtending scale whereas *NLY* orthologs are excluded from the ovule and

are preferentially activated in the vasculature. In conifers at least, these divergent spatio-temporal expression patterns suggest that the two paralogs could contribute differently to the development of the cones (Vazquez-Lobo et al., 2007). However, genes regulated by LFY/NLY homologs in gymnosperms remains to be found. Indeed, the role of *LFY-like* genes has not been established in this group, as gymnosperms are long life-cycles species not easily amenable to functional studies. Gymnosperms *LFY* and *NLY* genes expressed in *Arabidopis* or tobacco (*Nicotiana tabacum*) can partially compensate for the absence of their angiosperm counterparts but LFY orthologs always perform better than the NLY ones (Maizel et al., 2005; Dornelas and Rodriguez, 2006; Shiokawa et al., 2008). A specification of each paralog may thus have affected not only their expression patterns but also the properties of the corresponding proteins.

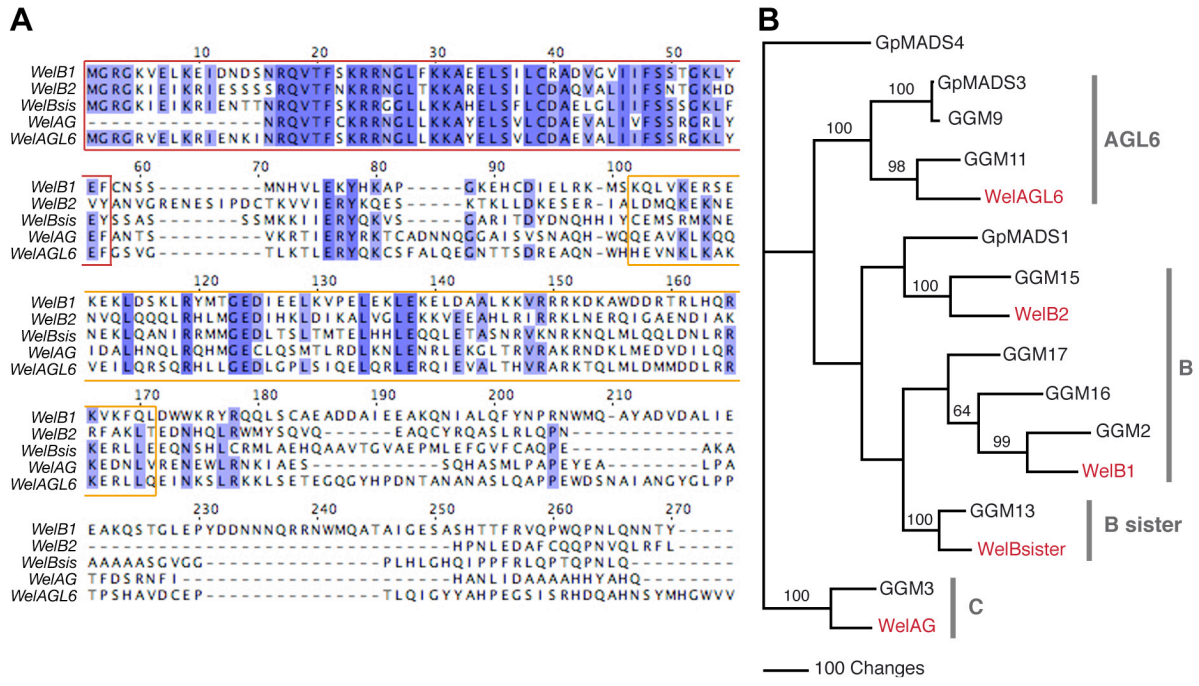
To test if a regulation of B and/or C genes by *LFY-like* genes was possible in gymnosperms, we characterized the properties of LFY/NLY proteins in *Welwitschia mirabilis*, a species widely spread during Mesozoic era but now limited to the costal deserts of Africa (Crane, 1996). *Welwitschia* belongs to the gnetophytes, one of the four groups of extant gymnosperms, whose members display angiosperms-like features such as vessel elements or cones with a flower appearance. For this reason, this group was regarded as 'an evolutionary link' between gymnosperms and flowering plants until the use of molecular markers clearly established their position within the gymnosperms. First, we identified several homologs of MADS-box genes in *Welwitschia mirabilis* and we showed that their expression patterns were compatible with a differential regulation by *LFY-like* genes (*WelLFY* and *WelNdly*) in this species. Then, we studied the ability of WelLFY and WelNdly proteins to interact with DNA *in vitro* and showed that the two paralogs exhibit different binding specificities. Next, we used Surface Plasmon Resonance (SPR) to test their capacity to recognize the upstream regions of B genes homologs from *Welwitschia*. Unlike WelNdly, WelLFY appears able to contact the promoters of both B genes. These data suggest that LFY homologs could have participated in the transcriptional regulation of B genes before the appearance of the flower and bring new evidence that both paralogs may have fulfilled different functions in the pre-floral network. The generalization of this model to gymnosperms as a whole and its consequences on the creation of the flower are discussed.

Results

Identification of MADS-box genes and their associated expression domains in *Welwitschia mirabilis*

The orthologs of *LFY* and *NLY* (*WelLFY* and *WelNdly*) have been isolated in *Welwitschia mirabilis* (Frohlich and Meyerowitz, 1997) and several EST corresponding to MADS-box genes have been identified in this species (Albert et al., 2005) but no detailed characterization of these genes has been reported. Using the sequences of B, C and *AGL6* clades members in other gymnosperms, we isolated 4 homologs of the homeotic selectors of floral development from reproductive cones of *W. mirabilis*. (Fig.1A) Our phylogenetic analysis indicated that they correspond to the *Welwitschia* homologs of *AG*, *AGL6* and B genes of *Gnetum* (Fig.1B) so we named them *WelAG*, *WelAGL6*, *WelB1* and *WelB2*. Gymnosperms also possess *Bsister* genes, a clade closely related to B genes, which could participate in the development of female organs of gymnosperms (Becker et al., 2002a; Becker et al., 2002b). Using a similar strategy, we isolated *WelBsister*, a homolog of Bsister gene in *W. mirabilis* (Fig.1A).

Fig.1 – Identification of 5 MADS-box genes in *W. mirabilis*.

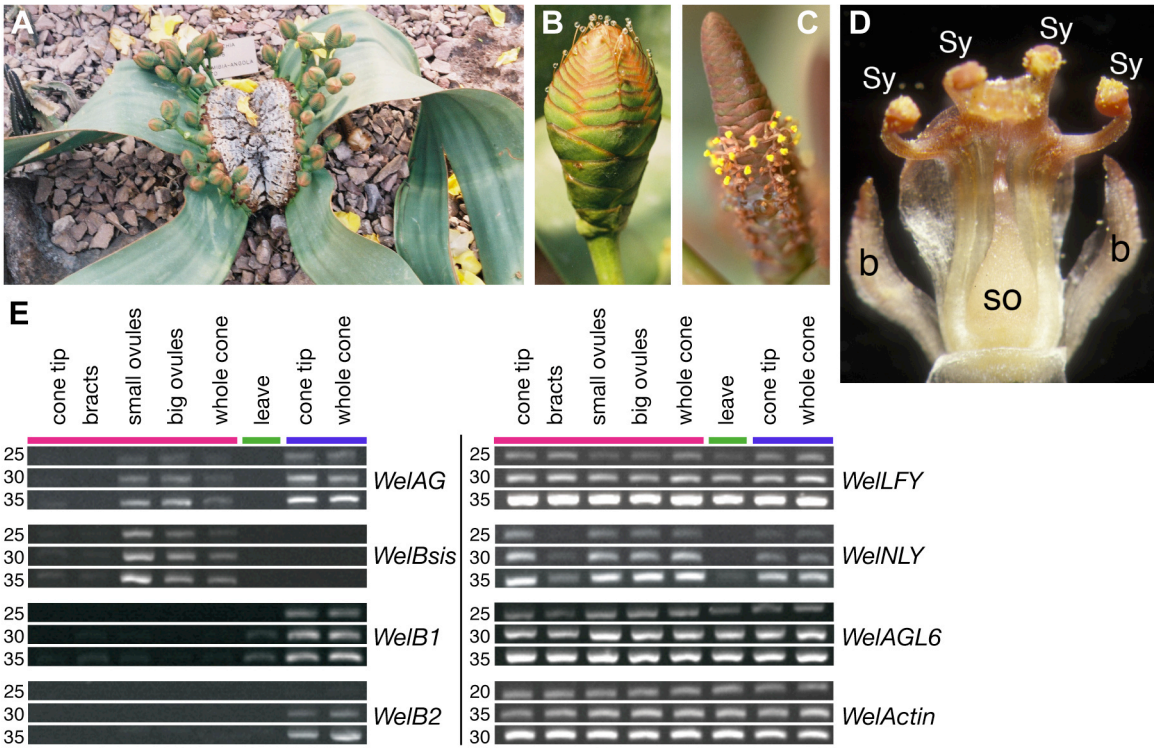


(A) Alignment of the predicted amino acid sequences of the *W. mirabilis* genes identified reveals a conserved domain structure of MIKC-type MADS-box proteins. The red box indicates the MADS-domain and the orange box indicates the K-domain according to (Becker et al., 2000). *WelAG* sequence is incomplete at both ends and the C-terminus of the *WelBsister* sequence is missing. Identical positions are shown in dark blue and positions conserved in more than 75% of the sequences are highlighted in light blue. (B) Relationships between a subset of MADS-domain proteins of *Gnetum parvifolium* (GpMADS), *Gnetum gnemon* (GGM) and *W. mirabilis* (Wel). *Welwitschia* proteins are indicated in red. The numbers next to some nodes give bootstrap percentages, shown only for relevant

nodes and those defining gene families (Winter et al., 1999; Becker et al., 2002b). Subfamilies are labelled in grey at the right margin.

We tested whether these genes were expressed in the reproductive structures of *W. mirabilis*, by semi-quantitative RT-PCR (Reverse Transcription of mRNA followed by a PCR) on various plant tissues. Except for the cotyledons, *Welwitschia* plants only produce two large leaves that persist during their entire life. Individuals of both sexes display their reproductive structures on stems that emerge from the base of the leaves (Fig.2A). Such a structure is a compound strobilus or cone that consists of a primary axis with bracts at each node (Fig.2B-C). Axillary units arise in the axils of the bracts, so that the oldest units are found at the base of the cone and the newly formed units emerge at the tip. The axillary units of the female cone consist of three pairs of opposite bracts, which surround a central fertile ovule. In the male cones, the situation is more complex (Fig.2D): two pairs of opposite bracts surround a tubular structure, called antherophore, which bears six stalked synangia, i.e., the pollen producing organs. This antherophore surrounds an ovule, therefore resembling a flower but this ovule is sterile and only participates in the attraction of pollinating insects by producing a sugar droplet. Thus, the plant is functionally dioecious.

Fig.2 – Identification of the *W. mirabilis* tissues expressing *WelB1*, *WelB2*, *WelAG*, *WelBsister*, *WelAGL6*, *WelLFY* and *WelNdly*.



(A) *Welwitschia mirabilis* plant with female cones on stalk emerging from the base of the leaves (B) Female cone (C) Male cone (D) Close-up of an axillary unit of a male cone showing the bracts (b), the pollen-producing synangia (sy) fused into an antherophore and a sterile ovule (so) with its integument. One bract and half of the antherophore bearing 2 synangia have been removed in order to show the central ovule (E) Semi-quantitative RT-PCR profiles for *WelLFY*, *WelNdly* and the *Welwitschia* MADS-box genes identified. The *Welwitschia* actin gene (*WelActin*) is used as a control. PCR cycles are given on the left. Names of the different tissues used are indicated above each lane. Female tissues, male tissues and leaves are underlined with a pink, blue or green line respectively. M.W Frohlich took the pictures A to D.

Expression of both *WelB1* and *WelB2* were detected in male cones, including the cone tip where axillary units are just forming, but absent from female strobili (Fig.2E). *WelBsister* was expressed in female cones exclusively, especially in ovules but also in sterile bracts and at the tip of the cone bearing only very young axillary units. The C-class homolog, *WelAG*, was expressed in both male and female strobili and seems strongly associated to reproductive tissues, as no transcripts was detected in female bracts and cone tips where ovules are yet to develop. In contrast, *WelAGL6* expression was detected in all tissues examined, including the ever-growing leaves (Fig.2E).

Early expression of *WelLFY* and *WelNdly* during cone development and coincidence with B and C genes domains

To test whether *LFY-like* genes are candidates to regulate B and C genes in *W. mirabilis*, we analyzed *WelLFY* and *WelNdly* transcripts in the tissues expressing the MADS-box genes. The semi quantitative RT-PCR results showed that the two paralogs are widely expressed in male and female strobili reproductive tissues and to a weaker extent, in sterile appendages: mRNA corresponding to both genes are present in female bracts and *WelLFY* is also active in mature leaves (Fig.2E). These results indicate that *LFY-like* genes are both expressed in male tissues where B and C genes are activated but do not allow to detect any correlations between LFY-like and MADS-box genes. Therefore, we performed *in situ* hybridizations to characterize the spatiotemporal expression pattern of these genes during reproductive development.

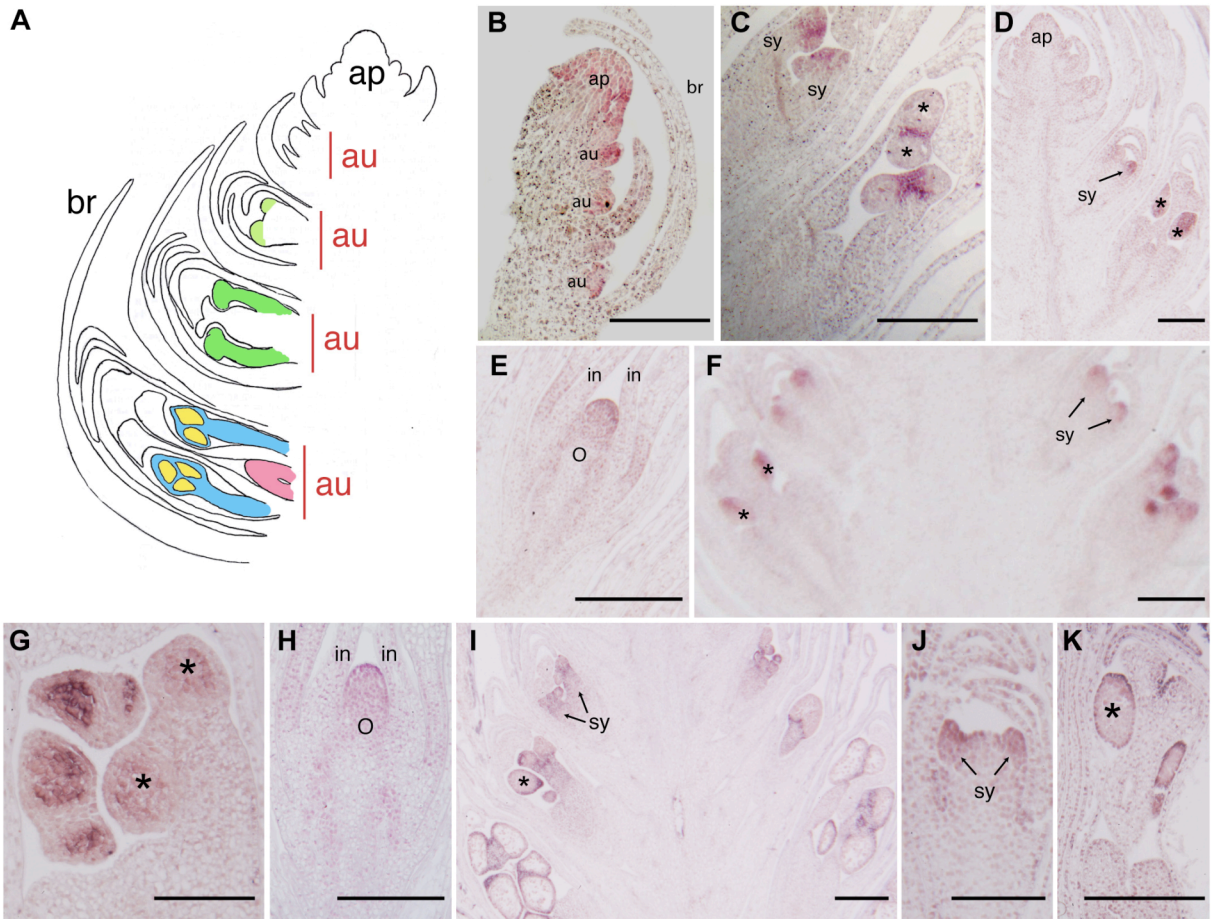
In male cones, expression of *WelLFY* was detected on the flanks of the cone apex where axillary units are about to form (Fig.3A-B) and in very young stages emerging from the apical dome (Fig.3B) whereas *WelNdly* transcripts were not detectable at the apex (Fig.3D). As secondary units develop, *WelLFY* signal disappeared from emerging bracts but remains high

in the primordia corresponding to the synangia (Fig.3C) and later in the upper part of the synangia as they elongate (Fig.3C). When the pollen-producing tissues start to differentiate (called the sporangia, there is 3 sporangia per synangium), the signal is still very strong but restricted to the tissue surrounding the pollen sacs and *WelLFY* is clearly excluded from the tissue that will later generate the pollen grains (Fig.3C). *WelNdly* transcripts are also visible in the upper part of the synangia primordia but the signal soon becomes restricted to the tissues that will form the sporangia (Fig.3D). Later, a signal is observed in at the top of the sterile ovule (Fig.3E).

B and C genes expression appear as axillary units develop. A strong *WelAG* signal is visible when primordia corresponding to the synangia start raising (Fig.3F). Interestingly, this signal is maintained as the primordia grow but it becomes restricted to the cells that will give rise to pollen grain, even before they adopt a morphology distinct from their neighbouring cells, therefore indicating the location of pollen-producing tissues before they become visible (Fig.3F-G). Additionally, *WelAG* transcripts are also present at the top of the sterile ovule (Fig.3H). *WelB1* and *WelB2* genes displayed nearly identical expression patterns, *WelB1* giving the strongest signal (Fig.3I): both genes are first expressed as male organs primordia become visible (Fig.3I-J) but no signal is observed in the center of the axillary unit where the ovule is going to emerge. The B signal is clearly visible as a gradual coloration along the stalk of the synangia to reach its maximum intensity at the top of the pollen-producing organs. However, as sporangia differentiate, B genes expression becomes restricted to the tissues where the sporangia are embedded, i.e., in the same tissue as the ones expressing the *LFY* ortholog (Fig.3I,K). The two *LFY-like* genes expression precede the expression of B and C genes and expression patterns of *WelLFY* and B genes on one hand and *WelNdly* and *WelAG* on the other hand are mutually exclusive in later stages of male cone development. Therefore, we postulated that *WelLFY* and *WelNdly* could regulate differentially the expression of B and C genes in *W. mirabilis*.

Due to the particular morphology of female cones with numerous bracts packed together, we did not succeed detecting in situ hybridization signals.

Fig.3 – Expression domains of *WelB1*, *WelB2*, *WelAG*, *WelLFY* and *WelNdly* in the male cones of *W. mirabilis*.



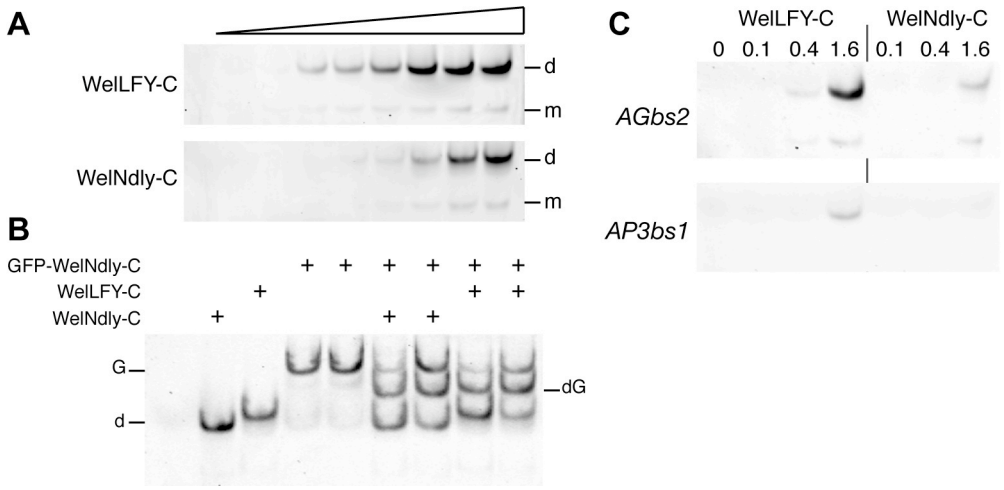
(A) Schematic representation of a longitudinal section of a male cone showing the apex (ap), the axillary units (au) and the bracts (br). The synangia primordia (light green) and young developing synangia (dark green) are indicated. In late developmental stages, the pollen-producing tissues (the sporangia, yellow) differentiate from the rest of the synangia (blue) and the sterile ovule (pink) starts to grow. *In situ* hybridization using probes against *WelLFY* (B,C), *WelNdly* (D,E), *WelAG* (F-H), *WelB1* (I) and *WelB2* (J,K) genes. All scales bars are 200 µm, except for G: 100 µm. ap, apex; br, bract; sy, synangia; o, sterile ovule; in, integument of the ovule. A (*) symbol indicates the sporangia location.

The DNA binding domains of *WelLFY* and *WelNdly* have different DNA binding specificities

To understand how LFY and NLY could regulate different genes, we tested the possibility that they differ in their DNA binding specificity. We produced recombinant versions of their DNA binding domain (WelLFY-C, residues 247-411 and WelNdly-C, residues 245-407). Both proteins appeared monomeric in solution (Table 1) and bind to *AP1bs1*, (a DNA probe bearing the *Arabidopsis* LFY binding site from *AP1* promoter) with a profile reminiscent of Arabidopsis LFY-C (Hames et al., 2008) (Fig.4A), suggesting they bind as dimer. We

confirmed this hypothesis by mixing WelNdly-C to a GFP-tagged WelNdly-C protein (Fig.4B): a single new complex of intermediate mobility formed corresponding to the WelNdly-C/GFP-WelNdly-C bound to *AP1bs1*. Taken together, these results strongly suggest that *Welwitschia* LFY homologs dimerize on DNA as observed with LFY-C from Arabidopsis. Mixing WelLFY-C and GFP-WelNdly-C also gave rise to a novel intermediate complex showing that WelLFY-C and WelNdly-C can heterodimerize *in vitro* (Fig.4B).

Fig.4 – DNA binding specificities of WelLFY-C and WelNdly-C.



(**A**) Electromobility shift assay (EMSA) with 10 nM fluorescent *AP1bs1* DNA and increasing concentrations of WelLFY-C or WelNdly-C. Protein concentrations from left to right are 0, 25, 50, 100, 200, 400, 800, 1500, 3000 and 5000 nM. The complexes with one monomer (m) or a dimer (d) are indicated. (**B**) EMSA with 10 nM fluorescent *AP1bs1* DNA and various combinations of WelLFY-C (400 nM), WelNdly-C (1.5 µM) and GFP-WelNdly-C (1.5 µM). WelLFY-C or WelNdly-C homodimer (d), GFP-WelNdly-C homodimer (G) and heterodimer involving a GFP and a non-GFP labelled protein (dG) are indicated. (**C**) EMSA with 1à nM *AGbs2* or 10 nM *AP3bs1* and three concentrations of WelLFY-C or WelNdly-C. Concentrations are given in µM above each lane.

