



Predicting and understanding transcriptional regulation of floral development by LEAFY

Laura Turchi

► To cite this version:

Laura Turchi. Predicting and understanding transcriptional regulation of floral development by LEAFY. Vegetal Biology. Université Grenoble Alpes [2020-..], 2023. English. NNT : 2023GRALV064 . tel-04465424

HAL Id: tel-04465424

<https://theses.hal.science/tel-04465424v1>

Submitted on 19 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : CSV- Chimie et Sciences du Vivant

Spécialité : Biologie Végétale

Unité de recherche : LPCV - Laboratoire de Physiologie Cellulaire Végétale

Prédire et comprendre la régulation transcriptionnelle du développement floral par LEAFY

Predicting and understanding transcriptional regulation of floral development by LEAFY

Présentée par :

Laura TURCHI

Direction de thèse :

François PARCY

Directeur de Recherche CNRS, Université Grenoble Alpes

Antoine FRENOY

MAITRE DE CONFERENCE, Université Grenoble Alpes

Directeur de thèse

Co-encadrant de thèse

Rapporteurs :

Klaas VANDEPOELE

FULL PROFESSOR, Universiteit Gent

Marie-Laure MARTIN

DIRECTRICE DE RECHERCHE, INRAE Ile-de-France - Versailles-Saclay

Thèse soutenue publiquement le **26 octobre 2023**, devant le jury composé de :

François PARCY

DIRECTEUR DE RECHERCHE, CNRS délégation Alpes

Klaas VANDEPOELE

FULL PROFESSOR, Universiteit Gent

Marie-Laure MARTIN

DIRECTRICE DE RECHERCHE, INRAE Ile-de-France - Versailles-Saclay

Cristel CARLES

PROFESSEURE DES UNIVERSITES, Université Grenoble-Alpes

Gabriel KROUK

DIRECTEUR DE RECHERCHE, CNRS délégation Occitane Est

Directeur de thèse

Rapporteur

Rapporteuse

Présidente

Examineur

Invités :

Romain Blanc-Mathieu

CHARGE DE RECHERCHE,

Nicolas Thierry-Mieg

CHARGE DE RECHERCHE,



Predicting and understanding transcriptional regulation of floral development by LEAFY

The role of genomic context and cofactors, and the elusive impact of evolutionary conservation

Acknowledgements

First and foremost, I would like to sincerely thank all jury members for agreeing to serve on this committee and for taking the time to evaluate my work, I truly enjoyed discussing with all of you.

I would also like to thank the members of my PhD committee, who helped me shape and steer my PhD project(s) in the past three years by offering their expertise and an external point of view while constantly supporting my work.

I am also extremely grateful to my team of supervisors, without whom I would not be here today. Thank you all for believing in me since the very beginning, for supporting and challenging me in the past three years while giving me time and space to figure things out on my own. I am amazed by how we always managed to find common ground for fruitful scientific discussions despite our different yet complementary expertise. I cannot fully express my gratitude in words, but I will do my best. Merci François pour ta curiosité et ta motivation inépuisables, pour l'énergie et la passion avec lesquelles tu te travailles pour rendre la science et la recherche accessibles au plus grand nombre. Surtout, merci d'avoir eu confiance en moi et en mes capacités dès mon stage de « découverte bioinformatique », sans lequel je n'aurais pas pu entreprendre le chemin de cette thèse. Merci Antoine pour ta disponibilité et ta porte toujours ouverte, pour ta patience, pour ton investissement dans ton rôle de formateur sous tous ses aspects et pour l'attention que tu as toujours portée à tous les membres de l'équipe. Tu as été un modèle de science et de recherche sereine, rigoureuse, engagée et attentive. Merci Romain d'avoir toujours été là quand j'avais besoin de toi, surtout ces derniers mois, que ce soit pour de petites questions ou de plus grandes discussions, ainsi que pour des moments de détente. Merci Nico pour tes conseils et ton souci du détail, qui m'ont toujours aidée à prendre du recul par rapport au projet et à porter une attention particulière aux analyses. Merci Jérémy d'avoir été si patient avec moi dès le début, en m'accueillant comme stagiaire malgré ma totale inexpérience en bio-informatique et en programmation. Merci pour les discussions au bureau, pour ton aide quand j'en avais besoin et plus généralement pour m'avoir montré, petit à petit, l'importance et la puissance de cette discipline que j'adore aujourd'hui.