As suggested by EMSA and confirmed by fluorescence anisotropy, the apparent affinity of WelLFY-C for *AP1bs1* is twice the affinity of WelNdly-C for the same DNA (Table 1). We also tested other Arabidopsis LFY binding sites (*AGbs2* and *AP3bs1*, located in the regulatory second intron of *AG* and in the upstream promoter of *AP3*, respectively) and found that WelLFY-C DNA preferences are closer to Arabidopsis LFY than WelNdly-C (Fig.4C).
It therefore seems that the two paralogs may have developed separate DNA binding specificities along with divergent expression pattern.

Table 1 – Characterization of the properties of WelLFY-C and WelNdly-C *in vitro*.

	Size-exclusion chromatography						Fluorescence anisotropy	
	Elution V (ml)		MW (kDa)		Nb. of molecules		K_D apparent (nM)	
	P1	P2	P1	P2	P1	P2	*AP1bs1*	*AGbs2*
WelLFY-C	88.5	91.5	24.0	18.6	1.05	0.81	350	1900
WelNLY-C	91.5	90	18.6	21.1	0.82	0.93	850	> 2500

The results of two independent (P1 and P2) size exclusion chromatography assays indicate that WelLFY-C and WelNdly-C are present as monomers in solution. Elution volume, corresponding molecular weight (MW) and the associated number of molecules are given in column. Apparent dissociation constants (K_D^{App}) of WelLFY-C and WelNdly-C bound to *AP1bs1* and *AGbs2*. Binding was modelled with a 1:1 equilibrium reaction.

WelLFY specifically recognizes B genes upstream sequences from *Welwitschia mirabilis*

We previously observed a correlation between B genes and *WelLFY* (but not *WelNdly*) expression. We wondered whether this could be linked to the different DNA binding preferences of the two paralogs. We thus isolated sequences upstream of the B genes coding region (2.8 kb for *WelB1* and 3.3 kb for *WelB2*). Since surface plasmon resonance allows to study the interaction between a transcription factor and a DNA regulatory region of several kilobases (Moyroud et al., 2009), we use it to test if WelLFY and WelNdly were able to interact specifically with the upstream sequences of *W. mirabilis* B genes. We recorded sensograms corresponding to the association and dissociation between recombinant versions of WelLFY and WelNdly and 3 DNA molecules (B genes promoter regions and a 2.2 kb genomic fragment from the *WelTubulin* gene as negative control). The sensogram curves were best fitted with a model that takes into account the existence of two types of sites (heterogenous ligand model), consistently with the presence in a promoter of a few high affinity sites among many low affinity sites.

Table 2 – Kinetic constants for interactions of WelLFYmin58 and WelNdlymin56 with *Welwitschia* genomic regions. bp, base pair; K_D^{App}, apparent dissociation constant.

		Length (bp)	$K_D^{App}1$ (µM)	$K_D^{App}2$ (µM)	χ^2
WelLFYmin58	*WelTubulin*	2147	1030	22.3	18.2
	WelB2	3377	235	0.045	2.6
	WelB1	2878	427	0.732	4.5
WelNLYmin56	*WelTubulin*	2147	796	0.157	14.2
	WelB2	3377	884	0.175	13.8
	WelB1	2878	2570	76.0	4.12

None of the two proteins was able to bind efficiently to the *WelTubulin* genomic locus (K_D^{App} best site ≈ 22 (M) and the quality of the fits was low (Chi2 > 14, a Chi2 < 10 is generally considered as satisfying), as it is often the case when only non-specific binding occurs (Table

2). On the contrary, some sites of high affinity for WelLFY ($K_D^{App} \approx 45$ nM) were detected in the upstream sequence of *WelB2*. This DNA region also appeared to display a binding site for WelNdly but with a lower affinity ($K_D^{App} \approx 175$ nM). Moreover, the low quality of the fit (Chi2=13.8) prevented us from establishing with confidence the presence of a site of high affinity for WelNdly in the upstream region of *WelB2* (Table 2). Similarly, we fail to detect a *cis*-element specifically bound by WelNdly in the promoter of *WelB1*, whereas the analysis indicates the presence of binding sites for WelLFY ($K_D \approx 730$ nM) with a good confidence (Chi2= 4.5) but a lower affinity than for the sites present in *WelB1* (Table 2). Taken together, these results strongly suggest that WelLFY interacts more efficiently than WelNdly with the upstream region of both B genes in *W. mirabilis*.

Discussion

WelLFY and *WelNdly* expression patterns are consistent with a role in reproductive organs.

Combining RT-PCR and in situ hybridization, we have shown that *WelLFY* and *WelNdly* are expressed in early stages of reproductive development consistent with a possible role in this process. *NLY* expression is absent from leaves suggesting its presence in young cones might contribute to confer its reproductive fate. A weak expression of *LFY* and/or *NLY* homologs also occurs in the shoot apex, vegetative buds or leaves of other gymnosperms such as Monterey pine (*Pinus radiata*), Japanese cedar (*Cryptomeria japonica*) and *Gnetum parvifolium* and their expression raises as the development progresses towards the reproductive phase (Mellerowicz et al., 1998; Mouradov et al., 1998a; Shindo et al., 2001; Shiokawa et al., 2008). It differs from Norway spruce (*Picea abies*), where *LFY* homologs are expressed at high level in both juvenile and reproductive tissues (Carlsbecker et al., 2004). In this latter case, it has been proposed that *DAL1*, the homolog of *AGL6* in this species, is expressed at higher levels in reproductive cones and could be responsible for triggering the reproductive fate. This does not seem to be the case in *Welwitschia* as *WelAGL6* was equally detected in leaves, and all male or female tissues examined.

Combined activities of B and C genes could specify male versus female reproductive organs in *Welwitschia*.

The identification of homologs of B and C genes in several gymnosperms led to the proposal that they could contribute to specify their reproductive structures (Theissen and Becker, 2004). In *Welwitschia mirabilis*, we found that *WelAG* is expressed in both male and female cones, whereas *WelB1* and *WelB2* are exclusively activated in male ones. This result

complements previous observations: in *Gnetum gnemon* (*G.gnemon*), Norway spruce (*Picea abies*) and cycas (*Cycas edentata*), C genes homologs are exclusively activated in the reproductive structures, independently of their sex, while B genes expression is limited to the male cones in *G.gnemon*, Norway spruce, Monterey pine and Japanese cedar (Tandre et al., 1995; Rutledge et al., 1998; Tandre et al., 1998; Mouradov et al., 1999; Winter et al., 1999; Becker et al., 2000; Fukui et al., 2001; Sundstrom and Engstrom, 2002; Jager et al., 2003; Zhang et al., 2004). Since this feature is shared among gymnosperms and angiosperms, it is likely that the function of these genes has been established in the last common ancestor of seed plants.

The case of *Welwitschia* is particularly interesting as each axillary unit of the male cone forms a central sterile ovule surrounded by pollen-producing organs. We found that *B* and *C* gene expression overlap in early male cone developmental stages but that *WelB1* and *WelB2* expressions stop once the pollen-producing organs are fully formed, while *WelAG* remains active during ovule development. This situation, also observed in *G.gnemon* (Becker et al., 2003) illustrates how, on the same cone, territories can form that strongly differ in identity and B/C gene expression, reminiscent of the situation in the angiosperm flower.

Based on the specific expression of the Bsister gene of *G.gnemon* in female individuals, it has been proposed that female identity could be specified by a combined expression of C+Bsister instead of C only (Becker et al., 2002b; Theissen and Becker, 2004). In *Welwitschia*, *WelBsister* expression is also restricted to the ovules of the female cones but, interestingly, it is absent from the sterile ovules developing on the male cones suggesting that Bsister genes are not essential for ovule formation but rather play a role in their later development (as shown for the *Bsister TRANSPARENT TESTA16* and *GORDITA*, in *Arabidopsis* or *FBP24* in petunia) (Prasad et al.; Nesi et al., 2002; de Folter et al., 2006).

Evidence of a pre-floral network in *W. mirabilis*.

LFY homologs are conserved in the four groups of land plants. To help understanding the origin of flowers, we investigated whether part of the network they orchestrate in angiosperms could have been established before flowers appeared. For this purpose, it is essential to establish the expression patterns of *LFY*, *NLY*, *B* and *C* homologs in a given species. Here, we show that *WelNdly* and *WelLFY* transcripts are detected at the tip of female cones just before expression of *WelAG* and *WelBsister* start and that *WelLFY* expression in the young male cone precedes that of *WelB1*, *WelB2* and *WelAG*. These early expression patterns are consistent with LFY-like proteins playing a role in B and C genes activation in *W. mirabilis*.

As described in three conifer species (Vazquez-Lobo et al., 2007), *WelLFY* and *WelNdly* display mutually exclusive expression patterns at more advanced developmental stages. However, these later territories differ between *Welwischia* and the conifers. Interestingly, *WelLFY* expression patterns are remarkably similar to those observed for *WelB1* and *WelB2,* whereas *WelNdly* and *WelAG* are expressed in similar domains. This observation suggests that each paralog could regulate these genes differentially within a pre-floral network. This hypothesis is supported by the SPR analysis revealing that WelLFY better binds genomic regions of *WelB1* and *WelB2*, than WelNdly does.

GGM15, GGM2 and *DAL11*, *DAL12*, *DAL13* are B genes homologs in *G.gnemon* and Norway spruce respectively and their expression is also restricted to the tissues surrounding the pollen-producing cells (Sundstrom et al., 1999; Becker et al., 2003). In Norway spruce, it is the *NLY* copy (*PaNLY*) that is expressed in these tissues and it would be interesting to test whether or not PaNLY is more able than PaLFY to recognize *DAL11*, *DAL12* and *DAL13* promoters. If it is the case, one can imagine that both proteins could bind to B genes promoters in the last common ancestor of seed plants and this property would have been retained by only one of the paralogs as gymnosperms diversified.

WelLFY and WelNdly evolved specific intrinsic properties.

A specialization of the role of *LFY* and *NLY* homologs in gymnosperms has long been postulated due to the divergent spatio-temporal expression of each paralog (Frohlich, 2000; Vazquez-Lobo et al., 2007). Unlike angiosperms homologs and despite the apparent conservation of the amino acids responsible for protein-DNA contacts (Hames et al., 2008), gymnosperms *LFY* expressed in Arabidopsis better rescues *lfy* mutants phenotype than *NLY* (Mouradov et al., 1998a; Maizel et al., 2005; Shiokawa et al., 2008) suggesting that the intrinsic properties of the two paralogs differ. Here, we established that WelLFY and WelNdly have indeed different DNA binding properties as WelLFY binds more efficiently than WelNdly to the *cis*-elements normally recognized by Arabidopsis LFY. This result is consistent with a yeast assay that showed that WelNdly failed to activate as efficiently as PRFLL (homolog of WelLFY in Monterey pine) a reporter gene under the control of the LFY binding sites from *AP1* or *AG* regulatory regions (Maizel et al., 2005).

The low affinity of WelNdly for the canonical LFY binding sites could be explained by differences in DNA binding specificities between NLY and LFY homologs. This way, each paralog could regulate its own set of targets genes during the development of the cones. Using SPR, we showed that WelLFY better recognizes than WelNdly the upstream region of two B

genes from *W. mirabilis* suggesting that WelLFY could be responsible for *WelB1* and *WelB2* induction in *Welwitschia* male primordia. Noticeably, the DNA-binding domains of WelLFY and WelNdly can interact *in vitro* to bind DNA so that a heterodimer could form *in vivo* when the expression domains of both paralogs largely overlap.

Towards development of functional tools in gymnosperms.

Many evolutionary scenarios have been proposed to shed the light on the molecular phenomenon that created the first flower (Frohlich, 2000; Albert et al., 2002; Theissen and Becker, 2004; Baum and Hileman, 2006). A modification of the behaviour of *LFY-like* genes and the homologs of floral homeotic regulators on the lineage leading to angiosperms has often been invoked. However, most of these hypotheses remain to be verified due to the lack of experimental techniques to study the function of these proteins in gymnosperms. In a previous study, we showed the usefulness of surface plasmon resonance to identify the promoters bound by a transcription factor of interest. Here, we combined this technique with a characterization of expression patterns and biochemical properties of the two *LFY* homologs from *W. mirabilis* to establish a first physical link between LFY homologs and B genes in gymnosperms. Our results bring new evidence supporting the existence a pre-floral genetic circuit governing the development of the reproductive structures in the last common ancestor of seed plants. Ultimately, establishing the function of each paralog in a gymnosperm species would allow testing if the observed protein-DNA interaction does result in the transcriptional activation of B genes. As experimental methods enabling the silencing of a given gene are becoming available in a broad range of flowering plants (Becker and Lange, 2010), it would be worth testing these techniques in gymnosperms and examining, for instance, the effect of a local suppression of LFY activity during male cone development. As 2 years old *Welwitschia* plants can already produce some cones, this species appears as an ideal gymnosperm organism to start with.

Methods

Plant material

Cones and leaves of *W. mirabilis* plants were collected in Pasadena (California) by MW.Frohlich. They were immediately frozen in liquid nitrogen and subseqeuntly used for mRNA extraction and *in situ* hybridization.

Isolation of cDNA clones

Total RNA was extracted using TRIzol reagent (Invitrogen) and 100 to 200 mg tissues from leaves or dissected cones (approximately 0.5 to 2 cm long) ground under liquid nitrogen. Samples (2 μg) of total RNA were treated with RNAse-free DNAse (Ambion) and quantified using a NanoDrop-ND100 spectrophotometer (NanoDrop Technologies). cDNA was synthesized from 1 μg of total RNA with

the RevertAid M-MuLV RT (Fermentas), using oligo-dT primers and other reaction components as described by the manufacturer.
WelAG, *WelB1*, *WelB2* and *WelBsister* were isolated by 3' and 5'-rapid amplification of cDNA ends (RACE) as described (Münster et al., 1997). Different oligonucleotides derived from previously identified gymnosperms MADS-box sequences (Rutledge et al., 1998; Tandre et al., 1998; Mouradov et al., 1999; Winter et al., 1999; Becker et al., 2000; Becker et al., 2002b; Jager et al., 2003; Zhang et al., 2004) were used as specific primers. The abbreviations of the ambiguous base positions follow the IUPAC nomenclature.

Primer name	Sequence
WelAG deg1F	ATGGGMCGHGGVAARATYGA
WelAG deg2F	AACMGACARGTHACWTTYTG
WelAG deg1R	TCYTGWTGDGCRTARTGRTG
WelB1 aF	CCGAGAAATTGGATGCAAGC
WelB1 bF	ACTGGATTAGAACCCTATGACGAC

Primer name	Sequence
WelBsis deg1F	ATGGGHMGAGGMAARATHGA
WelBsis deg2F	AAYAGRCARGTBACHTTYTC
WelBsis deg1R	TCYTGVAGRTTWGGYTGWGT
WelB2 aF	GTCATAGAGAGGTACAAGCAA
WelB2 bF	GATATGCAAAAAGAGAAAAATGAGAAT

In some cases, a second round of PCR with the TITANIUM Taq DNA polymerase (Clontech) and nested MADS-box specific primers was carried out on the primary amplification products. The resulting products of the RACE procedure were cloned with the TA cloning kit (Invitrogen) and sequenced. The complete sequences of *WelB1* and *WelB2* cDNA have been isolated. A partial sequence corresponding to *WelAG* (a few amino acid missing at both ends) and *WelBsister* (a few amino acid missing at the C-terminus) have been obtained.

Picea abies actin (ACP1972), *Picea wilsonii* tubulin (ABX57816) and *Gnetum gnemon AGL6* (*GGM13*, CAB44459) sequences were used to perform a BLAST search against a *Welwitschia* EST database (Albert et al., 2005) (http://blast.ncbi.nlm.nih.gov/). EST contigs assembly was performed with DNA BASER Sequence Assembler 2.6. *WelActin*, *WelAGL6,* and *WelNdly* (AF108227) sequences were then isolated by PCR with mixed cDNA from male and female cones using Phusion® DNA polymerase (Ozyme) and specific primers. *WelLFY* (AF109130) sequence was amplified with the same method but using TITANIUM Taq DNA polymerase (Clontech).

WelActin	ATGGCCGATGCTGAGGACATTCAA	CTAAAAGCACTTTCTGTGGACAATAG
WelAGL6	ATGGGTCGAGGCAGAGTTGAACT	TCAGACTACCCATCCATGCATGT
WelLFY	GCGGGATCCTCAGGGATGGCTCCTGAAAGTT	GGGAGGATCCTGTTCTTGACAAATCAAAGATGGGACTGCT
WelNdly	TAAAGCTTTCTGCATCCATGGTCG	CCCTCGAGTCAACATAATTTGCTTTT

The PCR products corresponding to *WelActin*, *WelAGL6, WelNdly* were then cloned with the Zero Blunt®PCR cloning kit (Invitrogen) and the PCR product corresponding to *WelLFY* was cloned with the TA cloning kit (Invitrogen) to create the pEDW72, pEDW58, pEDW87 and pEDW84 vectors respectively

Molecular Phylogenetic Analysis

Gnetum gnemon and *Gnetum parvifolium* MADS-box sequences (BAA85631, BAA85630, BAA85629, AC13993, CAC13991, CAC13992, CAB44457, CAB44459, CAB44455, CAB44449, CAB44448) were retrieved from GenBank and aligned with the isolated *Welwitschia* sequences. Alignments of predicted amino acid sequences were performed with MUSCLE 3.6 and the default settings (Edgar, 2004). The nucleotide sequences were then forced into an alignment based on the amino acid alignment using the PAL2NAL software (Suyama et al., 2006). After removing poorly aligned region and downweighting the third codon position transitions, phylogeny analyses were performed with the PAUP* v4.0 software using parsimony as the criterion for the shortest tree and in bootstrapping. 1000 bootstrap replicates were done.

Semi-quantitative RT-PCR

cDNA derived from about 50ng of DNAse-treated total RNA was used as template in each PCR. The PCRs were carried out in 50 ul and 5 ul samples were taken after 25, 30 and 35 cycles and were separated on 2% agarose gels stained with SYBR®Safe (Invitrogen). As a control, *W. mirabilis* actin cDNA was amplified under the same conditions but samples were taken after 20, 25 and 30 cycles. The following primers combinations were used:

Gene	Primer forward	
	Name	Sequence
WelLFY	WelLFYprobe2-F	CGCAAGATCGATGAAAATG
WelNdly	WelNdlyprobe2-F	GGACTTGCAGAGGCTTGAAC
WelAG	ProbWelAG-F	CGCTCTGCACAATCAACT
WelB1	WelB1aF	CCGAGAAATTGGATGCAAGC
WelB2	WelB2bF	GATATGCAAAAAGAGAAAAATGAGAAT
WelBsister	ProbWelBsister-F	GGAGCAGCAGCTGGAAACCGCAT
WelAGL6	ProbeWelAGL6-For	GGAAATCAACAAATCTCTACGCAAA
WelActin	WelActin-F	TTGCAATTCAGGCAGTTTTG

Gene	Primer reverse		Product size (nt)	Tm (°C)
	Name	Sequence		
WelLFY	WelLFYprobe2-R	GGAAACCCGCAAATCCGCT	204	56
WelNdly	WelNdlyprobe2-R	ATCGCAGTTTCTTTGCAGGT	179	56
WelAG	WelAGprofil-R	CCTGATGAGCATAGTGATGGGCA	340	58
WelB1	WelB1profil-R	GTAAGTATTATTTTGAAGGTTG	197	54
WelB2	ProbWelB2-R	CTGAACGTTGGGTTGCT	339	55
WelBsister	ProbeWelBsister-R	GGCTGAGTGGGTTGCAG	299	56
WelAGL6	oEDW-WelAGL6-Stop	TCAGACTACCCATCCATGCATG	271	55
WelActin	WelActin-R	GGCCACATATGCCAACTTCT	260	56

In situ hybridization

Following a fixation step in 4% paraformaldehyde, tissues were dehydrated and progressively embedded in Paraplast X-Tra™ (Tyco/Healthcare). Cones were cut as 7 µm tick slices and mount on glass slides positively charged (ProbeOn Plus, Fisher Biotech). Wax removal, proteinase K treatment (Invitrogen), paraformaldehde fixation and probe hybridization were performed according to J.Long protocol (www.its.caltech.edu/~plantlab/protocols/insitu.pdf). Probes were designed with the PrimerQuest software (http://eu.idtdna.com/Scitools/Applications/Primerquest/) against the most specific region of each gene of interest. For *WelLFY* and *WelNdly*, a probe corresponding to the variable sequence spanning exon 1 and 2 of each gene was amplified. For *Welwitschia* MADS-box gene, a probe matching the characteristic C-terminal end of each gene was used. Probes were synthesized according to (Drea et al., 2005): cDNA corresponding to *WelB1*, *WelB2* and *WelAG* cloned in pCR2.1 vectors (Invitrogen) and pEDW84 or pEDW87 were used as template in a PCR reaction with GoTaq DNA polymerase (Promega) and the following primers:

	Probe size	PCR primers	
		Name	Sequence (T7 promoter in bold)
WelLFY	291 bp	WLFinsR	CAAGTCTTTCCATGTAAGTCCT
		WLFinsLT7	**TAATACGACTCACTATAGGG**AGCCGAAAAGAAAAGGTTG
WelNdly	263 bp	WNDinsR2	GTTTGTCCTCTGTTCATCACTGT
		WelfxT7	**TAATACGACTCACTATAGGG**ACTTGCAGAGGCTTG
WelB1	225 bp	WelB1 aF	CCGAGAAATTGGATGCAAGC
		ProbWelB1 RT7	**TAATACGACTCACTATAGGG**CCACTCTCTGAAAGAGT
WelB2	339 bp	WelB2 bF	GATATGCAAAAAGAGAAAAATGAGAAT
		ProbWelB2 RT7	**TAATACGACTCACTATAGGG**CTGAACGTTGGGTTGCT
WelAG	341 bp	ProbWelAG F	CGCTCTGCACAATCAACT
		ProbWelAG RT7	**TAATACGACTCACTATAGGG**CCTGATGAGCATAGTGAT

The 5 PCR products were used as template for the *in vitro* transcription step with DIG RNA Labelling Kit (Roche) according to manufacturer's instructions. The presence of T7 promoter allows the synthesis of antisens probes. Probe quality and concentration were checked with a BioAnalyzer 2100 (Agilent).

Expression plasmids construction

WelLFY-C and WelNdly-C plasmids. Residues 247-411 from *WelLFY* cDNA and residues 245-407 from *WelNdly* cDNA were amplified from pEDW84 and pEDW87 respectively with the Phusion® DNA polymerase (Ozyme) and primers FL1001 (5'CCGCCATGGGGGAAGACAGGCAGAGGGAA3') and FL1002 (5'GGCTCGAGTCAAAGATGGGACTGCTTGCT3') or oFL1003 (5'GGGCCATGGGAGAGGAGAGACCCAGAGAA3') and oFL1004 (5'CCCTCGAGTCAACATAATTTGCTTTTTTCCAAGTGAC3') respectively, subcloned into pCR-Blunt (Invitrogen) and shuttled to pETM-11 (Dummler et al., 2005) as *Nco*I/*Xho*I fragment to yield the

pFLO3 and pFLO4 expression vectors, thereby allowing the production of WelLFY-C and WelNdly-C fused to a N-terminal 6 x histidine tag.

GFP-WelNdly-C plasmid. A GFP fragment was amplified from pBS-GLFY plasmid obtained from X Wu (Wu et al., 2003) using primers oETH1001 (5'CCCACTACTGAGAATCTTTATTTTCAGGGCCAGTTCAG3') and oETH1002 (5'CCCAAACCACTACCTCCGTTGCCGTTATCCTGTTTGTATAGTTCATCCAT3'). The amplified fragment was subsequently used as a megaprimer to amplify plasmid pFLO4 and yield pETH25.

WelLFYmin58 and WelNdlymin56 plasmids. Residues 58-411 from *WelLFY* cDNA and residues 56-407 from *WelNdly* cDNA were amplified from pEDW84 and pEDW87 with Phusion® DNA polymerase (Ozyme) and primers 5'ACACATATGAAGGAAATGGTTTGCCTAGAGGAGC3' and 5'GGCTCGAGAAGATGGGACTGCTTGCT3' or 5'TTACATATGAAGGATCTGAAATCGCTTGAAGAT3' and 5'CCCTCGAGACATAATTTGCTTTTTTCCA3' respectively, subcloned into pCR-Blunt (Invitrogen) and shuttled to pETM-30a+ plasmid as a *Nde*I/*Xho*I fragment to yield the pEDW29 and pEDW49 expression vectors.

Protein expression and purification

The vectors pFLO3 and pFLO4 were used to produce and purify recombinant WelLFY-C and WelNdly-C proteins according to the protocol described for LFY-C from *A.thaliana* (Hamès et al., 2008).

WelLFYmin58 and WelNdly56 were expressed using *Escherichia coli* strain Rosetta™ 2(DE3)pLys (Novagen). After induction by 0.5 mM IPTG, cells were grown overnight at 17°C. The pellet corresponding to 1 l culture was sonicated in 50 ml lysis buffer (50 mM Tris-HCl pH8, 5 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and one protease inhibitor cocktail table Complete EDTA-free (Roche) and centrifuged for 45 minutes at 16500 rpm. The supernatant was loaded on a column with 1 ml Ni-NTA resin (Quiagen), then washed with 35 ml of wash buffer (50 mM Tris-HCl pH8, 20 mM imidazole, 5 mM TCEP) and eluted with the same buffer containing 350 mM Imidazole instead. The fractions containing the protein were pooled and dialysed overnight in 50 mM Tris-HCl pH8, 5 mM TCEP buffer at 4°C. The dialysed proteins were exposed to a salt shock with 0.8 M NaCl to remove the bacterial DNA bound to the protein and the Ni-NTA purification step was repeated as previously except that the 0.8 M NaCl was added to the wash and elution buffers. The fractions containing the protein were pooled and applied to a Hi-load Superdex-200 16/60 prep grade column (GE Healthcare) equilibrated with 20 mM Tris-HCl pH8, 0.8 M NaCl and 5 mM TCEP to eliminate aggregated protein by size exclusion chromatography. Proteins concentrations were estimated using the Bradford assay (Bradford, 1976).

EMSA assay

Fluorescent EMSA assays were performed as described (Hamès et al., 2008). The DNA motifs *AP1bs1*, *AGbs2* and *AP3bs1* were generated by annealing of single-stranded oligonucleotides, 5'-labeled with TAMRA (Sigma), to non-fluorescent complementary oligonucleotides in annealing buffer (10 nM Tris pH7.5, 150 mM NaCl and 1 mM EDTA). The following sequences were used:

	5'-labeled TAMRA	Non-fluorescent
AP1bs1	TTGGGGAAGGACCAGTGGTCCGTACAATGT	ACATTGTACGGACCACTGGTCCTTCCCCAA
AGbs2	TGGATTTATACCCAATGTGTTAATGGGTTGT	ACAACCCATTAACACATTGGGTATAAATCCA
AP3bs1	CCTTCTTAAACCCTAGGGGTAATATTCTAT	ATAGAATATTACCCCTAGGGTTTAAGAAGG

For each reaction, 10 nM fluorescent dsDNA was incubated with WelLFY-C or WelNdly-C in 20 ul binding buffer (20 mM Tris–HCl pH 7.5, 150 mM NaCl, 1% glycerol, 0.25 mM EDTA, 2 mM MgCl2, 28 ng/ml fish sperm DNA (Roche) and 3 mM DTT). After 5 min incubation on ice, binding reactions were loaded onto native 6% polyacrylamide gels 0.5X TBE and electrophoresed at 90 V for 90 min at 4°C. Gels were scanned on a Typhoon 9400 scanner (Molecular Dynamics, Sunnyvale, CA; excitation light 532 nm, emission filter 580 BP 30).

Fluorescent anisotropy

Equilibrium binding of WelLFY-C and WelNdly-C to *AP1bs1*, *AGbs2* and *AP3bs1* were monitored by fluorescence anisotropy (LeTilly and Royer, 1993). Serial dilutions of recombinant proteins were

mixed with 10 nM of TAMRA-labeled (same as the one used for EMSA assays) in a binding buffer containing 20 mM Tris pH 7.5, 0.25 mM EDTA, 1 mM DTT, 150 mM NaCl, 2 mM $MgCl_2$, 1%v/v glycerol and 1 μM dIdC. Anisotropy was measured on a Safire^2 microplate reader (TECAN) with excitation and emission wavelengths of 530 and 585 nm, respectively. Each fluorescence anisotropy data point represents the average of 100 reading. Anisotropy values were automatically calculated with the data reduction software magellan V6.43 TECAN). The apparent equilibrium dissociation constant was determined by plotting anisotropy as a function of protein concentration, fitting the data to a 1:1 binding model and nonlinear least square analysis with the KaleidaGraph software (Synergy Software).

Cloning of the genomic upstream sequences of *WelB1* and *WelB2*

We use a P^{32} labelled DNA probe corresponding to the *WelB1* and *WelB2* PCR products (see *in situ* hybridization) to probe four lambda genomic sublibraries as in (Frohlich and Meyerowitz, 1997).
The upstream sequence of both corresponding genes was retrieved by PCR with Phusion® DNA polymerase (Ozyme), using the selected lambda phages plaque as a template and primer Lambda786 matching within the lambda vector (5'CGGAGTGGCTCACAGTCGGTGGTCCGGCAGTACAA3') and a gene-specific primer matching either a region within the K-domain (WelB1screenR2, 5'CGGAGTTTGGAGTCCAGCTTTTCCTTTTCG3') or the end of the coding sequence (WelB2screenR3, 5'TAAGAACCGTAACTGAACGTTGGGTTG3'). We obtained two DNA fragments of 3.7 kb and 8 kb respectively. The smaller PCR product contains the first two exons of *WelB1* and 2.3kb upstream the ATG. The larger fragment corresponds to the complete genomic sequence of the *WelB2*, confirming the identity of the gene, as well as 5.9 kb of sequence upstream to the start codon. Both PCR fragment were cloned in pCRBlunt plasmid vector to yield the new vectors pEDW131 and pEDW130, containing the *WelB1* or *WelB2* upstream sequence respectively. The *WelTubulin* genomic locus was amplified with Phusion® DNA polymerase (Ozyme) using genomic DNA previously prepared (Frohlich, 2000) as a template and the primers oEDW-WelTubulin-gene-F1 (5'GCCTTTAAACGACTTCTGTAAATA3') and oEDW-WelTubulin-gene-R1 (5'GGCTAACACAACAACAGAAGCAGAT3') designed to match the WelTubulin sequence previously identified (cf. Isolation of cDNA).

SPR analysis of DNA-protein interaction

Double-stranded DNA molecules for SPR analysis were synthesized by PCR amplification from pEDW131, pEDW130 and pEDW64 plasmids using the following primers, one primer of each pair being biotinylated:

	5'-Biotin	Non-labeled
WelB1	CGGATCTGGGTCGACTCTAGGCCT	CCAGTCCACGCAAATAATCAGA
WelB2	GTAAAACGACGGCCAG	GGAGTTCTTGCTTAGAAGATCAC
WelTubulin	GGAAACAGCTATGACCATG	AGTGAACAAATGCCCGCTTGG

PCR products were purified using a NucleoSpin® Extract II kit (Macherey-Nagel) and quantified with a NanoDrop-ND100 spectrophotometer (NanoDrop Technologies). Chip immobilisation and sensograms recording according to (Moyroud et al., 2009). Real-time SPR interaction curves were analysed using BIAevaluation v4 software (Biacore). For each analysis, the response of the reference channel was subtracted from the interaction curves obtained from the 3 experimental channels. These normalized curves were fitted globally to a heterogenous ligand interaction model, permitting the determination of two apparent dissociation constants (K_D^{App}), corresponding to the presence of two types of binding sites. The validity of the interaction model was verified by data fitting, with a good fit indicated by $\chi^2<10$.

Acknowledgments

We would like to thank Nicole Maturen and Kate Warner (Natural History Museum, London) for gene cloning advice, Karen James (Natural History Museum, London) for help with the in situ hybridization. Mathieu Reymond, Charlie Scutt from the RDP lab (ENS Lyon) and Annie

Chaboud (IBCP Lyon) are acknowledged for their advice related to the SPR experiments. We also would like to thank Renaud Dumas (LPCV, CEA Grenoble) for the help with recombinant protein purification.

References

Albert, V.A., Oppenheimer, D.G., and Lindqvist, C. (2002). Pleiotropy, redundancy and the evolution of flowers. Trends Plant Sci. **7,** 297-301.

Albert, V.A., Soltis, D.E., Carlson, J.E., Farmerie, W.G., Wall, P.K., Ilut, D.C., Solow, T.M., Mueller, L.A., Landherr, L.L., Hu, Y., Buzgo, M., Kim, S., Yoo, M.J., Frohlich, M.W., Perl-Treves, R., Schlarbaum, S.E., Bliss, B.J., Zhang, X., Tanksley, S.D., Oppenheimer, D.G., Soltis, P.S., Ma, H., DePamphilis, C.W., and Leebens-Mack, J.H. (2005). Floral gene resources from basal angiosperms for comparative genomics research. BMC Plant Biol **5,** 5.

Baum, D.A., and Hileman, L.C. (2006). A developmental genetic model for the origin of the flower. In: Flowering and its manipulation—Ainsworth C, ed., Sheffield, UK: Blackwell Publishing. 3-27.

Becker, A., Saedler, H., and Theissen, G. (2003). Distinct MADS-box gene expression patterns in the reproductive cones of the gymnosperm Gnetum gnemon. Dev Genes Evol **213,** 567-572.

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

Becker, A., Bey, M., Burglin, T.R., Saedler, H., and Theissen, G. (2002a). Ancestry and diversity of BEL1-like homeobox genes revealed by gymnosperm (Gnetum gnemon) homologs. Dev Genes Evol **212,** 452-457.

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

Busch, M.A., Bomblies, K., and Weigel, D. (1999). Activation of a floral homeotic gene in Arabidopsis. Science **285,** 585-587.

Carlsbecker, A., Tandre, K., Johanson, U., Englund, M., and Engstrom, P. (2004). The MADS-box gene DAL1 is a potential mediator of the juvenile-to-adult transition in Norway spruce (Picea abies). Plant J **40,** 546-557.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134,** 25-36.

Chaw, S.M., Parkinson, C.L., Cheng, Y., Vincent, T.M., and 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 U S A **97,** 4086-4091.

Crane, P.R. (1996). The fossil history of the gnetales. Int. J. Plant Sci. **157,** S50-S57.

de Folter, S., Shchennikova, A.V., Franken, J., Busscher, M., Baskar, R., Grossniklaus, U., Angenent, G.C., and Immink, R.G. (2006). A Bsister MADS-box gene involved in ovule and seed development in petunia and Arabidopsis. Plant J **47,** 934-946.

Dornelas, M.C., and Rodriguez, A.P. (2006). The tropical cedar tree (Cedrela fissilis Vell., Meliaceae) homolog of the Arabidopsis LEAFY gene is expressed in reproductive tissues and can complement Arabidopsis leafy mutants. Planta **223,** 306-314.

Drea, S., Corsar, J., Crawford, B., Shaw, P., Dolan, L., and Doonan, J.H. (2005). A streamlined method for systematic, high resolution in situ analysis of mRNA distribution in plants. Plant Methods **1,** 8.

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

Frohlich, M.W., and 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**.

Frohlich, M.W., and Chase, M.W. (2007). After a dozen years of progress the origin of angiosperms is still a great mystery. Nature **450,** 1184-1189.

Frohlich, M.W., and Parker,D.S. (2000). The mostly male theory of flower evolutionary origins: from genes to fossils. Systematic Botany **25,** 155-170.

Fukui, M., Futamura, N., Mukai, Y., Wang, Y., Nagao, A., and Shinohara, K. (2001). Ancestral MADS box genes in Sugi, Cryptomeria japonica D. Don (Taxodiaceae), homologous to the B function genes in angiosperms. Plant Cell Physiol **42,** 566-575.

Guo, C.L., Chen, L.G., He, X.H., Dai, Z., and Yuan, H.Y. (2005). [Expressions of LEAFY homologous genes in different organs and

stages of Ginkgo biloba]. Yi Chuan **27,** 241-244.

Hames, C., Ptchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gerard, F., Martiel, J.L., Benlloch, R., Parcy, F., and Muller, C.W. (2008). Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. Embo J **27,** 2628-2637.

Jager, M., Hassanin, A., Manuel, M., Le Guyader, H., and Deutsch, J. (2003). MADS-box genes in Ginkgo biloba and the evolution of the AGAMOUS family. Mol Biol Evol **20,** 842-854.

Lamb, R.S., Hill, T.A., Tan, Q.K., and Irish, V.F. (2002). Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. Development **129,** 2079-2086.

LeTilly, V., and Royer, C.A. (1993). Fluorescence anisotropy assays implicate protein-protein interactions in regulating trp repressor DNA binding. Biochemistry **32,** 7753-7758.

Litt, A., and Irish, V.F. (2003). Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. Genetics **165,** 821-833.

Lohmann, J.U., Hong, R.L., Hobe, M., Busch, M.A., Parcy, F., Simon, R., and Weigel, D. (2001). A molecular link between stem cell regulation and floral patterning in Arabidopsis. Cell **105,** 793-803.

Maizel, A., Busch, M.A., Tanahashi, T., Perkovic, J., Kato, M., Hasebe, M., and Weigel, D. (2005). The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science **308,** 260-263.

Mellerowicz, E.J., Horgan, K., Walden, A., Coker, A., and Walter, C. (1998). PRFLL--a Pinus radiata homologue of FLORICAULA and LEAFY is expressed in buds containing vegetative shoot and undifferentiated male cone primordia. Planta **206,** 619-629.

Mouradov, A., Hamdorf, B., Teasdale, R.D., Kim, J.T., Winter, K.U., and Theissen, G. (1999). 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,** 245-252.

Mouradov, A., Glassick, T., Hamdorf, B., Murphy, L., Fowler, B., Marla, S., and Teasdale, R.D. (1998a). NEEDLY, a Pinus radiata ortholog of FLORICAULA/LEAFY genes, expressed in both reproductive and vegetative meristems. Proc Natl Acad Sci U S A **95,** 6537-6542.

Mouradov, A., Glassick, T.V., Hamdorf, B.A., Murphy, L.C., Marla, S.S., Yang, Y., and Teasdale, R.D. (1998b). Family of MADS-Box genes expressed early in male and female reproductive structures of monterey pine. Plant Physiol **117,** 55-62.

Moyroud, E., Reymond, M.C., Hames, C., Parcy, F., and Scutt, C.P. (2009). The analysis of entire gene promoters by surface plasmon resonance. Plant J **59,** 851-858.

Moyroud, E., Kusters, E., Monniaux, M., Koes, R., and Parcy, F. (2010). LEAFY blossoms. Trends Plant Sci.

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

Parcy, F., Nilsson, O., Busch, M.A., Lee, I., and Weigel, D. (1998). A genetic framework for floral patterning. Nature **395,** 561-566.

Prasad, K., Zhang, X., Tobon, E., and Ambrose, B.A. The Arabidopsis B-sister MADS-box protein, GORDITA, represses fruit growth and contributes to integument development. Plant J **62,** 203-214.

Rutledge, R., Regan, S., Nicolas, O., Fobert, P., Cote, C., Bosnich, W., Kauffeldt, C., Sunohara, G., Seguin, A., and 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,** 625-634.

Shindo, S., Sakakibara, K., Sano, R., Ueda, K., and Hasebe, M. (2001). Characterization of a FLORICAULA/LEAFY homologue of Gnetum parvifolium and its implications for the evolution of reproductive organs in seed plants. J Plant Science **162,** 1199-1209.

Shiokawa, T., Yamada, S., Futamura, N., Osanai, K., Murasugi, D., Shinohara, K., Kawai, S., Morohoshi, N., Katayama, Y., and Kajita, S. (2008). Isolation and functional analysis of the CjNdly gene, a homolog in Cryptomeria japonica of FLORICAULA/LEAFY genes. Tree Physiol **28,** 21-28.

Sundstrom, J., and Engstrom, P. (2002). Conifer reproductive development involves B-type MADS-box genes with distinct and different activities in male organ primordia. Plant J **31,** 161-169.

Sundstrom, J., Carlsbecker, A., Svensson, M.E., Svenson, M., Johanson, U., Theissen, G., and Engstrom, 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,** 253-266.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein

sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34,** W609-612.

Tandre, K., Albert, V.A., Sundas, A., and Engstrom, 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., and Engstrom, P. (1998). Conservation of gene structure and activity in the regulation of reproductive organ development of conifers and angiosperms. Plant J **15,** 615-623.

Theissen, G., and Becker, A. (2004). Gymnosperms orthologs of class B floral homeotic genes and their impact on understanding flower origin. Critical Reviews in Plant Science **23,** 129-148.

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

Vazquez-Lobo, A., Carlsbecker, A., Vergara-Silva, F., Alvarez-Buylla, E.R., Pinero, D., and Engstrom, P. (2007). Characterization of the expression patterns of LEAFY/FLORICAULA and NEEDLY orthologs in female and male cones of the conifer genera Picea, Podocarpus, and Taxus: implications for current evo-devo hypotheses for gymnosperms. Evol Dev **9,** 446-459.

Winter, K.U., Saedler, H., and Theissen, G. (2002). On the origin of class B floral homeotic genes: functional substitution and dominant inhibition in Arabidopsis by expression of an orthologue from the gymnosperm Gnetum. Plant J **31,** 457-475.

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

Wu, X., Dinneny, J.R., Crawford, K.M., Rhee, Y., Citovsky, V., Zambryski, P.C., and Weigel, D. (2003). Modes of intercellular transcription factor movement in the Arabidopsis apex. Development **130,** 3735-3745.

Zhang, P., Tan, H.T., Pwee, K.H., and Kumar, P.P. (2004). Conservation of class C function of floral organ development during 300 million years of evolution from gymnosperms to angiosperms. Plant J **37,** 566-577.

Résultats complémentaires

Afin d'apporter des éléments de réponse aux interrogations soulevées par les premiers résultats obtenus, nous avons entrepris plusieurs approches expérimentales exposées ici.

Phylogénie des gènes MADS de gymnospermes

En 1995, le groupe de Peter Engström a identifié les premiers gènes à boîte MADS chez l'épicéa commun (*Picea abies*), démontrant ainsi l'existence de ces gènes dans les génomes de gymnospermes (Tandre et al., 1995). Des représentants de la famille MADS ont depuis été isolés chez le Ginkgo, le genre *Gnetum* et plusieurs espèces de Cycadophytes. L'étude des relations de parenté entre les gènes à boîte MADS des plantes à graines a permis d'établir que plusieurs sous-familles de gènes à boîte MADS incluaient à la fois des séquences angiospermes et gymnospermes. Nous avons extrait des bases de données publiques les 52 séquences gymnospermes appartenant à ces sous-groupes et nous les avons combinées aux séquences obtenues chez *Welwitschia* pour constituer un set de données que nous utiliserons pour réaliser une analyse phylogénétique complète.

Influence du domaine N-terminal sur les propriétés d'interaction de WelLFY et WelNdly

Afin de tester si le domaine N-terminal modifie le comportement de WelLFYmin58 et WelNdlymin56, nous avons amorcé l'étude des propriétés biochimiques de ces deux protéines.

De façon à analyser leur degré de multimérisation, nous avons réalisé une filtration sur gel pour ces deux protéines recombinantes. WelLFYmin58 s'élue dans un volume entre 76.5 et 74.8 ml et WelNdlymin56 entre 55.2 et 56.5 ml, ce qui permet d'exclure que les deux protéines soient monomériques en solution mais ne permet pas de déterminer avec précision la taille des complexes. Renaud Dumas et Camille Sayou, étudiante M2 dans notre équipe, ont utilisé la technique du SEC-MALLS (Size Exclusion Chromatography Multi-Angle Laser Light Scattering) pour étudier l'état de multimérisation d'homologues de LFY chez d'autres espèces. La présence du domaine N-terminal semble induire une dimérisation ou une tétramérisation de la protéine en solution selon les espèces. Dès lors, il est possible que les deux paralogues de *Welwitschia* possèdent une capacité différente de multimérisation. L'application de la technique du SEC-MALLS à WelLFYmin58 et WelNdlymin56 devrait permettre de tester cette hypothèse.

En parallèle, nous avons utilisé le retard sur gel pour tester la capacité de WelLFYmin58 et WelNdlymin56 à reconnaître *AP1bs1* (Fig.4.1A). Comme observé avec les domaines C-terminaux des deux protéines, WelLFYmin58 présente une meilleure affinité pour ce motif ADN que son paralogue ; la présence du domaine N-terminal ne modifierait donc pas les préférences de liaison de chaque paralogue de manière drastique. En revanche, un seul type de complexe protéine-ADN semble se former ; ce complexe implique probablement le même nombre de protéines quel que soit le paralogue utilisé car les distances de migration observées sont identiques. Ainsi, même s'il s'avère que WelLFYmin58 et WelNdlymin56 ont des degrés de multimérisation différents en solution, les complexes constitués avec l'ADN pourraient être similaires.