Science is a collective effort, and support from colleagues and friends has been crucial to get over the disappointments and slow patches that one inevitably faces while doing a PhD. Un grand merci à toute l'équipe de Flo_Re pour m'avoir accueillie à bras ouverts, pour les échanges scientifiques lors des réunions et pour les bons moments passés lors des journées et des retraites d'équipe. Un merci spécial à Gaby, pour son soutien et sa patience depuis le master, puis pendant les expériences que nous avons menées ensemble. Merci à Manu, Renaud, Moïra, Philippe, Loïc, Marianne et tous les stagiaires qui ont travaillé avec nous pendant ces trois ans. Un grand merci à l'équipe MAGE, pour l'ambiance toujours détendue et les goûters partagés, ainsi que pour l'effort constant de trouver un terrain commun de discussion et d'entraide malgré nos expertises parfois éloignées. Une mention spéciale pour Amandine, avec laquelle j'ai partagé l'intégralité de mon parcours de thèse, entre repas gourmands et accrobranche, aide technique et support moral dans les moments les plus

difficiles. Merci à Olivier, Mag, Nagi, Florence, Zakaria, Chloe, Elise et tous ceux qui ont fait partie de l'équipe au cours de ces trois ans. Merci à mes laboratoires d'accueil, LPCV et TIMC, et en particulier merci à la communauté « Science for dummies » du LPCV, qui a été une source de soutien et motivation à plusieurs reprises. Merci aussi à Tiffany et Sophie (la meilleure team admin du monde !) de m'avoir toujours aidée, jusqu'au dernier moment. Un grand merci à Camille, pour les apéros, l'escalade, les escapades dans le sud de la France, mais aussi pour avoir écouté et partagé dans les moments de stress de la thèse.

Since it is vital to have interests and sources of joy and fulfillment out of the lab, and I am grateful for the wide network of friends that have supported me in a variety of ways throughout this journey. Grazie Emma per avermi affiancata, di fatto, dall'inizio del dottorato fino alla fine, per aver sempre rappresentato un'isola felice di interessi, attività e creatività, nonché la mia unica parentesi italiana in terra francese. Merci à Zoé et Irene, best colocs ever, et au groupe du jardin, qui a commencé comme une drôle aventure et qui m'accompagne toujours.

Un grazie speciale alla famiglia e agli amici che mi hanno supportata (e sopportata) in questi ultimi anni, rispettando le mie scelte nonostante la distanza. Grazie per esserci sempre e per accogliermi ogni volta come se non fossi mai partita, spero che non cambi mai.

Last but not least, I am not sure where I'd be today if it weren't for Vangeli. Thank you for always being there for me since the beginning, for your endless patience and support (especially in the past few months), and more generally for being the best partner I could ever ask for.

Summary

Ensuring correct gene expression is crucial for living organisms, as its disruption can compromise survival. Transcription factors (TFs) regulate gene expression through the binding of specific DNA sequences called transcription factor binding sites (TFBSs). LEAFY (LFY) is a plant-specific TF with a crucial role in floral development, and it is highly conserved in sequence and binding specificity throughout flowering plant evolution.

LFY's central role in flowering has been studied for decades, and yet it remains unclear why only a subset of the genomic regions bound by LFY are regulated. To elucidate this point, I present, in the first part of this manuscript, an approach to predict transcriptional regulation of LFY TFBSs in the model plant *Arabidopsis thaliana*. I used state-of-the-art LFY TFBS models and the genetically-encoded genomic context of LFY sites to successfully build a classifier that can distinguish functional LFY sites (i.e. TFBSs that are bound and have an effect on gene expression *in vivo*) from nonfunctional ones (i.e. TFBSs that are not bound and are not associated with gene expression changes *in vivo*). My results suggest that the presence of surrounding LFY TFBSs and, to a lesser extent, the level of non-LFY TFBS diversity around LFY sites, are important to distinguish functional and nonfunctional LFY sites. Moreover, this approach reveals a number of co-occurring TFs that contribute to set apart LFY-regulated sites from nonfunctional ones. Despite previous evidence of the functional importance of conserved regions in gene regulation, including conservation of LFY sites in our model did not improve predictions, and I discuss some possible reasons behind this result. Overall, this approach allowed me to further characterize LFY's binding to DNA, and it can be used on new genomic sequences to predict transcriptional regulation of LFY sites, as well as with new TFs.