Spécificité de liaison propre à chaque paralogue chez *Welwitschia*

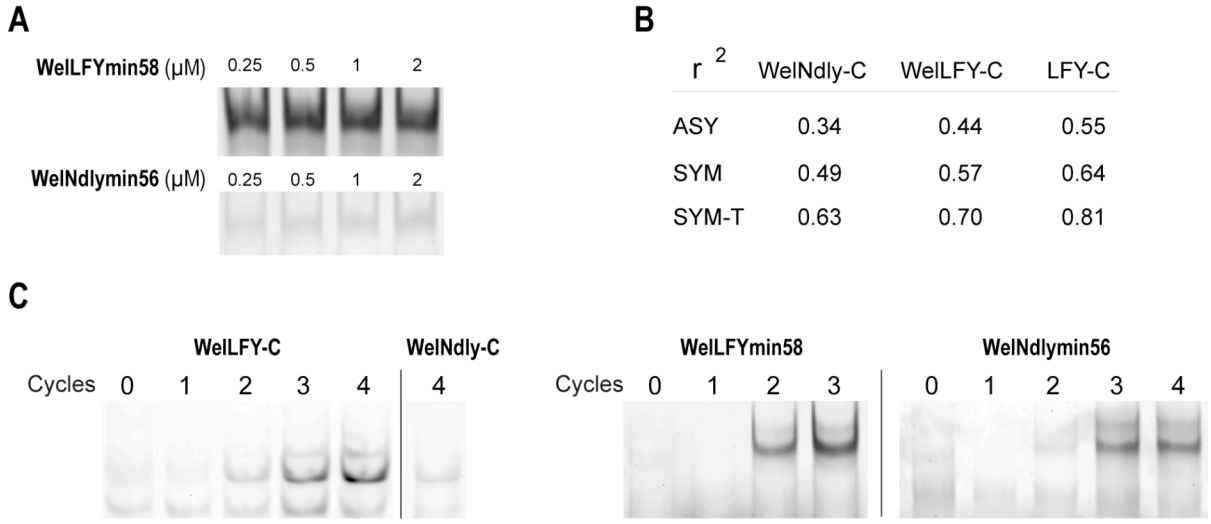
WelLFY et son paralogue WelNdly semblent se comporter différemment vis-à-vis de l'ADN. Il reste maintenant à établir s'il s'agit d'une différence d'affinité ou si les spécificités de liaison des deux paralogues ont divergé. Dans le premier cas, les deux protéines présenteraient les mêmes préférences de liaison mais WelNdly aurait une affinité globale moindre pour l'ADN (par exemple en établissant moins d'interaction avec les groupements phosphates des nucléotides). Si la seconde hypothèse est vraie en revanche, alors les deux protéines contacteraient l'ADN aussi efficacement l'une que l'autre mais leurs préférences iraient à des motifs différents.

Les résultats présentés dans l'article préliminaire ont montré que WelLFY-C reconnaissait mieux que WelNdly-C trois éléments *cis* retrouvés dans les régions régulatrices d'*AP1*, *AG* et *AP3*. Afin de comparer la capacité des domaines de liaison à l'ADN des deux paralogues sur un plus grand échantillon, nous avons mesuré par un essai QuMFRA (cf. Chapitre 2) l'affinité relative de chacune des deux protéines pour 26 motifs ADN de 30 paires de bases chacun. La protéine LFY-C a été utilisée comme témoin. Les valeurs mesurées ont ensuite été confrontées aux valeurs prédites par les trois matrices établies pour LFY-C. Les coefficients de détermination obtenus indiquent que comme LFY-C, le modèle SYM-T explique mieux les résultats observés que les deux autres matrices qui ne tiennent pas compte de la dépendance (Fig.4.1B). Cependant, les coefficients de détermination obtenus pour WelNdly-C (0.63) et WelLFY-C (0.70) sont inférieurs à celui estimé pour LFY-C (0.81) ce qui suggère que les protéines de *Welwitschia*, en particulier WelNdly, présentent des préférences de liaison un peu différentes. Cependant, la corrélation observée n'est pas négligeable et WelNdly

reconnaît la plupart des séquences efficacement liées par LFY-C, le motif préféré par WelNdly-C est donc probablement assez proche du motif favori de son homologue chez *Arabidopsis*.

Figure 4.1 | Caractérisation des spécificités de liaison de WelLFY et WelNdly.

(A) Retard sur gel en présence de 10 nM *AP1bs1* et une concentration croissante de WelLFYmin58 ou WelNdlymin56 (B) Coefficient de détermination entre les scores mesurés (-Ln(K_D relatif)) avec chaque protéine et les scores prédits par chacune des trois matrices ASY, SYM et SYM-T (cf. Chapitre 2). (C) Sélection des séquences reconnues par WelLFY-C, WelNdly-C, WelLFYmin58 ou WelNdlymin56 au cours des cycles Selex successifs (indiqués au-dessus de chaque piste). Pour chaque piste, 400 nM de protéine sont incubés avec 10 nM d'ADN double brin fluorescent correspondant à chaque cycle Selex.



Afin de décrire de manière exhaustive les préférences des protéines de *W.mirabilis*, nous avons réalisé un Selex sur WelLFY-C, WelNdly-C, WelLFYmin58 et WelNdlymin56. Dans chacun des cas, un groupe d'ADN double brin présentant une forte affinité pour la protéine testée à été obtenu après 3 à 4 tours d'enrichissement (Fig.4.1C). Le séquençage des aptamères sélectionnés vient d'être réalisé (cf. Table 2-5, Chapitre 2) et les séquences obtenues permettront de construire un modèle de liaison propre à chaque homologue. Ces modèles seront ensuite comparés à la matrice SYM-T afin d'établir si les spécificités de liaison ont divergé entre les différents homologues.

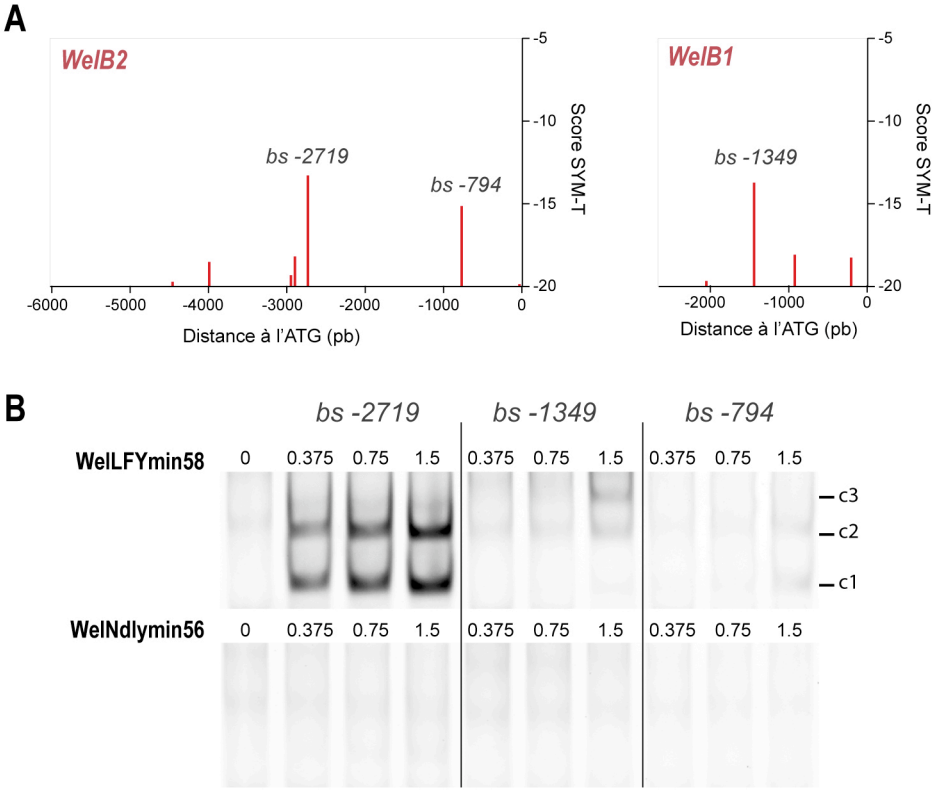
Identification des éléments *cis* reconnus par WelLFY/WelNdly

Les résultats de SPR suggèrent qu'il existe des éléments *cis* spécifiquement reconnus par *WelLFY* au sein des gènes B de *Welwitschia*. En l'absence de test fonctionnel, identifier les éléments *cis* responsables du phénomène observé et confirmer par une méthode indépendante leur capacité à être reconnus préférentiellement par *WelLFY* contribuerait à étayer les résultats de SPR.

Ne disposant pas encore des matrices associées aux deux protéines de *Welwitschia*, nous avons utilisé la matrice SYM-T pour essayer de détecter des sites de liaison potentiels au sein des régions génomiques situées en amont de *WelB1* et *WelB2* (Fig.4.2A). Les trois meilleurs motifs identifiés ont ensuite été individuellement testés par retard sur gel en présence de WelLFYmin58 ou WelNdlymin56 (Fig.4.2B). L'élément *bs-2719*, situé dans la région génomique de *WelB2* utilisée pour la SPR, s'est révélé particulièrement bien reconnu par WelLFYmin58 et deux complexes (c1 et c2) ADN-protéine sont formés. Les deux autres éléments testés présentent une affinité moindre pour WelLFYmin58, puisqu'un retard d'intensité plus faible est observé pour des concentrations de protéine plus importantes. Il est intéressant de noter qu'un troisième complexe, migrant moins loin dans le gel d'acrylamide, semble se former en présence de *bs-1349*. La protéine WelNdlymin56 en revanche, semble incapable de contacter les trois ADNs aux concentrations testées.

Figure 4.2 | Eléments *cis* reconnus par WelLFYmin58 dans les régions génomiques en amont des gènes B.

(A) Prédictions des scores des régions génomiques en amont du codon start de *WelB2* et *WelB1* avec la matrice SYM-T (pb, paires de bases depuis l'ATG). Les 3 meilleurs sites prédits ont été nommés *bs-2719*, *bs-794* (*WelB2*) et *bs-1349* (*WelB1*) d'après leur position par rapport au codon start. (B) Retard sur gel avec WelLFYmin58 ou WelNdlymin56 en présence de 10 nM d'ADN fluorescent correspondant aux sites *bs*. Trois complexes ADN/protéines différents, nommés c1, c2 et C3 sont observés.



Les résultats d'hybridation *in situ* montrent une bonne coïncidence des domaines d'expression de *WelNdly* et *WelAG* au cours des stades avancés du développement des cônes. Afin de tester si cette coexpression découle d'une régulation NEEDLY-spécifique de l'homologue d'*AG* chez *Welwitschia*, nous avons isolé l'intron suivant le premier exon de *WelAG* par criblage des banques d'ADN génomiques de *W. mirabilis*. Cet intron est particulièrement grand (15kb) comme cela est souvent observé pour les orthologues du gène (3kb pour *AG*, 6kb pour l'homologue d'*AG* chez *Amborella*, plusieurs kilo bases chez *Cycas*). Le comportement des deux paralogues sur cet intron pourra être étudié par SPR et/ou par une recherche de sites individuels à l'aide des matrices WelLFY et WelNdly spécifiques.

Conclusion

Un réseau pré-floral, impliquant LEAFY et son paralogue NEEDLY pourrait fonctionner chez les gymnospermes. Les deux paralogues remplissent probablement des fonctions distinctes au sein de ce circuit génétique, notamment quant à l'activation des gènes B. D'autres travaux seront nécessaires chez *Welwitschia* et autres espèces gymnospermes pour obtenir une image précise des fonctions de ce réseau pré-floral et des conséquences de son remaniement sur la lignée menant aux angiospermes. Les données que nous avons obtenues montrent que des méthodes récentes d'étude des interactions protéines/ADN peuvent apporter des informations pertinentes sur des protéines méconnues. Nous espérons que cette analyse participera au développement d'approches nouvelles pour étudier un groupe incontournable de l'histoire évolutive des plantes terrestres mais toujours peu compatible avec les contraintes de laboratoire.

Discussion

Discussion et Perspectives

Pour la plupart, les résultats obtenus au cours de cette thèse ont fait l'objet de discussions dans les articles associés. Cependant, les différentes approches entreprises s'inscrivent dans une perspective plus large que la problématique de chaque article. Ainsi, nos conclusions ont soulevé des interrogations nouvelles, à l'origine de perspectives de travail. Je présente ici quelques-unes des stratégies qui pourront être mise en œuvre pour les aborder.

Le dialogue LEAFY-ADN : une protéine, un génome, plusieurs possibilités

Bien que très étudiés, les mécanismes par lesquels les protéines interagissent avec l'ADN pour contrôler la transcription sont encore loin d'être élucidés (Bulyk, 2003). Identifier et modéliser le code de reconnaissance ADN/protéine est l'un des défis les plus ambitieux de la bioinformatique (Segal and Widom, 2009; Rister and Desplan, 2010). LEAFY est une protéine unique du règne végétal, son site de liaison étendu et son rôle de régulateur-clé du développement en font un candidat idéal pour aborder ces aspects.

Modéliser la spécificité de liaison de ce facteur original

Dès son identification, le gène *LEAFY* (*LFY*) a été proposé comme codant pour un facteur de transcription (Coen et al., 1990). Bien qu'étant un régulateur central de l'organogenèse florale, son mode d'interaction avec la molécule d'ADN et sa spécificité de liaison sont pourtant restés inconnus. En participant à la caractérisation structurale du complexe LFY/ADN et en modélisant la spécificité de liaison de LFY, ce travail de thèse a révélé les bases biochimiques lui permettant de remplir son rôle.

Plusieurs méthodes sophistiquées permettent désormais d'accéder aux spécificités de liaison *in vitro* des facteurs de transcription. La technique de 'protein binding microarrays' (PBM), basée sur l'hybridation d'un facteur de transcription contre une puce recouverte d'oligonucléotides (Bulyk et al., 1999; Badis et al., 2009) ou l'utilisation de plateformes microfluidiques (Maerkl and Quake, 2007) ont été éprouvées avec succès. Cependant, ces méthodes utilisent des petits ADNs (8 à 10 paires de bases) et ne sont pas adéquates pour étudier les facteurs de transcription qui, comme LFY, possèdent un site de liaison étendu. Alternativement, il existe des approches atomiques pour parvenir à un modèle prédictif de la liaison : ces stratégies effectuent des calculs d'énergie d'interaction directement à partir de la

structure d'un facteur de transcription lié à son ADN (Jamal Rahi et al., 2008). Ces méthodes présentent un inconvénient majeur puisqu'elles nécessitent de connaître au préalable la structure du complexe ADN/protéine pour chaque facteur étudié.
Bien que reposant sur un principe simple (Ellington and Szostak, 1990; Tuerk and Gold, 1990), des approches Selex très élaborées ont récemment vu le jour en biologie animale (Jolma et al.; Gopinath, 2007; Stoltenburg et al., 2007). Aussi, nous avons mis au point un protocole de Selex combiné aux techniques de séquençage nouvelle génération (454 SOLEXA ou Illumina) pour obtenir rapidement des ADNs de haute affinité pour LFY et ses homologues. Ces séquences ont servi tout d'abord à la construction d'un modèle prédictif de la liaison de LFY-C d'*Arabidopsis*. La modélisation de la conservation des acides aminés entre les différents homologues de LFY laissait supposer l'existence d'interactions supplémentaires entre LFY et l'ADN, à l'extérieur du motif de 19 nt utilisé pour la cristallographie (Chapitre 1). La matrice SYM-T obtenue décrit un motif de 19 nt également, indiquant que d'éventuels contacts supplémentaires ne dépendent pas de la séquence d'ADN et pourraient plutôt avoir lieu avec le squelette phosphate.

Les techniques d'immunoprécipitation de la chromatine se sont récemment développées pour établir *in vivo* le répertoire de sites reconnus par un facteur de transcription. Peut-on engendrer un modèle prédictif de liaison du facteur d'intérêt à partir de ces données collectées en présence d'une multitude d'autres facteurs pouvant influer sur l'interaction ? La technique originelle du ChIP-on-ChIP ne présente pas une résolution suffisante pour identifier précisément dans les promoteurs contactés, les sites reconnus par le facteur étudié (Segal and Widom, 2009). En revanche, la technique du ChIP-seq génère des fragments de quelques centaines de paires de bases seulement et les algorithmes de recherche de motifs sont aujourd'hui suffisamment efficaces pour parvenir à distinguer, dans certains cas, le motif reconnu par le facteur (Tompa et al., 2005; Wu et al., 2010). En utilisant uniquement les résultats du ChIP-seq contre LFY, nous avons obtenu ici un logo proche de celui établi par Selex mais sans détecter la dépendance existant entre les différentes positions et le modèle issu du ChIP-seq est moins performant (Chapitre 2). Dans certains cas, il est plus difficile de déduire un modèle prédictif de liaison à partir des données *in vivo* : le premier ChIP-seq réalisé sur une protéine de plante a visé SEPALLATA3 (SEP3), un facteur de transcription interagissant avec plusieurs protéines ABC pour former divers complexes du modèle quartet (Immink et al., 2009; Kaufmann et al., 2009). L'immunoprécipitation de la chromatine contre SEP3 ne permettant pas de distinguer les régions contactées par chaque type de complexe

auquel la protéine participe, le logo obtenu est une moyenne des spécificités de chaque multimère et n'est pas utilisable en tant que tel pour scanner un génome.

Le modèle SYM-T permet de prédire l'affinité de LFY pour tout motif ADN. Le génome ayant une architecture plus complexe que l'ADN nu, il reste à évaluer dans quelle mesure cette spécificité détermine le répertoire de site reconnus par LFY dans le génome de l'Arabette.

Comprendre les règles qui régissent la liaison de LFY aux éléments du génome

Par sa fonction régulatrice de l'expression des gènes, un facteur de transcription possède souvent une affinité de base pour toute molécule d'ADN (Manke et al., 2008). De plus, les matrices de liaison des facteurs eucaryotes pluricellulaires contiennent peu d'information ($I \approx 12.1$ bits en moyenne) comparé à celles des protéines régulatrices de bactéries ($I \approx 23$ bits en moyenne). Or, 30 bits d'information seraient nécessaires pour distinguer un site d'intérêt du bruit de fond des génomes eucaryotes (Wunderlich and Mirny, 2009). Par conséquent, l'information contenue dans ces matrices n'est pas suffisante à elle seule pour guider la protéine vers des sites précis du génome. Une recherche de sites de liaison avec une matrice, même bonne, prédira donc l'occupation de nombreux fragments génomiques qui ne seront pas retrouvés parmi les résultats de ChIP-seq (Vingron et al., 2009). La quantité d'information contenue dans la matrice SYM-T est de 17.3 bits, ce qui en fait un modèle plus informatif que la moyenne mais toujours insuffisant. Appliqué au génome, ce modèle va identifier comme 'occupés' des milliers de sites qui ne sont pas trouvés par ChIP-seq (faux positifs). L'analyse ROC indique par exemple que pour détecter 80% des fragments du ChIP, le modèle SYM-T retient comme positif 25% des séquences non liées soit plusieurs milliers de loci génomiques. Pour comprendre le guidage de LFY dans le génome et améliorer la spécificité et la sensibilité de notre modèle, il faut donc identifier les autres supports de l'information de liaison.

En effet, plusieurs molécules structurent l'ADN génomique et conditionnent son accessibilité aux facteurs de transcription. La présence de nucléosomes (Segal et al., 2006), les modifications de la chromatine ou des nucléotides (méthylation des cystéines ou méthylgénome) sont autant de paramètres qui modulent la disponibilité des sites constituant ainsi un second niveau de régulation. Des avancées majeures ont été réalisées ces cinq dernières années pour essayer de caractériser ces paramètres chez les modèles animaux mais également chez Arabidopsis. Ainsi, plusieurs modifications des histones ont été

cartographiées à l'échelle génomique (Zhang et al., 2007), l'épigénome a été déterminé à une base près (Lister et al., 2008) et la carte des nucléosomes d'Arabidopsis vient d'être dévoilée (Chodavarapu et al., 2010). En réalisant un premier scan du génome avec la matrice SYM-T pour établir la carte des probabilités d'occupation par LFY puis, en superposant à cette carte la position des nucléosomes, les méthylations de l'ADN et les modifications des histones, des corrélations pourront peut-être être établies pour tenter de comprendre pourquoi deux sites ayant une même probabilité d'occupation sont reconnus différentiellement par la protéine. Ces corrélations constitueront de nouvelles règles à implémenter dans le modèle biophysique pour améliorer sa spécificité.

Des ressources importantes sont en train d'être générées à l'échelle génomique ; il est donc important de développer les méthodes d'analyses qui permettront de combiner ces informations aux modèles biophysiques de liaison des facteurs de transcription. Ainsi, on pourra peut-être bientôt prédire in silico *le répertoire de sites du génome liés par une protéine régulatrice.*

Les paysages chromatiniens sont dynamiques et leur remodelage a été étudié lors de certains processus développementaux ou situations physiologiques (Charron et al., 2009). Aussi, certains sites apparemment non accessibles lorsque l'individu tout entier est considéré, pourraient être liés dans des tissus donnés, à des moment précis du développement ou en réaction à des stress environnementaux. Le gène *APETALA3* (*AP3*) par exemple est une cible directe de LFY mais nous n'avons pas identifiés par ChIP-seq les sites de liaison présents dans sa région promotrice, peut-être parce qu'ils ne sont pas accessibles au stade plantule utilisé pour le ChIP. Deux approches peuvent être employées pour tester cette hypothèse : une première solution consiste à reproduire la manip de ChIP-seq dans des inflorescences ou méristèmes floraux. Cela permettrait de comparer les répertoires de sites liés par la protéine pendant la phase végétative et la phase reproductive. Dans un second temps, une approche plus fine pourra être déployée: deux technologies principales sont disponibles pour isoler un tissu d'intérêt. La première réalise un tri cellulaire basé sur la fluorescence afin de récupérer uniquement les cellules exprimant un signal GFP. À l'aide d'une lignée exprimant la GFP dans un tissu donné, cette technique permet d'obtenir une population de cellules homogène (Birnbaum et al., 2005; Brady et al., 2007). La seconde est basée sur la réalisation de microdissections laser pour isoler les cellules recherchées (Nakazono et al., 2003; Dembinsky

et al., 2007; Spencer et al., 2007). Couplées au ChIP-seq, ces techniques pourraient permettre de déterminer les sites accessibles à LFY dans un type cellulaire donné.

Les résultats de ChIP-seq ont également distingué plusieurs fragments qui ne contiennent aucun 'bon' site individuel puisque plus de 8% des 1564 séquences retenues par les deux expériences de ChIP ne possèdent aucun motif avec un score supérieur à -20. Comment LFY contacte-t-il efficacement ces séquences ? Pour savoir si ces contacts ont lieu grâce à des facteurs présents seulement *in vivo*, on peut tester si ces régions sont toujours reconnues lorsque la protéine est utilisée seule et que les autres facteurs cellulaires (contexte chromatinien, cofacteurs) sont absents. La SPR étant applicable à l'étude des longs fragments ADN (Chapitre 3), elle constitue une méthode de premier choix pour ce test. Si les interactions sont vérifiées, deux possibilités sont envisageables : soit la protéine est capable de reconnaître un second type de site, soit il existe un mécanisme qui lui permet de recruter plusieurs sites de mauvaise affinité pour établir un contact efficace.
La caractérisation préliminaire de plusieurs homologues de LFY indique que leur domaine N-terminal permet la formation de tétramères sur l'ADN (données de Camille Sayou et Renaud Dumas). Ces complexes pourraient contacter simultanément deux sites de liaison avec une coopérativité intersites. Ce mécanisme a été décrit pour des facteurs qui, comme LFY, contrôlent des transitions développementales marquées (Lebrecht et al., 2005; Gregor et al., 2007). Quelles sont les évidences suggérant l'existence d'un tel mécanisme avec LFY ? Plusieurs sites d'affinité proches sont regroupés dans les régions régulatrices d'*AGAMOUS* (*AG*) ou de *TERMINAL FLOWER1* (*TFL1*) reconnues par LFY, mais une liaison coopérative de la protéine entre ces sites n'a jamais été démontrée. La SPR constitue une méthode potentielle pour tester l'existence d'une telle coopérativité.