In addition to working on its own, LFY interacts with UNUSUAL FLORAL ORGANS (UFO), an F-box protein, to ensure correct petal and stamen development. While the LFY-UFO interaction and their implication in flower development were already known, the exact role of UFO in this process had yet to be determined. In the second part of this manuscript, I include a recently published article on the transcriptional role of the LFY-UFO complex in flower development that allows access to genomic regions distinct from those bound by LFY alone. Moreover, I present some additional results suggesting the implication of LFY and UFO in floral meristem establishment in the early stages of flower development, broadening their importance in this crucial developmental process.

Table of contents

1	Introduction.....	9
1.1	Complexity of gene regulation in living organisms and open challenges	9
1.2	Transcription factors in gene regulation	11
1.2.1	Experimental strategies to map TF-DNA binding genome-wide.....	14
1.2.2	Statistical modeling of TFBSs.....	20
1.2.3	Combinatorial nature of the <i>cis</i> -regulatory code	29
1.2.4	Role of evolutionary conservation in the study of gene regulation.....	31
1.3	Flowering in Arabidopsis: an ideal system to study gene regulation.....	36
1.3.1	Flowering in Arabidopsis and the ABCDE model.....	37
1.3.2	LEAFY, a master floral regulator	39
1.3.3	LFY is a highly conserved TF	41
1.3.4	LFY binding and expression resources in Arabidopsis.....	42
2	Objectives.....	44
3	Chapter 1: A machine-learning model to predict transcriptional regulation of LFY sites genome-wide based on genomic context and evolutionary conservation	45
3.1	Introduction.....	45
3.2	Results	48
3.2.1	Genomic context features better predict functional LFY sites than features based on state-of-the-art LFY-DNA binding models	48
3.2.2	Inclusion of evolutionary conservation does not improve predictions	51
3.2.3	The model reveals important information about how LFY binds the DNA.....	54
3.2.4	The model can be used to make predictions of LFY functional sites on ‘unknown’ sites.....	57
3.3	Discussion.....	59
3.3.1	Genomic context carries key regulatory information	59
3.3.2	Several possible reasons to why conservation of LFY sites does not help predictions.....	61
3.3.3	New potential targets of LFY revealed by the model.....	63
3.4	Supplementary figures	65
4	Chapter 2: The LFY-UFO complex regulates distinct genes from LFY.....	67
4.1	Introduction.....	67
4.1.1	LFY and UFO are involved in petal and stamen development	67
4.1.2	UFO has been proposed to interact with LFY to promote its ubiquitination and degradation	67
4.1.3	LFY and UFO form a transcriptional complex that recognizes new genomic regions.....	68
4.2	Article	70
4.3	Additional results and discussion.....	102

4.3.1	The role of LFY-UFO extends beyond petals and stamens development	102
4.3.2	UFO could have LFY-independent targets.....	104
4.3.3	Double action of UFO as a transcriptional cofactor in the nucleus and an F-box in the cytoplasm	105
5	Conclusions and perspectives	107
5.1	Genomic context can be used to characterize transcriptional regulation of LFY sites and shed new light on LFY's regulatory properties	107
5.2	Could conservation still be useful to model transcriptional regulation?	109
5.3	The action of LFY-UFO in plant development could extend beyond petal and stamen development	110
6	Materials and Methods	113
6.1	ChIP-seq and ampDAP-seq analysis	113
6.2	Microarray analysis	113
6.3	RNA-seq experiments and analyses	114
6.3.1	WT vs <i>lfy</i> RNA-seq experiment.....	114
6.3.2	<i>lfy rev</i> vs <i>rev</i> , <i>ufo rev</i> vs <i>rev</i> RNA-seq experiment.....	114
6.4	Data matrix to classify LFY sites based on genomic context and conservation	115
6.4.1	Definition of LFY sites genome-wide.....	115
6.4.2	Integration of binding and expression data and definition of nonfunctional/functional/'unknown' LFY sites	115
6.4.3	Computing POcc around LFY TFBS	116
6.4.4	Computing co-occurrence and LFY-LFY distances	117
6.4.5	Computing LFY-TSS distances.....	118
6.4.6	Encoding sequence type.....	118
6.4.7	Computing TFBS density and diversity around LFY TFBSs.....	118
6.4.8	Computing average conservation at LFY sites.....	121
6.5	Calculating CNS enrichment at LFY sites	123
6.6	Training and testing Random Forest models	124
6.7	Extracting feature importance from Random Forest models	126
6.8	Using trained Random Forest models to make predictions on 'unknown' sites	126
7	References.....	127