Alternativement, ces régions détectées en ChIP-seq mais dépourvues de bon site pour LFY pourraient être reconnues exclusivement *in planta* si la présence d'un partenaire est nécessaire. Plusieurs cofacteurs de LFY sont eux-mêmes des facteurs de transcription capables de reconnaître spécifiquement des séquences ADN propres. Ainsi, des expériences de ChIP-seq viennent d'être réalisées sur SEP3 (Kaufmann et al., 2009) et WUSCHEL (WUS) (Busch et al., 2010). Plusieurs loci contactés par LFY, notamment ceux reliés à la floraison ou aux communications hormonales, le sont également par SEP3. Une idée intéressante serait d'imaginer la formation d'hétérocomplexes LFY-SEP3 ou LFY-WUS, capables de lier des sites différents du motif reconnu par le dimère de LFY. Les cofacteurs de

LFY joueraient ainsi le rôle d'aiguilleur de la protéine en orientant le choix des cibles régulées. Une recherche de motifs à l'aide du programme MEME pourra être effectuée sur les 8% de séquences de ChIP-seq dépourvues d'un 'bon' site de liaison pour LFY pour identifier éventuellement un motif ADN correspondant à la signature d'un hétérocomplexe LFY-partenaire. Quels sont les autres moyens à disposition des cofacteurs pour influencer *in vivo* la liaison de LFY à l'ADN ? La protéine activée LFY-VP16 ne modifie pas le domaine d'expression du gène *AP3* qui reste limité à la zone exprimant *UNUSUAL FLORAL ORGAN* (*UFO*) (verticilles 2 et 3, (Parcy et al., 1998)). Ces données suggèrent que UFO est nécessaire pour le recrutement de LFY au niveau des séquences régulatrices d'*AP3*, par exemple en modifiant la spécificité du facteur ou en induisant un remodelage de la chromatine dévoilant les sites de liaison d'*AP3*. Réaliser une expérience de ChIP-seq sur des plantules *35S::LFY 35S::UFO* ou sur des plantules *35S::LFY ufo* puis comparer les résultats enregistrés aux données de ChIP que nous avons déjà obtenues permettrait de tester l'influence de l'activité UFO sur le répertoire de sites liés par LFY.

Liaison n'étant pas synonyme de régulation, comment identifier parmi les sites reconnus par LFY ceux qui conduisent à une modulation de l'expression des gènes environnants ?

Identifier la signature génomique des sites régulateurs

Identifier les sites de régulation de l'expression des gènes présente des enjeux considérables que ce soit pour prédire les effets phénotypiques de variations d'une séquence entre individus d'une même population ou pour réaliser une lecture des nouveaux génomes séquencés (Farnham, 2009; Janky et al., 2009). Les sites régulateurs peuvent être de forte ou de faible affinité (Tanay, 2006; Gertz et al., 2009; Meijsing et al., 2009). De plus, la majorité des sites liés par un facteur individuel n'aurait aucun rôle fonctionnel (Wasserman and Sandelin, 2004). Plusieurs études ont testé la validité de ce postulat, célèbre sous le nom de 'Futility theorem'. Certains auteurs affirment qu'une prédiction suffisamment fine des sites de liaison permet d'expliquer les domaines d'expression observés (Segal et al., 2008; Zinzen et al., 2009). D'autres analyses en revanche rapportent une corrélation limitée entre les patrons d'occupation des régions régulatrices et les profils d'expression suggérant que seul un sous-ensemble des gènes liés par un facteur de transcription permet un contrôle de l'expression (Gao et al., 2004; Hu et al., 2007). Dès lors, comment identifier parmi les sites reconnus par LFY ceux qui conduisent à une régulation des gènes cibles ?

Une première approche consiste à utiliser l'évolution pour distinguer une partie des sites fonctionnels des sites artefactuels (Macisaac et al., 2006; Ward and Bussemaker, 2008; Vingron et al., 2009) : le génome d'*Arabidopsis lyrata*, une autre espèce de Brassicacées très proche d'*Arabidopsis thaliana* a récemment été séquencé. La proximité des deux espèces permet d'identifier sans aucune difficulté les orthologues. Scanner le génome de chaque Arabidopsis avec le modèle SYM-T devrait permettre de repérer les sites non significatifs plus versatiles (et donc non retrouvés simultanément chez les deux espèces apparentées) tandis que les principaux sites contactés par LFY pour exercer sa fonction doivent être conservés. D'autres génomes de Brassicacées sont en cours de séquençage puisque cette famille possède de nombreuses espèces à intérêt agronomique fort (chou, radis, colza). En constituant des 'filtres évolutifs' basés sur les génomes d'espèces de plus en plus éloignées, ce type d'approche devrait permettre un repérage efficace de sites fonctionnellement pertinents.

Il a aussi été proposé que la position des éléments cis par rapport au site d'initiation de la transcription est un paramètre important pour faire d'un motif un site fonctionnel (Farnham, 2009). Dans le cas de LFY, la pertinence de la localisation des éléments cis reste à déterminer puisque ceux-ci sont indifféremment retrouvés en amont (ex :promoteur AP1), en aval (ex :TFL1) ou dans les séquences codantes (ex : AG) des gènes régulés.

Retracer 400 Millions d'années d'évolution

Évolution des interactions LFY/ADN

LFY est une protéine très conservée aux travers de la lignée verte et qui, contrairement à la plupart des facteurs de transcription, n'a pas engendré de famille multigénique. Il est donc facile d'identifier ses homologues chez des espèces très éloignées pour retracer l'évolution de cette protéine régulatrice. Savoir si les nouveautés évolutives émergent principalement de modifications des séquences protéiques ou de transformations des éléments *cis* régulateurs fait toujours débat (Hoekstra and Coyne, 2007; Wray, 2007). Les résultats que nous avons obtenus indiquent que les deux 'moteurs de l'innovation moléculaire', changement de spécificité et dynamique des éléments *cis*, ont pu contribuer à l'évolution du rôle de LFY.

Le changement de spécificité

La résolution des premières structures ADN/facteurs de transcription a révélé qu'il n'existait pas de correspondance simple entre un acide aminé donné et une base nucléotidique puisque des molécules ayant une séquence protéique très proche et des structures secondaires conservées reconnaissent chacune leur propre motif ADN (Pabo and Sauer, 1984). En dépit d'une conservation stricte de tous les résidus impliqués dans les contacts avec l'ADN, LEAFY et ses homologues ne partagent pas une même spécificité. PpLFY1 de la mousse *Physcomitrella patens* est incapable de reconnaître les motifs préférés par LFY et reconnaît des séquences différentes. L'absence des résidus H394 et R427, remplacés par un acide aspartique et une cystéine respectivement, supprime probablement la liaison qui positionne la grande hélice α3 (Chapitre 1). Cette hélice porte 2 des 4 acides aminés qui contactent spécifiquement le motif ADN ; une variation de sa position peut donc modifier la liaison du motif et induire un changement de spécificité même si les résidus au contact direct de l'ADN sont conservés.

Les pertes et gains d'éléments cis

Il est possible que la spécificité de liaison de LFY ait été figée chez les plantes à fleurs, la protéine jouant désormais un rôle crucial pour la survie de l'espèce. Les homologues angiospermes que nous avons testés jusqu'ici semblent capables de reconnaître les mêmes séquences que la protéine d'Arabidopsis mais des différences de spécificité *in vitro*, même mineures, peuvent contribuer à la sélectivité du facteur *in vivo* (Wei et al., 2010). Les résultats du Selex détermineront si des micro-ajustements de spécificité on eu lieu entre homologues. En revanche, l'étude réalisée sur la sous-famille *AG* indique que plusieurs espèces ont adoptées des stratégies différentes pour retenir ou éliminer de leur séquences régulatrices des éléments reconnus par LFY (Chapitre 2). Les duplications survenues le plus précocement au sein des sous-familles AP3/PI, AG/STK et SEP coïncideraient avec l'émergence des plantes à fleurs (Aoki et al., 2004; Kramer et al., 2004; Stellari et al., 2004). Par conséquent, un événement clé pour l'évolution du réseau floral au travers de la diversification des angiospermes est la façon dont les différentes copies générées ont retenues ou perdu les éléments *cis* liés par LFY. En particulier, il serait intéressant de suivre le devenir des motifs des gènes de classe B : chez les eudicotylédonnes basales comme les Renonculacées, la lignée *AP3* a connu plusieurs duplications secondaires, entraînant la création de gènes B spécialisés dans le développement d'un seul type d'organe (pétale, étamine ou organe pétaloïde supplémentaire) (Kramer et al., 2007). Comprendre comment les éléments *cis* liés par LFY

sont passés au travers de ces duplications permettrait peut-être de proposer une explication à la spécialisation des différents paralogues.

Au cours de ce travail de thèse, nous avons surtout envisagé l'évolution de l'interaction LFY-ADN or il existe d'autres éléments ayant pu contribuer à la mutation fonctionnelle de ce facteur. Une restriction du domaine d'expression des homologues de LEAFY chez les plantes à graines par rapport aux mousses et fougères (Himi et al., 2001; Tanahashi et al., 2005) ainsi qu'une modification des éléments cis qui contrôlent l'expression de LFY chez plusieurs Brassicacées ont pu participer à l'évolution de sa fonction (Yoon and Baum, 2004; Sliwinski et al., 2006). Pour évaluer l'importance de ce phénomène, il serait nécessaire de disséquer les régions régulatrices de LFY et ses homologues et identifier les conséquences morphologiques des changements de son domaine d'expression.

Évolution fonctionnelle de LFY

Comment passer du réseau pré-floral au réseau floral ?

À terme, l'origine des angiospermes pourra être résolue par une combinaison des travaux de paléobotanique, systématique et biologie moléculaire. En attendant de trouver les évidences fossiles qui permettront d'éclairer cet 'abominable mystère', retracer le chemin évolutif ayant conduit à la création du circuit génétique floral offre une opportunité unique d'étudier les mécanismes moléculaires à l'origine de l'apparition de nouvelles formes.

LFY et les gènes ABC étant largement retrouvés chez le groupe frère des angiospermes, les scénarios génétiques proposés font le postulat d'un réseau pré-floral dont le recyclage aurait conduit à la création de la première fleur (Frohlich, 2000; Albert et al., 2002; Theissen and Becker, 2004; Baum and Hileman, 2006). Les patrons d'expression et les résultats de SPR que nous avons obtenus chez *Welwitschia mirabilis* indiquent une spécialisation du rôle de chaque paralogue et apportent les premières évidences expérimentales en faveur d'une régulation des gènes B par LFY chez les gymnospermes. Ainsi, LFY pourrait assurer la régulation des gènes B chez l'ancêtre commun des plantes à graines. La coexpression de *WelAG* et *WelNdly*, particulièrement visible au niveau du développement tardif des tissus producteurs de pollen et de l'ovule, laisse supposer une régulation de *WelAG* par WelNdly.

Étudier par SPR l'existence d'interaction entre les protéines WelLFY, WelNdly et le second intron de *WelAG* récemment isolé permettra de tester cette hypothèse.
Comme *NEEDLY* (*NLY*) est absent des génomes angiospermes et que LFY régule l'expression des gènes B et C chez les plantes à fleurs, nous pouvons proposer que LFY ait acquis la capacité à réguler de nouveaux gènes, dont *WelAG*, après la divergence des gymnospermes. Alors que deux gènes assureraient l'expression de B et C dans des territoires chevauchant puis mutuellement exclusifs chez les gymnospermes, un seul gène réaliserait ce travail chez les premières angiospermes. Placer sous contrôle d'un même régulateur les activités B et C permet de lier les cascades développementales qui génèrent les organes mâles et femelles, ce qui a pu rendre la formation de structures bisexuelles obligatoire. Ceci est cohérent avec le modèle de fleur primitive proposé. D'autre part, les gènes B et C de plusieurs espèces du grade ANA ont des domaines d'expression étendus et très chevauchants (Kim et al., 2005) alors que chez la plupart des eudicotylédones centrales, les domaines d'expression des gènes B et C sont limités aux verticilles dictés par le modèle ABCDE et ses variantes. Parallèlement à la capacité à activer de nouveaux gènes, LFY a pu commencer à recruter des corégulateurs différents, autorisant une régulation découplée de B et C. Ainsi, des territoires mutuellement exclusifs auraient pu être rétablis au cours de la diversification des angiospermes.

Le précurseur de la fleur étant toujours inconnu, on ignore si celle-ci dérive d'un cône mâle, d'un cône femelle ou d'une sous-partie d'une de ces structures reproductrices (Frohlich and Chase, 2007). Identifier la structure ancestrale et élucider les mécanismes qui ont rendu possible l'apparition de la fleur nécessite donc de comprendre le développement des organes sexuels des gymnospermes. Or, l'absence d'organisme modèle gymnosperme pour lequel des ressources génétiques conséquentes (génome séquencé, banque de mutants) seraient disponibles constitue une limite importante à la réalisation de progrès significatifs. Il est possible de transformer plusieurs variétés de pins et épicéa mais ces espèces ligneuses ont des cycles de vie très long et plusieurs années sont nécessaires pour qu'elles atteignent la maturité et commencent à former leurs premiers cônes (Tang et al., 2007). Comment tester alors le réseau pré-floral proposé ?

Les matrices de liaison à l'ADN de WeLFY et WelNLY une fois disponibles permettront de détecter l'ensemble des sites préférés par chacun des paralogues au sein des loci B et C. Il n'existe pas de séquences promotrices connues pour les gènes B d'une autre gymnosperme et

il est souvent difficile de cloner les régions génomiques en amont des gènes d'une espèce non séquencée. En revanche, les homologues d'*AG* ont été identifiés chez les 4 groupes de gymnospermes dont *Gnetum*, une espèce proche de *Welwitschia* (Tandre et al., 1995; Rutledge et al., 1998; Winter et al., 1999; Zhang et al., 2004). Il est possible d'amplifier facilement le second intron de ces homologues puis de le comparer avec celui de *WelAG* pour regarder si les sites de hautes affinités identifiés par les matrices sont conservés chez une espèce apparentée ; cela constituerait une première évidence de l'importance fonctionnelle de ces sites. L'utilisation de l'interférence ARN (Becker and Lange, 2010) pourrait aider à contourner le problème du temps de génération très long des gymnospermes. Ainsi, en perturbant ou supprimant l'expression d'un gène directement dans un tissu ou organe donné, ces méthodes une fois adaptées à *Welwitschia*, permettraient de tester l'effet d'une absence d'activité de WelLFY ou WelNdly sur le développement et la morphologie des cônes.

LFY sans son ABC : à la recherche du rôle ancestral de LFY

Identifier précisément la fonction de LFY chez les mousses et fougères représente un véritable challenge. Chez ces deux groupes, il n'existe en effet aucun homologue des gènes régulés par LFY chez les plantes à fleurs (Münster et al., 2002; Singer et al., 2007), il n'y a donc pas de gènes cibles candidats *a priori*. De plus, un mutant pour l'activité PpLFY chez *Physcomitrella patens* présente un phénotype embryonnaire létal, ce qui rend impossible l'identification directe des gènes dérégulés.

Des ressources génétiques suffisantes combinées aux méthodes transcriptomiques et génétiques à haut débit peuvent cependant permettre d'obtenir une vue d'ensemble du réseau génétique d'intérêt (Long et al., 2008). À ce titre, *Physcomitrella patens* constitue un excellent modèle pour aborder le rôle de LFY en l'absence de gènes ABC : il s'agit d'une plante relativement facile à cultiver, transformable, dont le génome est séquencé depuis 2008 (Quatrano et al., 2007; Rensing et al., 2008). Ces ressources permettent deux approches complémentaires : d'une part, en utilisant la matrice 'PpLFY1' générée grâce au résultats du Selex, un scan du génome pourra être réalisé pour identifier les régions ayant une probabilité élevée d'occupation par PpLFY. L'anticorps anti-LFY-C que nous avons produit reconnaît également PpLFY1 (observation de Mylène Robert), ainsi une expérience de ChIP-seq pourra être effectuée à partir de plants entiers de *P.patens* puisque les deux paralogues *PpLFY1* et *PpLFY2* ayant une expression assez large. Si la précipitation est difficile, des lignées

surexpresseurs *35S ::PpLFY1* et *35S ::PpLFY2* disponibles chez la mousse pourront être utilisées (Tanahashi et al., 2005).

Ces deux stratégies (scan du génome et ChIP-seq) devraient générer une liste de sites potentiellement reconnus par le facteur, comment faire émerger de cette liste les cibles de PpLFY ? Dans un premier temps, une confrontation des résultats des deux approches permettra d'obtenir un nombre plus restreint de sites ayant une bonne affinité pour le facteur et reconnus *in vivo*. Plus de 35000 gènes ont été prédits ou annotés chez *Physcomitrella* (Rensing et al., 2008) et les candidats pour lesquels une fonction est connue seront examinés en priorité. En particulier, on pourra rechercher les gènes ayant trait à la stimulation des divisions cellulaires puisqu'il s'agit de la fonction ancestrale proposée pour LFY. À l'issue de ces différentes étapes, un groupe minimum de cibles pourra être retenu : si ces gènes sont effectivement régulés par PpLFY, leur expression doit être modifiée entre une plante sauvage et une plante *35S::PpLFY* surexprimant le facteur de transcription. Détecter les variations d'expression entre ces deux lignées par RT-PCR quantitative ou RNA-seq (si les gènes suivis sont nombreux, (Wang et al., 2009)) permettra de conforter les cibles. *Physcomitrella patens* présente un fort taux de recombinaison homologue, similaire à celui de levure (Schaefer, 2001). Cette capacité unique parmi les plantes modèles actuelles, permet de muter précisément un gène ou un élément *cis* d'intérêt directement au locus endogène. Ainsi, il sera possible de vérifier que la perte de la liaison par LFY entraîne bien la dérégulation du gène testé.

La mousse étant un excellent modèle génétique, cet organisme pourrait également permettre des avancées majeures concernant les règles qui gouvernent l'expression des génomes en général et constitue un système de choix pour élucider les raisons qui font d'un bon site de liaison un site fonctionnel.

En conclusion, ce travail sur le facteur de transcription LEAFY m'a permis d'aborder des questions de nature très générale et d'autres spécifiques aux plantes comme le développement et l'origine des fleurs. Il est vraisemblable que les progrès accomplis (structure du domaine C-terminal, modèle de liaison à l'ADN, première évidence d'un réseau pré-floral) contribueront aux travaux futurs de l'équipe.

MATERIEL ET METHODES

Matériel et Méthodes

Les techniques ayant contribuées aux résultats publiés sont exposées dans les Matériel et Méthodes des articles correspondants et seuls sont détaillés ici les techniques que j'ai utilisées ou mises au point pour obtenir les résultats complémentaires.

Matériel

1-Matériel végétal

Les tissus de *Welwitschia mirabilis* (*W.mirabilis*) ont été collecté par Mike Frohlich en Californie (Article 4). Ceux d'*Amborella trichopoda* (*A. trichopoda*) ont été collectés en Nouvelle-Calédonie par Charlie Scutt (Laboratoire RDP, ENS Lyon). Les tissues d'*Arabidopsis thaliana* (*A. thaliana*) proviennent de plantes écotype *Columbia* cultivées au laboratoire (22°C, 16h de lumière, 8h d'obscurité, intensité lumineuse de 120 µE $m^{-2}s^{-1}$) sur support autoclavé 50% vermiculite exfoliée, 50% terreau. J'ai collecté les tissus de *Lunularia cruciata* (*L. cruciata*) et *Marchantia polymorpha* (*M. polymorpha*) en Avril 2007 au Chelsea Physical Garden (Londres).

2-Matériel bactérien

Les clonages classiques ont été effectués dans des bactéries *Escherichia Coli* de souche DH5α F^- ϕ80 *lacZ ΔM15 Δ(lacZYA-argF)U169 deoR recA1 endA1 hsdR17*(r_k^-, m_k^+) *phoA supE44 thi-1 gyrA96 relA1* λ^- (Invitrogen) ou TOP10 F^- *mcrA D(mrr-hsdRMS-mcrBC) f80lacZDM15 DlacX74 deoR recA1 araD139 D(ara-leu)7697 galU galK rpsL (StrR) endA1 nupG* (Invitrogen).

Les expressions de protéines ont été effectuées à partir de bactéries *E. coli* souche RosettaBlue(DE3)pLysS F^- *ompT hsdS_B(r_B^- m_B^-) gal dcm* λ *(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5]) pLysSRARE (Cam^R ;*Novagen) ou *E. coli* souche Rosetta2(DE3)pLysS F^- *ompT hsdS_B(r_B^- m_B^-) gal dcm (DE3) pLysSRARE2 (Cm^R ;* Novagen).

3-Levures

Les expériences d'induction chez la levure ont été réalisées à partir de la souche *Saccharomyces cerevisiae* EGY48 *Mat*α *his3 leu2::3LexAop-LEU2 ura3 trp1 LYS2* (Invitrogen).

4-Solutions d'usage courant

LB (*Luria Bertani*) préparation manuelle: 1% [m/v] bacto tryptone, 0.5% [m/v] extrait de levure, 1% [m/v] NaCl, pH7.0, + *optionnel 1% agar pour obtenir un milieu solide*
LB commercial: 2% [m/v] poudre LB Broth Base Lennox (Invitrogen) + *optionnel 1% agar pour obtenir un milieu solide*
LB liquide 0,5M NaCl: 2% [m/v] poudre LB Broth Base Lennox (Invitrogen), 2.42% [m/v] NaCl
Laemmli 1X: 2.5% β-mercaptoethanol, 2% SDS, 80mM Tris pH6.8, 10% glycérol, 0.004% bleu de bromophénol
TB (*Terrific Broth*) liquide préparation maison (protocole initial): 1.2% [m/v] Bacto Tryptone, 2.4% [m/v] extrait de levure, 1% [m/v] NaCl, 0.04% [v/v] Glycérol, 50 mM KPO_4 pH7.2
TB liquide commercial: 4.76% [m/v] poudre TB modified (Sigma), 0.8% [v/v] glycérol
TAE (Tris Acetate EDTA) 1X: 40mM Tris, 20mM acide acétique, 1mM EDTA pH8.0
TAE modifié 1X: 40mM Tris, 20mM acide acétique, 0.1mM EDTA pH8.0
TBE 10X: 10.8% [m/v] Tris, 5.5% [m/v] acide borique, 20mM EDTA 0.5M pH8.0
TE (Tris EDTA): Tris 10mM pH8.0, EDTA 1mM

Méthodes

1-Biologie moléculaire

Extraction d'ADN génomique

ADN génomique d'*A thaliana*
L'ADN génomique d'*A. thaliana* a été extrait d'après le protocole d'Edwards et al., 1991. 'équivalent d'un demi-capuchon de tube eppendorf de jeunes feuilles est écrasé à température ambiante pendant 15 sec à l'aide d'un pilon adapté aux tubes eppendorf 1,5 mL. 200 µL de tampon d'extraction (200 mM Tris-HCl pH7.5; 250 mM NaCl; 25 mM EDTA; 0,5% SDS) sont ajoutés puis le mélange est vortexé 5sec. Après centrifugation à 10000 g pendant 1min, 150 µL de surnageant sont transférés dans un tube eppendorf propre. 150 µL d'isopropanol sont ajoutés et le tube est incubé 2 min à température ambiante. Après centrifugation à 10000g pendant 5 min, le culot est lavé avec 100µL d'EtOH 70% et centrifugé à 10000g pendant 5min. Le culot est finalement séché au speedvac puis dissout dans 50µL de TE et stocké à -20°C.

ADN génomique de *W. mirabilis* et *A. trichopoda*
Les ADN génomiques de *W. mirabilis* ont été obtenus par Mike Frohlich, ceux d'*A. trichopoda* par Charlie Scutt.

Extraction d'ARNm et synthèse d'ADNc

ARNm et ADNc de *W. mirabilis*, *M. polymorpha* et *L. cruciata*
L'extraction des ARNm et la synthèse des ADNc de *M. polymorpha* et *L. cruciata* ont été obtenus à partir de thalles comme ceux de *We. mirabilis* décrits dans l'article 4.

ADNc d'*A. trichopoda*, *Illicium parviflorum* et *Ginkgo biloba*
Le même protocole a été utilisé pour *A. trichopoda* à partir de boutons de fleurs mâles et femelles récoltés par C. Scutt. Les ADNc utilisés pour amplifier l'homologue de LFY chez le ginkgo (*G. biloba*) et *Illicium parviflorum* (*I. parviflorum*) ont été préparés par C. Scutt et M. Frohlich respectivement.

Clonage

Clonage des homologues de *LEAFY*
Les homologues de LFY déjà identifiés chez *G. biloba* (AF108228), *W. mirabilis* (AF109130, AF108227) et *M. polymorpha* (AF286056, uniquement séquence partielle disponible) ont été amplifiés avec la Phusion®DNA polymerase (Ozyme) à partir d'ADNc de l'espèce correspondante en utilisant des couples d'amorces spécifiques. Les produits PCR obtenus ont été clonés dans le vecteur pCRBlunt (Invitrogen) pour donner les vecteurs pEDW60 (*GinLFY*), pEDW86 (*WelLFY*), pEDW87 (*WelNdly*) ou dans le vecteur PCR4 TOPO (Invitrogen) pour donner pEDW100 (*MarpoFLO*). Tous les vecteurs ont été vérifiés par séquençage (Cogenics).

Les séquences partielles d'*IlliLFY* et *LuLFY* ont été obtenues par PCR sur ADNc d'*I. Parviflorum* et *L. cruciata* respectivement, à l'aide des amorces dégénérées LFL1 et LFR1 (Frohlich and Meyerowitz, 1997) puis amplification de l'extrémité 3' de l'ADNc correspondant (3' RACE system for rapid amplification of cDNA ends (Invitrogen)). Les produits PCR ont été directement séquencé (Cogenics).

Une séquence partielle d'*AmboLFY* a été obtenue par criblage d'une banque d'ADNc préparée à partir de fleurs femelles d'*A. trichopoda* (Fourquin et al., 2005). Après transfert des plages de lyse incluant les bactériophages λ sur membrane d'hybridation (Amersham Hybond N), les membranes sont incubées avec une sonde hétérologue radioactive correspondant à l'ADNc du gène LFY d'*A. thaliana* (AtLFY) selon le protocole λ Zap (Stratagene). Les clones positifs ont été purifiés par un clonage secondaire permettant d'obtenir des plages de lyse individualisées. Les ADNc sont obtenus dans le

vecteur pBluescript II par excision in vivo (protocole Stratagene λZapII) par l'intermédiaire du phage helper M13 ExAssist (Stratagene). La séquence complète a ensuite été obtenue à l'aide du kit Marathon® cDNA Amplification (Clontech) à l'aide de la TITANIUM® Taq DNA polymerase (Clontech) et cloné dans le vecteur pGEM-T Easy (Promega) pour donner le vecteur pEDW71.

Les séquences de PpLFY1 et PpLFY1-D394H ont été introduites par Alexis Maizel dans le vecteur pCH1 qui comporte une étiquette Glutathione S-transférase (GST) en 54 et une étiquette 6X histidine (6his) en 3'. Les vecteurs pAM422 (PpLFY1) et pAM426 (PpLFY1-D394H) ont ainsi été obtenus.

Clonage des gènes à boîte MADS et des homologues d'actine et tubuline chez *W. mirabilis*
Les homologues des gènes MADS floraux, de l'actine et de la tubuline chez *W. mirabilis* ont été isolés suivant la procédure décrite dans l'article 4.

Clonage des régions régulatrices chez *W. mirabilis* et *A. trichopoda*
Les régions génomiques en amont du codon Start de *WelB1* et *WelB2* ont été isolées comme indiqué dans l'article 4. Pour le second intron de *WelAG*, une méthode identique a été utilisée mais 2 plaques de lyses donnant un signal positif ont été identifiées. Une PCR sur une solution de chacun des 2 plaques (article 4) a permis le clonage de deux fragments chevauchants, de 6.3 kb (pEDW120) et 11 kb (pEDW132) dans le vecteur pCRBlunt (Invitrogen). Une fois rassemblées, ces deux séquences ont formé le second intron complet de *WelAG*.
Le 2nd intron d'*AmboAG*, l'homologue d'*AGAMOUS* chez *A. trichopoda* (AY936231), a été cloné avec la Phusion®DNA polymerase (Ozyme) à partir d'ADN génomique d'*A. trichopoda, et* d'oligonucléotides spécifiques (oEDWAmboAG1F:5'CGCATAGAGAACACAACTAATAGGCAGGT3'; oEDWAmboAG3R:5'GAGAATTGGCCTCTGAAACAGTTCCGGA3').

Mutagenèse

Les vecteurs pEDW50 (LFY-C H312D-6his), pEDW124 (LFY-C D364E R390K), pEDW125 (LFY-C D364E), pEDW126 (LFY-C R390K) ont été obtenus par stratégie de mutagenèse 'Quick change' (Stratagène) à partir de la matrice pCH28 (Article 1).

Construction des vecteurs d'expression

La méthode consiste à amplifier la séquence codante du gène d'intérêt à partir d'ADNc, puis d'ajouter des sites de restriction de part et d'autre pour l'insertion dans un vecteur d'expression permettant de fusionner à la séquence une ou plusieurs étiquettes autorisant la purification. Cette méthode a été décrite en détail pour LFY-C et GFP-LFY-C (Article 1) et WelLFY-C, WelNdly-C, GFP-WelNdly-C, WelLFYmin58 et WelNdlymin56 (Article 4). Pour les autres protéines produites, seuls sont indiqués ci-dessous le nom des vecteurs d'expression finals et les protéines correspondantes.

Vecteur d'expression	Protéine correspondante	Article/Chapitre associé	Remarque
pCH2	GST-LFY-6his	Chapitre 1, Fig1.1	Etiquette GST clivée, protéine LFY-6his utilisée
pEDW50	LFY-C H312D	Chapitre 1, Fig1.2	
pEDW124	LFY-C D364E R390K	Chapitre 1, Fig1.2	
pEDW125	LFY-C D364E	Chapitre 1, Fig1.2	
pEDW126	LFY-C R390K	Chapitre 1, Fig1.2	
pETH79	AtLFYmin40-6his	Chapitre 2, Fig2.1, Table 2.5	
pETH148	PpLFY1-6his	Chapitre 2, Fig2.2A, Fig2.3, Table 2.5	
pETH149	PpLFY1-D394H-6his	Chapitre 2, Fig2.2A, Fig2.3, Table 2.5	
pETH152	PpLFY1-C427R-6his	Chapitre 2, Fig2.3	
pAM426	GST-PpLFY1-D394H-6his	Chapitre 2, Fig2.2B	Etiquette GST clivée, protéine PpLFY1-6his utilisée
pEDW112	6his-PpLFY1-C	Chapitre 2	
pEDW111	6his-PpLFY1-C D394H	Chapitre 2, Fig2.2C	
pETH153	PpLFY1-D394H C427R-6his	Chapitre 2, Fig2.3	
pEDW102	6his-Trx-MarpoFLO-C	Chapitre 2, Fig2.4	
pSAB6	6his-GinLFY-C	Chapitre 2, Table 2.5	
pETH96	GinLFYmin40-6his	Chapitre 2, Table 2.5	
pEDW95	6his-AmboLFY-C	Chapitre 2, Table 2.5	

pETH91	AmboLFYmin40-6his	Chapitre 2, Table 2.5

2-Biochimie

Surproduction des protéines en système bactérien *E.coli* et purification

Protéines recombinantes correspondant au domaine C-terminal
Les protéines LFY-C H312D, LFY-C D 364E, LFY-C R390K, LFY-C D364E R390K, PpLFY1-C, PpLFY1-C D394H, MarpoFLO-C, AmboLFY-C, GinLFY-C ont été produites selon le protocole mis au point pour LFY-C (Matériel et Méthode Article 1) en utilisant les vecteurs adéquats donnés dans la table ci-dessus (cf. Construction des vecteurs d'expression).

Protéines recombinantes comportant les domaines N- et C-terminaux
Les protéines WelLFYmin58 et WelNdlymin56 ont été produites et purifiées selon le protocole détaillé dans les Matériel et Méthodes de l'Article 4. Les protéines GinLFYmin40, AmboLFYmin40 et AtLFYmin40 ont été produites par Renaud Dumas à partir des vecteurs indiqués ci-dessus (cf. Construction des vecteurs d'expression).

Protéines recombinantes entières : PpLFY1 sauvage et versions mutées
Les protéines GST-PpLFY1-6his et GST-PpLFY1-D394H-6his (Chapitre 2, Fig2.2) ont été produites à partir des vecteurs pAM422 et pAM426 respectivement, introduits dans les bactéries *E.coli* RosettaBlue(DE3)pLysS (Novagen). À partir d'une préculture saturée, les bactéries sont diluées au $1/50^{ème}$ dans 1L de milieu TB+antibiotiques à 37°C en agitation continue. A DO_{600} comprise entre 0.8 et 1, l'expression de LFY est induite sur 12 à 15h à 22°C par ajout de 0,5mM IPTG. Les bactéries induites sont centrifugées 30min à 4500g.
Chaque culot de 1 L de culture initiale est resuspendu dans 35mL de tampon de lyse GST (20mM Tris pH8, 250mM NaCl, 1mM EDTA, 1mM azide Na, 5% [v/v] glycérol, 0,2% [v/v] triton X-100, 2mM DTT, une tablette de cocktail d'inhibiteur de protéases (Roche)), soniqué, et centrifugé 40min à 30000g. Le surnageant correspond à la fraction des protéines solubles. Une purification reposant sur la présence de l'étiquette GST est réalisée par chromatographie d'affinité sur résine Glutathione sepharose 4B (Amersham BioSciences). Les protéines GST-PpLFY1-6his ou GST-PpLFY1-D394H-6his sont fixées à la résine par incubation 1h à 4°C (à raison de 2 mL de résine/L de culture initiale), puis centrifugées à 300g pendant 5min. La résine est déposée sur colonne et lavée avec 10 volumes de TBSE (20mM Tris pH8, 250mM NaCl, 1mM EDTA, 5% [v/v] glycérol, 2mM DTT). Les protéines recombinantes sont éluées après digestion pendant 2h (temps optimal de la digestion) à la thrombine (à raison de 25 unités/ml de résine) qui reconnaît un site de clivage spécifique entre la GST et PpLFY1. La thrombine est inactivée avec du PMSF et éliminée avec des billes de p-aminobenzamidine. Une deuxième étape de purification basée sur la présence de l'étiquette 6his est réalisée par chromatographie d'affinité sur résine Ni sepharose (Amersham BioSciences). Les fractions contenant les protéines PpLFY1-6his ou PpLFY1-D394H-6his sont dialysées une nuit à 4°C dans 2L de tampon de charge histidine (20mM Tris pH8, 500mM NaCl, 5% [v/v] glycérol, 1mM DTT, 5mM imidazole), incubées 1h à 4°C avec des billes Ni-NTA préalablement lavées et équilibrées (à raison de 250µL de résine Ni-NTA/L de culture initiale), puis centrifugées à 300g pendant 5min. Après dépôt sur colonne, les billes sont lavées à 5mM et 60mM en imidazole (concentration optimisée par gradient). L'élution est réalisée à 300mM en imidazole.

Les protéines PpLFY1-6his, PpLFY1-D394H-6his, PpLFY1-C427R-6his et PpLFY1-D394H C427R-6his ont été produites par Renaud Dumas à partir des vecteurs indiqués ci-dessus (cf. Construction des vecteurs d'expression).

Chromatographie d'exclusion de taille (*SEC, Size Exclusion Chromatography*)

Le protocole de SEC analytique a été décrit dans l'Article 1 et l'Article 4. La calibration a été réalisée dans les mêmes conditions de tampon à partir du kit Gel filtration calibration (GE Healthcare; thyroglobuline 669 kDa, Ferritine 440 kDa, Aldolase 158 kDa, Conalbumine 75 kDa, Ovalbumin 44

kDa). La masse moléculaire des protéines WelLFY-C, WelNdly-C, WelLFYmin58 et WelNdlymin56 est déduite de l'équation de référence $y = be^{ax}$ où y= poids moléculaire et x=volume d'élution.

Suivi des purifications et dosage

Les différentes fractions (soluble, insoluble, lavages et élution) sont analysées par électrophorèse sur gel SDS-page pour repérer la protéine d'intérêt et évaluer le rendement de la purification.

Gels SDS-page de contrôle de la purification
Gel de concentration: 5% acrylamide/bisacrylamide 37.5:1 (Merck), 62.5mM Tris-HCl pH6.8, 0.1% SDS, 0.1% persulfate d'ammonium (PSA), 0.01% TEMED
Gel de séparation: 10.5 ou 15% acrylamide/bisacrylamide 37.5:1 (Merck), 375mM Tris-HCl, 0.1% SDS, 0.1% PSA, 0.01% TEMED
Les concentrations en acrylamide//bisacrylamide sont ajustées en fonction de la masse moléculaire de la protéine à observer: 10,5% pour le suivi des étapes de purification des protéines entières ou contenant les 2 domaines N- et C-terminal (de masse moléculaire supérieure à 45kDa) ou 15% pour le suivi des étapes de purification des protéines ne contenant que le domaine C-terminal (de masse moléculaire de l'ordre de 25kDa)

Dosage de Bradford
Les protéines pures à l'issue de la filtration sur gel (cf.SEC) ont été quantifiées par dosage au réactif de Bradford (BIO RAD): les valeurs de DO_{595} de trois dilutions de la protéine sont confrontées à une gamme étalon de BSA titrée réalisée dans les mêmes conditions de mesures.

Dosage sur gel d'acrylamide SDS-page
Cette technique de dosage consiste à déposer un volume croissant de protéines sur gel SDS-PAGE et comparer l'intensité de la bande correspondante à la protéine d'intérêt à celle d'une gamme étalon d'actine titrée mise à migrer dans les mêmes conditions. Ces intensités sont estimées à l'aide du logiciel ImageJ. Cette technique a été sollicitée pour le dosage de protéines impures où seule la bande correspondante à la protéine d'intérêt a été quantifiée.

Stockage des protéines

Les protéines correspondant au domaine C-terminal de LFY et ses homologues sont stockées en tampon Tris-HCl pH8, 150 mM NaCl, 2mM MgCl2, 5mM DTT, 5% Glycérol à -80°C après congélation instantanée par passage dans l'azote liquide. Les protéines WelLFYmin58, WelNdlymin56, GinLFYmin40 et AmboLFYmin40 sont stockées en tampon Tri-HCl pH8, 0.8M NaCl, 5mMTCEP à -80°C après congélation instantanée par passage dans l'azote liquide. Les protéines AtLFYmin40 sont stockées en tampon Tri-HCl pH8, 5mM TCEP après congélation instantanée par passage dans l'azote liquide.

3-Caractérisation des interactions ADN-Protéine

Techniques in vitro basées sur le suivi de l'ADN couplé à un fluorophore

Obtention des ADNs fluorescents

ADNs marqués au TAMRA (Tetramethyl-6-Carboxyrhodamine)
Les ADNs double brins *AP1bs1*, *AGbs2*, *AP3bs1*, *AP1bs1-m1 et AP1bs1-m2* (30 pb) ont été obtenus par hybridation entre un oligonucléotide commercial marqué par un fluorophore TAMRA en 5' (SIGMA) et un oligonucléotide complémentaire non marqué dans un tampon d'hybridation 1x (150 mM NaCl, 1 mM EDTA, 10 mM Tris HCl pH 7.5) par passage à 95°C pendant 5 minutes et refroidissement progressif jusqu'à température ambiante à l'abri de la lumière (cf. Matériel et Méthode Article 1 et Article 4). Les séquences correspondantes sont données Chapitre 2 Fig2.4 (AP2bs1, AP1bs1-*m1 et AP1bs1-m2*) ou Matériel et Méthode Article 4, Chapitre 4 (*AGbs2* et *AP3bs1*).

Les ADNs double brin utilisés pour le Selex (73 ou 81pb) ont été synthétisés par PCR à l'aide d'une amorce marquée en 5' par un fluorophore TAMRA (SIGMA) et dosés d'après le protocole décrit dans les Méthodes du Supplementary Online Material Article 2, Chapitre 2.

ADNs marqués au Cy3 ou Cy5 (Cyanine 3 ou 5)
Les ADN utilisés pour les essais QuMFRA (Chapitre 2 et 4) et les ADNS correspondant aux sites *bs-2719* (5'**G**TTAACTTCATAATTAGTGGGTAATAATTAC3'), *bs-1349* (5'**G**CGAACTTATCCCATGCTGGTTAATAGCTCC3') et *bs-794* (5'**G**TAAGTTATTTAAACAGTGGTTAAATGGTAT3') ont été obtenus par annealing d'un oligonucléotide de 30 paires de base et un oligonucléotide complémentaire de 31 paires de base (la séquence de cet oligonucléotide est donnée à côté du nom du site, d**G**TP supplémentaire en 5' indiqué en gras) permettant un marquage au Cy3-dCTP ou Cy5-dCTP (cf. Methods des Supplementary Online Data Article 2, Chapitre 2). Les séquences des différents oligonucléotides sont données dans les Méthodes du Supplementary Online Material Article 2.

Retards sur gel (EMSA)

Réalisation des gels en conditions non-dénaturantes (gels natifs)
Des gels d'acrylamide en conditions non-dénaturantes ont été utilisés pour tester l'interaction protéine/ADN. Les gels réalisés contiennent 6% acrylamide/bisacrylamide 29:1 (Sigma), 0.5X TBE, 0.1% PSA, 0.01% TEMED.

Préparation des échantillons et migration
Les réactions de liaison entre les protéines recombinantes correspondant au domaine de liaison (-C) de LFY et ses homologues ont été préparées dans le binding buffer comme expliqué dans les Articles correspondants : LFY-C (Articles 1, 2 et 3) ; WelLFY-C, WelNdly-C, GFP-WelNdly (Article 4). Les réactions de liaison impliquant des versions mutées de LFY-C (Compléments Chapitre 1), PpLFY-C ou PpLFY-C D394H ont été réalisées dans le binding buffer décrit Article 1.
Les réactions de liaison entre les protéines recombinantes AtLFYmin40, AmboLFYmin40, GinLFYmin40, PpLFY1, PpLFY1 D394H (Chapitre 2) ou WelLFYmin58 et WelNdlymin56 (Chapitre 4) ont été réalisées dans le Buffer 'Eugenio' (10 mM HEPES pH 7.2, 1 mM Spermidine, 14 mM EDTA, 0.3mg/mL BSA, 0.25% CHAPS, 1% Glycérol et 3 mM TCEP).
La migration des gels et la visualisation des résultats sont réalisées suivant les Matériels et Méthodes des Articles 1, 2 et 4.

Anisotropie de fluorescence (AF)

Protocole 1 (Chapitres 1 et 2)
Les premières données d'AF (Chapitres 1 et 2) ont été obtenues à partir du fluorimètre MOS-450 (BioLogic Science Instruments) où les échantillons analysés (ADN + protéines + binding buffer ajusté à 50 µl final) étaient transférés dans des cuves en quartz. Puis, les échantillons ont été excités par un laser émettant une lumière à 546 nm polarisée alternativement selon un axe vertical et horizontal à l'aide d'un modulateur de fréquence 100 kHz. Un photomultiplicateur avec filtre *cut off* centré à 578 nm, placé à 90° de la lumière d'excitation, permet de convertir la fluorescence de l'échantillon mesurée en signal électrique amplifié.

Protocole 2 (Chapitre 4)
Le laboratoire s'étant équipé d'un Tecan Safire^2 (MTX Lab Systems) au cours de ma thèse, nous avons pu procéder dès lors à l'analyse d'échantillons en plaques de 384 puits (cf. Matériel et Méthodes Chapitre 4). les échantillons sont excités par une diode électroluminescente (DEL) produisant un rayonnement à 530 nm. Cette lumière est ensuite polarisée et acheminée au monochromateur final par l'intermédiaire d'un système mettant en jeu trois lentilles optiques: la première lentille collecte et aligne selon un axe vertical la lumière produite par la DEL; la deuxième focalise cette lumière d'excitation polarisée sur les puits puis collecte la lumière plus ou moins polarisée émise par l'échantillon; la troisième focalise la lumière d'émission sur une fibre optique qui transmet le signal lumineux à un monochromateur centré sur 580 nm. Un photomultiplicateur convertit en dernier lieu le

signal lumineux en signal électrique amplifié. 100 mesures sont réalisées par puits, la valeur moyenne d'anisotropie est donnée.

Calcul des valeurs d'AF (Chapitres 1, 2 et 4)

Le calcul d'AF est effectué comme suit (canet et al., 2001) : $AF = (V_v-V_h)/(V_v+0,5V_h)$, où V_v est la lumière émise après excitation selon l'axe vertical, et V_h est la lumière émise après excitation selon l'axe vertical. Ces données permettent de déduire le K_D apparent de la réaction de liaison (en supposant qu'une molécule de LFY se lie à une molécule d'ADN).

En effet, $AF_{mesurée} = A_fF_f + A_bF_b$ avec A_f: AF des oligonucléotides marqués libres

F_f: fraction des oligonucléotides marqués libres

A_b: AF des oligonucléotides marqués liés

F_b: fraction des oligonucléotides marqués liés

LFYADN : complexes LFY/ADN

Sachant $F_f + F_b = 1$ et $F_b = \frac{[LFYADN]}{[ADN_{total}]}$ alors $AF_{mesurée} = A_f + (A_b - A_f)\frac{[LFYADN]}{[ADN_{total}]}$ (Eq1)

En considérant la loi d'action de masse définissant la constante de dissociation apparente (K_D^{App}) dans la réaction $[LFY] + [ADN] \leftrightarrow [LFYADN]$: $K_D = \frac{[LFY][ADN]}{[LFYADN]}$ et la loi de conservation des éléments : $[ADN_{total}] = [ADN] + [LFYADN]$ et $[LFY_{total}] = [LFY] + [LFYADN]$

on peut écrire :

$$[LFYADN]^2 + [LFYADN]\left(-K_D^{App} - [ADN_{total}] - [LFY_{total}]\right) + [ADN_{total}][LFY_{total}] = 0 \text{ (Eq2)}$$

D'après (Eq1) et par résolution de (Eq2) :

$$AF_{mesurée} = A_f + (A_b - A_f)\frac{\left(K_D^{App} + [ADN_{total}] + [LFY_{total}]\right) - \sqrt{\left(K_D^{App} + [ADN_{total}] + [LFY_{total}]\right)^2 - 4[ADN_{total}][LFY_{total}]}}{2[ADN_{total}]}$$

Selex

Protocole 1 (Article 2) : sur résine Ni-NTA

Le protocole de Selex conçu pour la protéine LFY-C est présenté en détail dans la partie Methodes des Supplementary Online Material de l'Article 2.