Table of figures

Figure 1.2-1 Factors influencing TF-DNA binding, adapted from (Héberlé & Bardet, 2019)	12
Figure 1.2-2 Schematic representation of experimental techniques to capture genome-wide TF-DNA binding, adapted from (Wang et al., 2023)	15
Figure 1.2-3 Schematic representation of the differences in TF-bound sequences retrieved with ChIP-seq, DAP-seq and ampDAP-seq techniques	19
Figure 1.2-4 How to build a Position Weight Matrix (PWM), from (Wasserman & Sandelin, 2004)	23
Figure 1.2-5 Example of supervised machine learning algorithm to predict whether a given sequence is centered on a transcription start site ('TSS') or not ('Not TSS'), from (Libbrecht & Noble, 2015)	26
Figure 1.3-1 Arabidopsis life cycle, from (Krämer, 2015)	37
Figure 1.3-2 Arabidopsis flower and the ABCDE model, from (Theißen et al., 2016)	39
Figure 1.3-3 Effects of <i>lfy</i> knock-out mutation and overexpression in Arabidopsis, adapted from (Siriwardana & Lamb, 2012)	40
Figure 3.1-1 Graphical abstract for Chapter 1: A machine-learning model to predict transcriptional regulation of LFY sites genome-wide based on genomic context and evolutionary conservation.....	47
Figure 3.2-1 Performance of random forest classifiers built with state-of-the-art LFY TFBS models (PWM with dependencies and POcc) and genomic context information	50
Figure 3.2-2 Including average conservation level of LFY sites does not improve predictions	53
Figure 3.2-3 Most important features included in our Random Forest models. Ten features with the highest Gini importance (highest median value over 100 iterations) in our Random Forest models..	55
Figure 3.2-4 The model can be used to predict whether 'unknown' LFY sites (which could not be confidently labeled as functional nor as nonfunctional) are functional or nonfunctional	58
Figure S3.4-1 Supplementary information about LFY sites genome-wide	65
Figure S3.4-2 Snapshots of LFY sites labeled as 'unknown' and predicted to be functional in Figure 3.2-4A, as indicated by their name at the top of each panel	66
Figure 4.1-1 LFY and UFO form a transcriptional complex to regulate flower developmental genes. Credits: Philippe Rieu.	69
Figure 4.3-1 Investigating the role of LFY and UFO in floral meristem identity establishment	102
Figure 4.3-2 Differentially expressed genes in RNA-seq experiments with <i>rev-c4</i> , <i>rev-c4 lfy-12</i> and <i>rev-c4 ufo-1</i> inflorescences.....	103
Figure 6.4-1 PR (top row) and ROC (bottom row) curves of random forest models trained with only one feature, POcc, computed over a window of ± 250 bp (graphs on the left) or ± 500 bp (graphs on the right) around each LFY site.....	117
Figure 6.4-2 Schematic example of methods to compute non-LFY TFBS density and diversity around LFY sites	120
Figure 6.4-3 AUC for PR (graphs on the left) and ROC (graphs on the right) curves of Random Forest models trained with different features representing TFBS density and diversity around LFY sites ...	121
Figure 6.4-4 PR (top row) and ROC (bottom row) curves of random forest models trained with two conservation features (PhastCons and PhyloP scores) computed with three different methods.....	123
Figure 6.6-1 PR curve to compare balanced vs stratified cross-validation strategies. Green and orange curves obtained when testing stratified and balanced Random Forest models, respectively	125
Figure 6.6-2 Schematic representation of the Repeated Stratified k-fold cross-validation strategy used in all Random Forest models shown in this manuscript	126

List of abbreviations

Abbreviation	Definition
(amp)DAP-seq	(amplified) DNA affinity purification sequencing
AUC	Area Under the Curve
CDS	Coding sequence
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CNS(s)	Conserved Noncoding Sequence(s)
DBD	DNA-binding domain
DEG(s)	Differentially Expressed Gene(s)
IC	Information Content
LFY	LEAFY, plant-specific TF
POcc	Predicted Occupancy
PR curve	Precision-Recall curve
PWM	Position Weight Matrix
ROC	Receiver Operating Characteristic
TF	Transcription Factor
TFBS(s)	Transcription Factor Binding Site(s)
TSS	Transcription Start Site
UFO	UNUSUAL FLORAL ORGANS, F-box protein

1 Introduction

1.1 Complexity of gene regulation in living organisms and open challenges

Spatiotemporal regulation of gene expression is required for correct development and to ensure species survival. While genetic information is the same in all cells, the set of expressed genes can vary depending on developmental stage and environmental cues.