Protocole 2 (Résultats complémentaires Chapitre 2) : sur billes magnétiques

Afin d'optimiser le protocole 1 pour en faire une méthode haut débit applicable aux homologues de LEAFY et autres protéines, un second protocole de Selex a été mis au point et utilisé avec succès sur les protéines AtLFYmin40, AmboLFY-C, AmboLFYmin40, GinLFY-C, GinLFYmin40, WelLFY-C WelLFYmin58, WelNdly-C, WelNdlymin56, PpLFY1 et PpLFY1 D394H. Marie Monniaux (étudiante en thèse 1ère année) a ensuite appliquée ce protocole aux homologues de vigne (VFLmin40) et riz (RFL-C et RFLmin40). Seules les variations par rapport au protocole 1 sont présentées ci-dessous :

-Synthèse de la banque d'ADN aléatoires: les ADNs double brins fluorescents sont synthétisés par PCR selon le protocole 1 mais un oligonucléotide 73 paires de bases est utilisé (5'TGGAGAAGAGGAGAGATCTAGC$(N)_{30}$CTTGTTCTTCTTCGATTCCGG3') comme matrice.

-Réactions de sélection: pour chaque cycle de sélection, 1 µM de protéine est mélangé à 10 nM d'ADN double brins fluorescent (73-mers) dans 225 µl de Buffer Selex (20 mM Tris pH8, 250 mM NaCl, 2 mM $MgCl_2$, 5 mM TCEP, 60 µg/ml Fish DNA and 1% glycerol). La réaction est incubée 15 min à 4°C sur roue puis 25 µl de billes magnétiques pré-équilibrées dans le Buffer Selex sont ajoutés.

Après 30 minutes d'incubation à 4°C sur roue, le tube eppendorf est placé sur un support aimanté et le surnageant est enlevé puis remplacé par 50 µl de Buffer Selex minus Fish DNA. 2ul de billes sont prélevés pour servir de matrice pour l'amplification des ADNs fluorescents du tour suivant. 250 µl de Buffer Selex contenant 20 µg/ml Fish DNA sont alors ajoutés et la réaction est incubée 2 min sur roue à 4°C. La procédure de lavage est répétée 5 fois, 2 µl de billes étant prélevés uniquement un tour sur deux (tour 0, 2, 4 et 6).

-Amplification des ADNs sélectionnés : 4 réactions PCR sont réalisées comme indiqué dans le protocole 1 en utilisant 1µl de solution de billes magnétiques des tours 0, 2, 4 et 6 comme matrice.

-Quantification des produits PCR : Les produits PCRs sont dosés sur gel d'agarose 3% coloré au SIBR®Safe (Invitrogen) contre une gamme d'ADN.

Clonage et test de séquences individuelles

Pour tester l'enrichissement, les ADNs retenus lors du dernier cycle de Selex sont amplifiés comme précédemment en utilisant le même couple d'amorces mais non-marqué. Le produit PCR est ensuite cloné dans le vecteur PCR Blunt (Invitrogen). Après transformation de bactéries thermocompétentes DH5α, les séquences correspondant à 10 clônes individuelles sont amplifiées directement par PCR sur colonie en utilisant le couple d'amorce Selex marqué au TAMRA. Les produits PCR sont dosés sur gel comme précédemment, puis utilisés directement pour le test d'interaction avec la protéine en retard sur gel.

Construction des matrices

Les étapes de nettoyage des séquences (suppression des bordures fixes, élimination de la redondance), alignement, analyses de la dépendance et calculs de fréquences et poids pour construire les matrices et logo associés ont été décrits dans la partie Methodes du Supplementary Online Data de l'Article 2.

Préparation des échantillons pour séquençage Illumina (Codes barres)

La séquence permettant l'hybridation de l'amorce de séquençage ainsi qu'un code barre de 6 paires de bases sont ajoutés aux ADNs (73 pb) correspondants aux cycles Selex retenus pour le séquençage massif par PCR (15 cycles) en utilisant la Phusion®DNA polymerase (Ozyme) et 0.5 à 2 µl de solution de billes magnétiques + ADN comme matrice. La séquence du code barre est spécifique de chaque cycle Selex. 3 cycles de PCR sont ensuite réalisés en utilisant 5 µl du produit de la PCR1 (146 pb), la Phusion®DNA polymerase (Ozyme) et un couple d'amorce identique pour tous les cycles Selex et permettant la fixation des ADNs à la puce Illumina pour le séquençage. Les produits PCR ainsi obtenus (171 pb) sont purifiés avec le kit NucleoSpin Extract II (Macherey-Nagel). La pureté et la quantité (>150 ng/µl) sont vérifiées sur gel d'agarose 3% puis les échantillons sont conservés à -20°C.

Mesure des affinités relatives par QuMFRA

La méthode du QuMFRA (Man and Stormo 2001) a été utilisée pour mesurer les affinités relatives de différents oligonucléotides pour LFY-C (Article 2) et WelLFY-C ou WelNdly-C (Compléments Chapitres 4). Le protocole suivi a été décrit dans la partie Méthode du Supplementary Online Material de l'Article 2.

Technique in vitro basée sur l'utilisation d'ADN biotinylés

Résonance plasmonique de surface (SPR)

La résonance plasmonique de surface à été utilisée pour tester l'interaction entre facteurs de transcription et longues molécules d'ADN dans les chapitres 2 et 4. La synthèse, biotinylation et purification des ADNs utilisés, la préparation de la puce, l'enregistrement des données et l'analyse des courbes d'association et dissociation des complexes ADN/protéines ont été décrits dans les Matériel et Méthodes des Articles 2 et 4.

Tests en levure

Milieux et cultures

La croissance des levures est obtenue en milieu liquide ou solide, à 30°C, pendant 48 heures. La sélection est réalisée en fonction de leur capacité à croître sur un milieu appauvrie : le(s) vecteur(s) qu'elles reçoivent leur permet en général de produire un acide aminé, du tryptophane par exemple, dont la présence n'est donc plus nécessaire dans le milieu de culture.
Milieu complet YPD: 1% [m/v] yeast extract, 2% [m/v] peptone, 2% [m/v] difco bacto agar, 2% [m/v] glucose filtré
SM (milieu minimum): 0.17% [m/v] YNB, 0.5% [m/v] NH_4SO_4, poudra AA (fonction de la prototrophie), 2% [m/v] glucose ou galacotse filtré + optionnel: 2% agar pour obtenir un milieu solide. Ici, le milieu minimum est UTH- (sans uracile, tryptophane et histidine). On rajoute alors 0,70g/L de poudre UTH- (BIO 101 Systems, Q-BIOgene).

Transformation des levures
Les levures sont mises en culture dans 2mL de milieu SM jusqu'à saturation. 20µL sont ensuite transférés dans 25mL de milieu SM. A DO_{600} égale à 1, les levures sont centrifugées pendant 5min à 3000g puis ressuspendues dans 50mL de milieu préchauffé SM (ou YPD). A DO_{600} égale à 0,6-0,7, les levures sont de nouveau centrifugées puis ressuspendues dans 15mL d'eau stérile. 3mL sont répartis dans 5 tubes eppendorf à raison de 2X1,5mL par tube, avec deux centrifugations à 6000rpm. Sont rajoutés successivement au culot de levures 480µL de PEG 3500 (polyethylène glycol) 50%, 72µL d'acétate de lithium 1M pH7.5, 50µL d'ADN simple brin, et 100µL d'eau stérile contenant entre 50ng et 5µg de plasmide. Le culot est ensuite vortexé vigoureusement jusqu'à son entière dissolution,puis incubé 30min à 30°C. Les levures subissent finalement un choc thermique de 15min à 42°C avant d'être centrifugées, ressuspendues dans 500µL d'eau stérile, et incubées sur boîte SM+agar pendant 2 jours.

Mesure de l'activité β-galactosidase
Pour chaque transformation, entre 5 et 10 colonies isolées sont individuellement repiquées dans 3 mL de milieu sélectif 2% galactose. Le tout est incubé sous agitation continue pendant 6h à 30°C. Les levures sont ensuite centrifugées à 3000g pendant 5min et ressuspendues dans 0,40mL de tampon Z (60mM Na_2HPO_4,$12H_2O$; 40mM NaH_2PO_4, $2H_2O$; 10mM KCl, 1mM $MgSO_4$, $7H_2O$, β-mercaptoéthanol 50mM) et perméabilisées par 3 séries de chocs thermiques '5 min incubation dans de l'azote liquide puis 5 min d'incubation à 37°C dans un bain marie'. Les tubes sont ensuite vigoureusement vortexés et 200 µl sont transférés dans une plaque transparente 96-puit (GRENIER) pour mesurer la densité optique à 595 nm (DO_{595}) au moyen du Safire^2 (Tecan). Les 200ul restants sont également transférés dans une plaque transparente 96-puit (GRENIER) et incubés 5 min à 30°C. Après l'incubation, 40µL d'ONPG à 4mg/mL (stocké dans du KPO_4 0,1M pH7.0) sont ajoutés puis l'activité *lacZ* de catalyse de l'ONPG (incolore) en ONP (jaune) + galactose est suivie par mesure de la densité optique à 420 nm (DO_{420}) en cinétique sur 3h à 30°C grâce au Safire^2 (Tecan).
L'activité β-galactosidase est alors estimée en 'Miller units' à partir du calcul suivant:
Miller Units=1000 x DO_{420}/((volume de cellules) x temps de réaction en min x DO_{595})
Soit Miller Units=1000X pente initiale de la réaction/(volume de cellule X DO_{595})

3-Caractérisation des patrons d'expression

RT-PCR semi Quantitative

L'expression de *WelLFY*, *WelNdly*, *WelActin* et des gènes à boîtes MADS identifiés chez *Welwitschia mirabilis* a été recherchée au niveaux des différents tissus des cônes mâles, femelles et des feuille de *Welwitschia* en suivant le protocole décrit dans le Matériel et Méthode de l'Article 4.

Hybridations *in situ*

Les domaines d'expression de *WelLFY*, *WelNdly*, *WelAG*, *WelB1* et *WelB2* ont été déterminés au niveau de cônes mâles de *W. mirabilis* par hybridation *in situ* comme décrit dans l'Article 4.

4-Analyse de séquences

Alignement et modélisation de la conservation entre homologues de *LEAFY*

Les séquences des homologues de *LEAFY* (>200) chez différentes espèces de plantes terrestres (angiospermes, gymnospermes, ptéridophytes et bryophytes) ont été extraits des bases de données GenBank (http://www.ncbi.nlm.nih.gov) et TIGR transcript assembly database ((Childs et al., 2007); http://plantta.tigr.org) puis alignés à l'aide du programme MUSCLE3.6 (Edgar, 2004). Ce groupe a été complété par les 2 nouveaux homologues clonés au cours de cette thèse (*AmboLFY*, *LuLFY*) ainsi que par les séquences de 2 paralogues, *SmLFY1* et *SmLFY2*, que nous avons identifiés chez *Selaginella moellendorffii* par BLAST de son génome accessible depuis le web : http://www.phytozome.net/search.php?show=blast&method=Org_Smoellendorffii.
L'alignement a ensuite été édité avec JalView v2.5 (Waterhouse et al., 2009) pour ne garder que les séquences correspondantes au domaine C-terminal de chaque homologue. L'alignement final ainsi obtenu a été utilisé pour représenter la conservation des positions sur la structure du complexe LFY/ADN à l'aide du programme ConSurf (Ashkenazy et al., 2010). La modélisation tridimensionnelle obtenue a été visualisée à l'aide du programme Chimera (Pettersen et al., 2004).

Homologues des gènes MADS floraux

Recherche des membres de la sous-famille AG chez les angiospermes séquencés
Nous avons identifié les loci génomiques des représentants des lignages *AGAMOUS* (*AG*), *SEEDSTICK* (*STK*), eu*AG* and *euPLE* chez plusieurs espèces angiospermes dont le génome à été séquencé afin de réaliser une analyse de la probabilité d'occupation de leur second intron par LFY. La procédure complète est décrire dans la partie Methods des Supplementary Online Material (Article 2).

Identification des gènes à boîte MADS chez *W. mirabilis*
La sous-famille d'appartenance de *WelAG*, *WelB1*, *WelB2*, *WelBsister* et *WelAGL6* ont été déterminées par analyses phylogénétiques contre les gènes à boîte MADS de *Gnetum gnemon* et *Gnetum parvifolium* comme explicité dans le Matériel et Méthodes de l'Article 4.

Analyse des données de ChIP-seq

La confrontation des données du ChIP-seq aux prédictions des différentes matrices décrivant les spécificités de liaison de LFY-C ont été effectuées à l'aide des programmes conçus en langage Python par Eugenio G. Minguet et Sandrine Blanchet principalement. Les différents programmes et bases mathématiques des modèles associés sont présentés dans la partie Méthode des Supplementary Online Material de l'Article 2.

REFERENCES

Achard, P., Herr, A., Baulcombe, D.C., and Harberd, N.P. (2004). Modulation of floral development by a gibberellin-regulated microRNA. Development **131,** 3357-3365.

Albert, V.A., Oppenheimer, D.G., and Lindqvist, C. (2002). Pleiotropy, redundancy and the evolution of flowers. Trends Plant Sci. **7,** 297-301.

Ambrose, B.A., Espinosa-Matiàs, S., Vazquez-Santana, S., Vergara-Silva, F., Martinez, E., Marquez-Guzman, J., and Alvarez-Buylla, E.R. (2006). Comparative develomental series of the Mexican triurids support a euanthial interpretation for the unusual reproductive axes of Lacandonia schismatica (Triuridaceae). Am J Bot **93,** 15-35.

Aoki, S., Uehara, K., Imafuku, M., Hasebe, M., and Ito, M. (2004). Phylogeny and divergence of basal angiosperms inferred from APETALA3- and PISTILLATA-like MADS-box genes. J Plant Res **117,** 229-244.

APGIII. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. Botanical Journal of the Linnean Society **161**.

Arendt, D. (2003). Evolution of eyes and photoreceptor cell types. Int J Dev Biol **47,** 563-571.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res **38 Suppl,** W529-533.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.F., Coburn, D., Newburger, D.E., Morris, Q., Hughes, T.R., and Bulyk, M.L. (2009). Diversity and complexity in DNA recognition by transcription factors. Science **324,** 1720-1723.

Barrett, S.C.H. (2008). Major evolutionary transitions in flowering plant reproduction: an overview. Int. J. Plant Sci. **169,** 1-5.

Bateman, R.M., Hilton, J., and Rudall, P.J. (2006). Morphological and molecular phylogenetic context of the angiosperms: contrasting the 'top-down' and 'bottom-up' approaches used to infer the likely characteristics of the first flowers. J Exp Bot **57,** 3471-3503.

Baum, D.A., and Hileman, L.C. (2006). A developmental genetic model for the origin of the flower. In: Flowering and its manipulation—Ainsworth C, ed., Sheffield, UK: Blackwell Publishing. 3-27.

Becker, A., and Lange, M. (2010). VIGS--genomics goes functional. Trends Plant Sci **15,** 1-4.

Bell, C.D., Soltis, D.E., and Soltis, P.S. (2005). The age of the angiosperms: a molecular timescale without a clock. Evolution **59,** 1245-1258.

Berg, O.G., and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol **193,** 723-750.

Birnbaum, K., Jung, J.W., Wang, J.Y., Lambert, G.M., Hirst, J.A., Galbraith, D.W., and Benfey, P.N. (2005). Cell type-specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. Nat Methods **2,** 615-619.

Blazquez, M.A., and Weigel, D. (2000). Integration of floral inductive signals in Arabidopsis. Nature **404,** 889-892.

Blazquez, M.A., Soowal, L.N., Lee, I., and Weigel, D. (1997). LEAFY expression and flower initiation in Arabidopsis. Development **124,** 3835-3844.

Blazquez, M.A., Ferrandiz, C., Madueno, F., and Parcy, F. (2006). How floral meristems are built. Plant Mol Biol **60,** 855-870.

Blazquez, M.A., Green, R., Nilsson, O., Sussman, M.R., and Weigel, D. (1998). Gibberellins promote flowering of arabidopsis by activating the LEAFY promoter. Plant Cell **10,** 791-800.

Bowman, J.L. (1997). Evolutionary conservation of angiosperm flower development at the molecular and genetic levels. Journal of Biosciences **22,** 515-527.

Bowman, J.L., and Meyerowitz, E.M. (1991). Genetic control of pattern formation during flower development in Arabidopsis. Symp Soc Exp Biol **45,** 89-115.

Bowman, J.L., Alvarez, J., Weigel, D., Meyerowitz, E.M., and Smyth, D.R. (1993). Control of Flower development in Arabidopsis thaliana by APETALA1 and interacting genes. Development **119,** 721-743.

Brady, S.M., Orlando, D.A., Lee, J.Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science **318,** 801-806.

Bramwell, D. (2002). Haw many plants are there ? Plant Talk **28,** 32-34.

Bulyk, M.L. (2003). Computational prediction of transcription-factor binding site locations. Genome Biol **5,** 201.

Bulyk, M.L., Gentalen, E., Lockhart, D.J., and Church, G.M. (1999). Quantifying DNA-protein interactions by double-stranded DNA arrays. Nat Biotechnol **17,** 573-577.

Busch, A., and Gleissberg, S. (2003). EcFLO, a FLORICAULA-like gene from Eschscholzia californica is expressed during organogenesis at the vegetative shoot apex. Planta **217,** 841-848.

Busch, M.A., Bomblies, K., and Weigel, D. (1999). Activation of a floral homeotic gene in Arabidopsis. Science **285,** 585-587.

Busch, W., Miotk, A., Ariel, F.D., Zhao, Z., Forner, J., Daum, G., Suzaki, T., Schuster, C., Schultheiss, S.J., Leibfried, A., Haubeiss, S., Ha, N., Chan, R.L., and Lohmann, J.U. (2010). Transcriptional control of a plant stem cell niche. Dev Cell **18,** 849-861.

Buzgo, M., Soltis, D.E., Soltis, P.S., and Ma, H. (2004). Towards a comprehensive integration of morphological and genetic studies of floral development. Trends Plant Sci **9,** 164-173.

Calonje, M., Martin-Bravo, S., Dober, C., Gong, W., Jordon-Thaden, I., Kiefer, C., Paule, J., Schmickl, R., and Koch, M.A. (2009). Non-coding nuclear DNA markers in phylogenetic reconstruction. Pl. Sys. evol **282,** 257-280.

Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. (2001). From DNA to diversity: molecular genetics and the evolution of animal design. (Wiley-Blackwell).

Chae, E., Tan, Q.K., Hill, T.A., and Irish, V.F. (2008). An Arabidopsis F-box protein acts as a transcriptional co-factor to regulate floral development. Development **135,** 1235-1245.

Charron, J.B., He, H., Elling, A.A., and Deng, X.W. (2009). Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. Plant Cell **21,** 3732-3748.

Chaw, S.M., Parkinson, C.L., Cheng, Y., Vincent, T.M., and 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 U S A **97,** 4086-4091.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet **8,** 93-103.

Childs, K.L., Hamilton, J.P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P.D., Town, C.D., Buell, C.R., and Chan, A.P. (2007). The TIGR Plant Transcript Assemblies database. Nucleic Acids Res **35,** D846-851.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., Casero, D., Bernal, M., Huijser, P., Clark, A.T., Kramer, U., Merchant, S.S., Zhang, X., Jacobsen, S.E., and Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. Nature **466,** 388-392.

Clark, R.M., Wagler, T.N., Quijada, P., and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet **38,** 594-597.

Coen, E.S., and Meyerowitz, E.M. (1991). The war of the whorls: genetic interactions controlling flower development. Nature **353,** 31-37.

Coen, E.S., Roemro, J.M., Doyle, S., Elliott, R., Murphy, G., and Carpenter, R. (1990). *Floricaula*: A homeotic gene required for flower development in Antirrhinum majus. Cell **63,** 1311-1322.

Crepet, W.L., and Nicklas, K.J. (2009). Darwin's second 'abominable mystery': Why are there so many angiosperm species?1. Am J Bot **96**.

Darwin, F. (1903). More letters of Charles Darwin, a record of his work in hitherto unpublished letters. (London, UK).

Dembinsky, D., Woll, K., Saleem, M., Liu, Y., Fu, Y., Borsuk, L.A., Lamkemeyer, T., Fladerer, C., Madlung, J., Barbazuk, B., Nordheim, A., Nettleton, D., Schnable, P.S., and Hochholdinger, F. (2007). Transcriptomic and proteomic analyses of pericycle cells of the maize primary root. Plant Physiol **145,** 575-588.

Doyle, J.A. (2006). Seed ferns and the origin of angiosperms. the journal of the Torrey Botanical Society **133**.

Doyle, J.A. (2008). Integrating molecular phylogenetic and paleobotanical evidence on the origin of the flower. Int. J. Plant Sci. **169,** 816-843.

Drinnan, A.N., Crane, P.R., and Hoot, S.B. (1994). Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). Pl. Sys. evol **8,** 93-122.

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

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

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature **346,** 818-822.

Endress, P.K. (2001). Origins of flower morphology. J Exp Zool **291,** 105-115.

Endress, P.K. (2010). The evolution of floral biology in basal angiosperms. Philos Trans R Soc Lond B Biol Sci **365,** 411-421.

Endress, P.K., and Friis, E.M. (1991). Archamamelis, hamamelidalean flowers from the Upper Cretaceous of Sweden. Plant Systematics and Evolution **175**.

Endress, P.K., and Doyle, J.A. (2007). Floral phyllotaxis in basal angiosperms: development and evolution. Curr Opin Plant Biol **10,** 52-57.

Endress, P.K., and Doyle, J.A. (2009). Reconstructing the ancestral angiosperm flower and its initial specializations. Am J Bot **96,** 22-66.

Eriksson, S., Bohlenius, H., Moritz, T., and Nilsson, O. (2006). GA4 is the active gibberellin in the regulation of LEAFY transcription and Arabidopsis floral initiation. Plant Cell **18,** 2172-2181.

Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nat Rev Genet **10,** 605-616.

Fourquin, C., Vinauger-Douard, M., Fogliani, B., Dumas, C., and Scutt, C.P. (2005). Evidence that CRABS CLAW and TOUSLED have conserved their roles in carpel development since the ancestor of the extant angiosperms. Proc Natl Acad Sci U S A **102,** 4649-4654.

Friis, E.M., and Crane, P. (2007). Botany: new home for tiny aquatics. Nature **446,** 269-270.

Friis, E.M., Pedersen, K.R., and Crane, P.R. (2001). Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. Nature **410,** 357-360.

Friis, E.M., RaunsgaardPedersen, K.R., and Crane, P.R. (2006). Cretaceous angiosperm flowers: innovation and evolution in plant reproduction. Palaeogeography, Palaeoclimatology, Palaeocology **232,** 251-293.

Friis, E.M., Pedersen, K.R., and Crane, P.R. (2010). Diversity in obscurity: fossil flowers and the early history of angiosperms. Philos Trans R Soc Lond B Biol Sci **365,** 369-382.

Friis, E.M., Doyle, J.A., Endress, P.K., and Leng, Q. (2003). Archaefructus--angiosperm precursor or specialized early angiosperm? Trends Plant Sci **8,** 369-373.

Frohlich, M.W. (2003). An evolutionary scenario for the origin of flowers. Nat Rev Genet **4,** 559-566.

Frohlich, M.W. (2006). Recent developments regarding the evolutionary origin of flowers. Advances in botanical research **44**.

Frohlich, M.W., and 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**.

Frohlich, M.W., and Chase, M.W. (2007). After a dozen years of progress the origin of angiosperms is still a great mystery. Nature **450,** 1184-1189.

Frohlich, M.W., and Parker,D.S. (2000). The mostly male theory of flower evolutionary origins: from genes to fossils. Systematic Botany **25,** 155-170.

Gandolfo, M.A., Nixon, K.C., and Crepet, W.L. (2004). Cretaceous flowers of Nymphaeaceae and implications for complex insect entrapment pollination mechanisms in early Angiosperms. PNAS **93,** 10274-10279.

Gao, F., Foat, B.C., and Bussemaker, H.J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics **5,** 31.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature **457,** 215-218.

Gopinath, S.C. (2007). Methods developed for SELEX. Anal Bioanal Chem **387,** 171-182.

Gould, S.J. (1977). Ontogeny and Phylogeny.

Grant, V. (1971). Plant speciation. (New York).

Gregor, T., Tank, D.W., Wieschaus, E.F., and Bialek, W. (2007). Probing the limits to positional information. Cell **130,** 153-164.

Himi, S., Sano, R., Nishiyama, T., Tanahashi, T., Kato, M., Ueda, K., and Hasebe, M. (2001). Evolution of MADS-box gene induction by FLO/LFY genes. J Mol Evol **53,** 387-393.

Hoekstra, H.E., and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. Evolution **61,** 995-1016.

Hofer, J., Turner, L., Hellens, R., Ambrose, M., Matthews, P., Michael, A., and Ellis, N. (1997). UNIFOLIATA regulates leaf and flower morphogenesis in pea. Curr Biol **7,** 581-587.

Hong, R.L., Hamaguchi, L., Busch, M.A., and Weigel, D. (2003). Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell **15,** 1296-1309.

Hu, Z., Killion, P.J., and Iyer, V.R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet **39,** 683-687.

Immink, R.G., Tonaco, I.A., de Folter, S., Shchennikova, A., van Dijk, A.D., Busscher-Lange, J., Borst, J.W., and Angenent, G.C. (2009). SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biol **10,** R24.

Irish, V.F. (2010). The flowering of Arabidopsis flower development. Plant J. **61,** 1014-1028.

Jamal Rahi, S., Virnau, P., Mirny, L.A., and Kardar, M. (2008). Predicting transcription factor specificity with all-atom models. Nucleic Acids Res **36,** 6209-6217.

Janky, R., Helden, J., and Babu, M.M. (2009). Investigating transcriptional regulation: from analysis of complex networks to discovery of cis-regulatory elements. Methods **48,** 277-286.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.B., Peery, R., McNeal, J.R., Kuehl, J.V., and Boore, J.L. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A **104,** 19369-19374.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J., Bonke, M., Palin, K., Talukder, S., Hughes, T.R., Luscombe, N.M., Ukkonen, E., and Taipale, J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res **20,** 861-873.

Kaufmann, K., Muino, J.M., Jauregui, R., Airoldi, C.A., Smaczniak, C., Krajewski, P., and Angenent, G.C. (2009). Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLoS Biol **7,** e1000090.

Kim, S., Koh, J., Yoo, M.J., Kong, H., Hu, Y., Ma, H., Soltis, P.S., and Soltis, D.E. (2005). Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators. Plant J **43,** 724-744.

Kramer, E.M., Di Stilio, V.S., and Schluter, P.M. (2003). Complex patterns of gene duplication in the APETALLA3 and PISTILLATA lineages of the Ranunculaceae. Int. J. Plant Sci. **164,** 1-11.

Kramer, E.M., Jaramillo, M.A., and Di Stilio, V.S. (2004). Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. Genetics **166,** 1011-1023.

Kramer, E.M., Holappa, L., Gould, B., Jaramillo, M.A., Setnikov, D., and Santiago, P.M. (2007). Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot Aquilegia. Plant Cell **19,** 750-766.

Lamb, R.S., Hill, T.A., Tan, Q.K., and Irish, V.F. (2002). Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. Development **129,** 2079-2086.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res **33,** W299-302.

Lebrecht, D., Foehr, M., Smith, E., Lopes, F.J., Vanario-Alonso, C.E., Reinitz, J., Burz, D.S., and Hanes, S.D. (2005). Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. Proc Natl Acad Sci U S A **102,** 13176-13181.

Lee, I., Wolfe, D.S., Nillson, O., and Weigel, D. (1997). A *LEAFY* co-regulator encoded by *UNUSUAL FLORAL ORGANS*. Current Biol. **7,** 95-104.

Lee, J., Oh, M., Park, H., and Lee, I. (2008). SOC1 translocated to the nucleus by interaction with AGL24 directly regulates leafy. Plant J **55,** 832-843.

LeTilly, V., and Royer, C.A. (1993). Fluorescence anisotropy assays implicate protein-protein interactions in regulating trp repressor DNA binding. Biochemistry **32,** 7753-7758.

Levin, J.Z., and Meyerowitz, E.M. (1995). UFO: an Arabidopsis gene involved in both floral meristem and floral organ development. Plant Cell **7,** 529-548.

Liljegren, S.J., Gustafson-Brown, C., Pinyopich, A., Ditta, G.S., and Yanofsky, M.F. (1999). Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate. Plant Cell **11,** 1007-1018.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133,** 523-536.

Litt, A., and Kramer, E.M. (2010). The ABC model and the diversification of floral organ identity. Semin Cell Dev Biol **21,** 129-137.

Liu, C., Thong, Z., and Yu, H. (2009a). Coming into bloom: the specification of floral meristems. Development **136,** 3379-3391.

Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H. (2009b). Regulation of floral patterning by flowering time genes. Dev Cell **16,** 711-722.

Lohmann, J.U., and Weigel, D. (2002). Building beauty: the genetic control of floral patterning. Dev Cell **2,** 135-142.

Lohmann, J.U., Hong, R.L., Hobe, M., Busch, M.A., Parcy, F., Simon, R., and Weigel, D. (2001). A molecular link between stem cell regulation and floral patterning in Arabidopsis. Cell **105,** 793-803.

Long, T.A., Brady, S.M., and Benfey, P.N. (2008). Systems approaches to identifying gene regulatory networks in plants. Annu Rev Cell Dev Biol **24,** 81-103.

Macisaac, K.D., Gordon, D.B., Nekludova, L., Odom, D.T., Schreiber, J., Gifford, D.K., Young, R.A., and Fraenkel, E. (2006). A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. Bioinformatics **22,** 423-429.

Maerkl, S.J., and Quake, S.R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. Science **315,** 233-237.

Maizel, A., Busch, M.A., Tanahashi, T., Perkovic, J., Kato, M., Hasebe, M., and Weigel, D. (2005). The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science **308,** 260-263.

Majka, J., and Speck, C. (2007). Analysis of protein-DNA interactions using surface plasmon resonance. Adv Biochem Eng Biotechnol **104,** 13-36.

Manke, T., Roider, H.G., and Vingron, M. (2008). Statistical modeling of transcription factor binding affinities predicts regulatory interactions. PLoS Comput Biol **4,** e1000039.

McDonnell, J.M. (2001). Surface plasmon resonance: towards an understanding of the mechanisms of biological molecular recognition. Curr Opin Chem Biol **5,** 572-577.

Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. Science **324,** 407-410.

Moczek, A.P., and Rose, D.J. (2009). Differential recruitment of limb patterning genes during development and diversification of beetle horns. Proc Natl Acad Sci U S A **106,** 8992-8997.

Moon, J., Lee, H., Kim, M., and Lee, I. (2005). Analysis of flowering pathway integrators in Arabidopsis. Plant Cell Physiol **46,** 292-299.

Moore, M.J., Bell, C.D., Soltis, P.S., and Soltis, D.E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A **104,** 19363-19368.

Moyroud, E., Tichtinsky, G., and Parcy, F. (2009). The LEAFY floral regulators in Angiosperms: Conserved proteins with diverse roles. J. Plant Biol. **52**.

Moyroud, E., Kusters, E., Monniaux, M., Koes, R., and Parcy, F. (2010). LEAFY blossoms. Trends Plant Sci.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A., and Bulyk, M.L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet **36,** 1331-1339.

Münster, T., Faigl, W., Saedler, H., and Theissen, G. (2002). Evolutionary aspects of MADS-box genes in the eusporangiate fern Ophioglossum. Plant Biology **4,** 474-483.

Nakazono, M., Qiu, F., Borsuk, L.A., and Schnable, P.S. (2003). Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: identification of genes expressed differentially in epidermal cells or vascular tissues of maize. Plant Cell **15,** 583-596.

Ng, M., and Yanofsky, M.F. (2001). Function and evolution of the plant MADS-box gene family. Nat Rev Genet **2,** 186-195.

Nilsson, O., Lee, I., Blazquez, M.A., and Weigel, D. (1998). Flowering-time genes modulate the response to LEAFY activity. Genetics **150,** 403-410.

Pabo, C.O., and Sauer, R.T. (1984). Protein-DNA recognition. Annu Rev Biochem **53,** 293-321.

Parcy, F. (2005). Flowering: a time for integration. Int J Dev Biol **49,** 585-593.

Parcy, F., Nilsson, O., Busch, M.A., Lee, I., and Weigel, D. (1998). A genetic framework for floral patterning. Nature **395,** 561-566.

Paton, A., Brummitt, N., Govaerts, R., Harman, K., Hinchcliffe, S., B., A., and Lughadha, E.N. (2008). Toward target 1 of the global strategy for Plant conservation: a working list of all known plant species - progress and prospects. Taxon **57,** 602-611.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem **25,** 1605-1612.

Preston, J.C., and Hileman, L.C. (2010). SQUAMOSA-PROMOTER BINDING PROTEIN 1 initiates flowering in Antirrhinum majus through the activation of meristem identity genes. Plant J **62,** 704-712.

Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. (1995). The CONSTANS gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. Cell **80,** 847-857.

Qiu, Y.L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., Dombrovska, O., Lee, J., Kent, L., Rest, J., Estabrook, G.F., Hendry, T.A., Taylor, D.W., Testa, C.M., Ambros, M., Crandall-Stotler, B., Duff, R.J., Stech, M., Frey, W., Quandt, D., and Davis, C.C. (2006). The deepest divergences in land plants inferred from phylogenomic evidence. Proc Natl Acad Sci U S A **103,** 15511-15516.

Quatrano, R.S., McDaniel, S.F., Khandelwal, A., Perroud, P.F., and Cove, D.J. (2007). Physcomitrella patens: mosses enter the genomic age. Curr Opin Plant Biol **10,** 182-189.

Quinn, J.G., O'Neill, S., Doyle, A., McAtamney, C., Diamond, D., MacCraith, B.D., and O'Kennedy, R. (2000). Development and application of surface plasmon resonance-based biosensors for the detection of cell-ligand interactions. Anal Biochem **281,** 135-143.

Rao, N.N., Prasad, K., Kumar, P.R., and Vijayraghavan, U. (2008). Distinct regulatory role for RFL, the rice LFY homolog, in determining flowering time and plant architecture. Proc Natl Acad Sci U S A **105,** 3646-3651.

Regal, P.J. (1977). Ecology and evolution of flowering plant dominance. Science **196,** 622-629.

Ren, N., Zhu, C., Lee, H., and Adler, P.N. (2005). Gene expression during Drosophila wing morphogenesis and differentiation. Genetics **171,** 625-638.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin, I.T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W.B., Barker, E., Bennetzen, J.L., Blankenship, R., Cho, S.H., Dutcher, S.K., Estelle, M., Fawcett, J.A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K.A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D.R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P.J., Sanderfoot, A., Schween, G., Shiu, S.H., Stueber, K., Theodoulou, F.L., Tu, H., Van de Peer, Y., Verrier, P.J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A.C., Hasebe, M., Lucas, S., Mishler, B.D., Reski, R., Grigoriev, I.V., Quatrano, R.S., and Boore, J.L. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319,** 64-69.

Retallack, G., and Dilcher, D.L. (1981). Arguments for a glossopterid ancestry of angiosperms. Paleobiology **7,** 54-67.

Rister, J., and Desplan, C. (2010). Deciphering the genome's regulatory code: the many languages of DNA. Bioessays **32,** 381-384.

Rudall, P.J., and Bateman, R.M. (2010). Defining the limits of flowers: the challenge of distinguishing between the evolutionary products of simple versus compound strobili. Philos Trans R Soc Lond B Biol Sci **365,** 397-409.

Rutledge, R., Regan, S., Nicolas, O., Fobert, P., Cote, C., Bosnich, W., Kauffeldt, C., Sunohara, G., Seguin, A., and 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,** 625-634.

Saarela, J.M., Rai, H.S., Doyle, J.A., Endress, P.K., Mathews, S., Marchant, A.D., Briggs, B.G., and Graham, S.W. (2007). Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature **446,** 312-315.

Schaefer, D.G. (2001). Gene targeting in Physcomitrella patens. Curr Opin Plant Biol **4,** 143-150.

Schmid, M., Uhlenhaut, N.H., Godard, F., Demar, M., Bressan, R., Weigel, D., and Lohmann, J.U. (2003). Dissection of floral induction pathways using global expression analysis. Development **130,** 6001-6012.

Schultz, E.A., and Haughn, G.W. (1991). *LEAFY*, a homeotic gene that regulates inflorescence development in Arabidopsis. Plant Cell **3,** 771-781.

Segal, E., and Widom, J. (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. Nat Rev Genet **10,** 443-456.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature **451,** 535-540.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature **442,** 772-778.

Sessions, A., Yanofsky, M.F., and Weigel, D. (2000). Cell-cell signaling and movement by the floral transcription factors LEAFY and APETALA1. Science **289,** 779-782.

Singer, S.D., Krogan, N.T., and Ashton, N.W. (2007). Clues about the ancestral roles of plant MADS-box genes from a functional analysis of moss homologues. Plant Cell Rep **26,** 1155-1169.

Sliwinski, M.K., White, M.A., Maizel, A., Weigel, D., and Baum, D.A. (2006). Evolutionary divergence of *LFY* function in the mustards *Arabidopsis thaliana* and *Leavenworthia crassa*. Plant Mol Biol **62,** 279-289.

Soltis, D.E., Soltis, P.S., Endress, P.K., and Chase, M.W. (2005). Phylogeny and evolution of angiosperms. (Sunderland, Massachusetts USA).

Soltis, D.E., Chanderbali, A.S., Kim, S., Buzgo, M., and Soltis, P.S. (2007). The ABC model and its applicability to basal angiosperms. Ann Bot **100,** 155-163.

Souer, E., Rebocho, A.B., Bliek, M., Kusters, E., de Bruin, R.A., and Koes, R. (2008). Patterning of inflorescences and flowers by the F-Box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of petunia. Plant Cell **20,** 2033-2048.

Souer, E., van der Krol, A., Kloos, D., Spelt, C., Bliek, M., Mol, J., and Koes, R. (1998). Genetic control of branching pattern and floral identity during Petunia inflorescence development. Development **125,** 733-742.

Specht, C.D., and Bartlett, M.E. (2009). Flower evolution: the origin and subsequent diversification of the angiosperm flower. Annu. Rev. Ecol. Evol. Syst. **40,** 217-243.

Spencer, M.W., Casson, S.A., and Lindsey, K. (2007). Transcriptional profiling of the Arabidopsis embryo. Plant Physiol **143,** 924-940.

Stellari, G.M., Jaramillo, M.A., and Kramer, E.M. (2004). Evolution of the APETALA3 and PISTILLATA lineages of MADS-box-containing genes in the basal angiosperms. Mol Biol Evol **21,** 506-519.

Stoltenburg, R., Reinemann, C., and Strehlitz, B. (2007). SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol Eng **24,** 381-403.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics **16,** 16-23.

Sun, G., Ji, Q., Dilcher, D.L., Zheng, S., Nixon, K.C., and Wang, X. (2002). Archaefructaceae, a new basal angiosperm family. Science **296,** 899-904.

Tanahashi, T., Sumikawa, N., Kato, M., and Hasebe, M. (2005). Diversification of gene function: homologs of the floral regulator FLO/LFY control the first zygotic cell division in the moss Physcomitrella patens. Development **132,** 1727-1736.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res **16,** 962-972.

Tandre, K., Albert, V.A., Sundas, A., and Engstrom, P. (1995). Conifer homologues to genes that control floral development in angiosperms. Plant Mol Biol **27,** 69-78.

Tang, W., Newton, R.J., and Weidner, D.A. (2007). Genetic transformation and gene silencing mediated by multiple copies of a transgene in eastern white pine. J Exp Bot **58,** 545-554.

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

Theissen, G., and Becker, A. (2004). Gymnosperms orthologs of class B floral homeotic genes and their impact on understanding flower origin. Critical Reviews in Plant Science **23,** 129-148.

Theissen, G., and Melzer, R. (2007). Molecular mechanisms underlying origin and diversification of the angiosperm flower. Ann. Bot. (London) **100,** 603-619.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23,** 137-144.

Triezenberg, S.J., Kingsbury, R.C., and McKnight, S.L. (1988). Functional dissection of VP16, the *trans*-activator of herpes simplex virus immediate early gene expression. Genes Dev. **2,** 718-729.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249,** 505-510.

Vingron, M., Brazma, A., Coulson, R., van Helden, J., Manke, T., Palin, K., Sand, O., and Ukkonen, E. (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. Genome Biol **10,** 202.

Wagner, G.P. (2007). The developmental genetics of homology. Nat Rev Genet **8,** 473-479.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10,** 57-63.

Ward, L.D., and Bussemaker, H.J. (2008). Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics **24,** i165-171.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet **5,** 276-287.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics **25,** 1189-1191.

Wei, G.H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H.G., Ukkonen, E., Hughes, T.R., Bulyk, M.L., and Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J **29,** 2147-2160.

Weigel, D., and Nilsson, O. (1995). A developmental switch sufficient for flower initiation in diverse plants. Nature **377,** 495-500.

Weigel, D., Alvarez, J., Smyth, D.R., Yanofsky, M.F., and Meyerowitz, E.M. (1992). LEAFY controls floral meristem identity in Arabidopsis. Cell **69,** 843-859.

Wilkins, A.S. (2002). The evolution of developmental pathways. (Sunderland, Massachusetts, USA).

William, D.A., Su, Y., Smith, M.R., Lu, M., Baldwin, D.A., and Wagner, D. (2004). Genomic identification of direct target genes of LEAFY. Proc Natl Acad Sci U S A **101,** 1775-1780.

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

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat Rev Genet **8,** 206-216.

Wu, S., Wang, J., Zhao, W., Pounds, S., and Cheng, C. (2010). ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data. Theor Biol Med Model **7,** 18.

Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet **25,** 434-440.

Yamaguchi, A., Wu, M.F., Yang, L., Wu, G., Poethig, R.S., and Wagner, D. (2009). The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. Dev Cell **17,** 268-278.

Yoon, H.S., and Baum, D.A. (2004). Transgenic study of parallelism in plant morphological evolution. Proc Natl Acad Sci U S A **101,** 6524-6529.

Yu, H., Ito, T., Wellmer, F., and Meyerowitz, E.M. (2004). Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. Nat Genet **36,** 157-161.

Zhang, P., Tan, H.T., Pwee, K.H., and Kumar, P.P. (2004). Conservation of class C function of floral organ development during 300 million years of evolution from gymnosperms to angiosperms. Plant J **37,** 566-577.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J., and Jacobsen, S.E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. PLoS Biol **5,** e129.

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature **462,** 65-70.

Abstract

Flowers are a key innovation in plant evolution and their origin remains a mystery. *LEAFY* (*LFY*) is a unique plant transcription factor regulating floral development, but this gene predates flowers. My thesis work aimed to understand how the evolution of LFY biochemical properties could help explaining flower origins.
First, I took part in the structural characterization of LFY DNA-binding domain, revealing a novel protein fold bound to DNA as a cooperative dimer (*Hamès et al., 2009*). To exhaustively characterize its DNA binding specificity, I set up a SELEX experiment that yielded hundreds of sequences with high-affinity for LFY. Based on these sequences, I built a LFY binding site predictive model (position weight matrix) that I validated *in vitro*. This matrix led to a biophysical model, able to explain the set of genomic regions bound by LFY *in vivo*, as established in a ChIP-seq experiment in *Arabidopsis* seedlings, but also to detect functional orthologs of LFY targets genes in other angiosperms genomes.
Since *LFY* is present in mosses, ferns and gymnosperms that do not flower, I tested if its DNA binding specificity is different in these groups. In parallel, I uncovered correlations between the expression patterns of *LFY* genes and several homologs of floral homeotic genes in the gymnosperm, *Welwitschia mirabilis.* After demonstrating that surface plasmon resonance can be used to analyse interactions between transcription factors and entire gene promoters (*Moyroud et al., 2009*), I tested the interplay between LFY and its potential targets in *Welwitschia* and established that a pre-floral network was already at work in non-flowering plants.

Résumé

Les fleurs sont une innovation cruciale du monde végétal et leur origine demeure mystérieuse. Le développement floral est contrôlé par LEAFY (LFY), un facteur de transcription original qui existait déjà chez les plantes pré-angiospermes. Le but de ma thèse a été d'étudier si l'évolution des propriétés de LFY pouvait aider à comprendre l'apparition de la structure florale.
J'ai contribué à la caractérisation structurale du domaine de liaison à l'ADN de LFY d'Arabidopsis, qui a révélé un repliement inédit contactant l'ADN sous forme coopérative (Hamès et al., 2008). Ensuite, j'ai utilisé la technique de SELEX pour isoler 500 séquences de haute affinité pour LFY et bâtir un modèle biophysique prédisant l'affinité de ce facteur pour toute séquence d'ADN. Ce modèle a été validé par des mesures *in vitro* et par confrontation à l'ensemble des cibles génomiques de LFY révélées par ChIP-seq. Appliqué aux génomes d'autres angiospermes, ce modèle permet de repérer les cibles de LFY au sein d'un groupe de paralogues.
LFY étant présent chez des plantes sans fleurs comme les mousses et les gymnospermes, j'ai testé par Selex si sa spécificité de liaison à l'ADN était différente chez ces plantes. Enfin, j'ai montré des corrélations d'expression entre les gènes LFY et plusieurs homologues des gènes homéotiques floraux chez la gymnosperme *Welwitschia mirabilis*. Après avoir établi que la technique de résonnance plasmonique de surface détectait des interactions entre une protéine et de grands fragments d'ADN (Moyroud et al., 2009), j'ai montré la capacité de LFY de Welwitschia à reconnaître les promoteurs de ses cibles potentielles, mettant ainsi en évidence l'existence d'un réseau pré-floral.