During gene expression, information encoded in the DNA is transcribed to RNA and then translated to proteins (Figure 1.1-1). This is also known as the central dogma of molecular biology, first proposed by Francis Crick in lectures in 1957 and then published the following year (Cobb, 2017; Crick, 1958). Transcription Factors (TFs) are a particular class of proteins that can bind DNA at specific sites and recruit the transcriptional machinery to control the expression of specific genes.

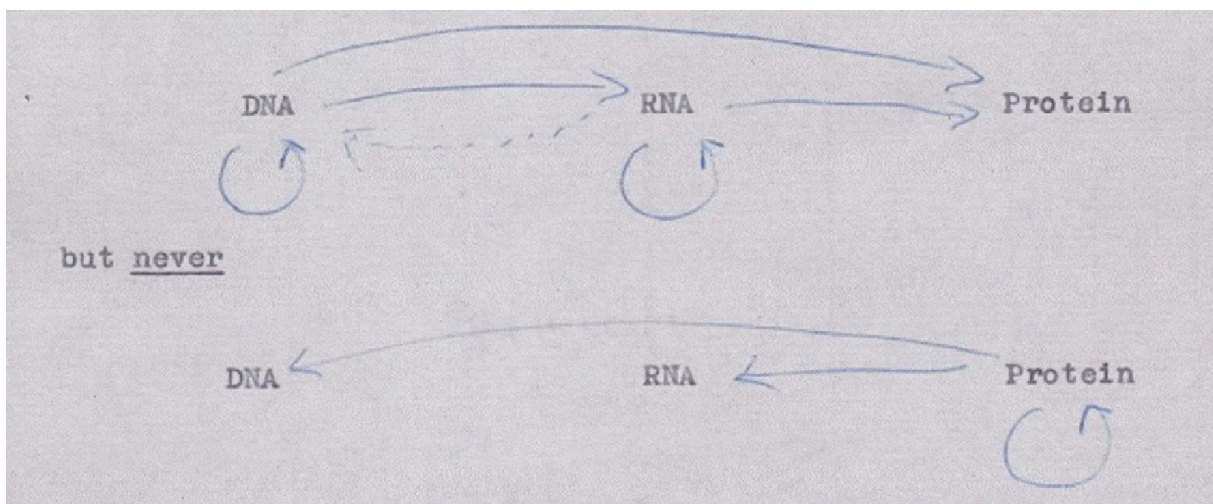


Figure 1.1-1 Francis Crick's first outline of the central dogma of molecular biology, later explained in (Crick, 1958). From (Cobb, 2017).

Knowing when and where genes are expressed is crucial to increase our understanding of the physiological and developmental processes in which they are involved. Recent technological advances have dramatically increased the availability of total mRNA or protein profiles in many species and at increasing resolution, from whole tissues to single cells and even single nuclei (Bennett et al., 2023; Noor et al., 2019; Slyper et al., 2020). These data are valuable to better understand complex biological processes at the systemic level (Greene & Troyanskaya, 2010).

In recent years, powerful computational approaches have leveraged such data to tackle fundamental questions in gene regulation biology, reaching remarkable accuracy in previously unattainable tasks such as the prediction of gene expression from DNA sequence (Avsec, Agarwal, et al., 2021). However, building such models requires extensive data, which are hardly available outside of model organisms or simplified systems (e.g. tissue culture). Moreover, they often trade performance and accuracy for interpretability and robustness on different systems (Meyer & Saez-Rodriguez, 2021).

As the general rules of gene expression presented in Figure 1.1-1 are common to all living organisms, leveraging data from multiple species can also offer new insights into fundamental questions about gene regulation in living systems. Multispecies comparisons have been instrumental in the study of noncoding regulatory regions, as highly conserved sequences have been shown to be functionally important (Berthelot et al., 2018; Lindblad-Toh et al., 2011; Siepel et al., 2005; Wittkopp & Kalay, 2012). Despite multiple disciplines coming together to decipher the complexity of gene regulation, our understanding of living systems at higher scales remains fragmented and is far from complete.

This manuscript mainly focuses on the early steps of gene regulation, namely the binding of TFs to DNA to regulate gene expression. In the rest of the introduction, I will first introduce the role of TFs in gene regulation in general, the factors influencing their binding and the main techniques to study their binding genome-wide, as well as the statistical models that can be used to study their binding specificity. As TFs can bind DNA with the help of other cofactors, I will go on to describe how different TF combinations can influence gene expression. Then, I will discuss how evolutionary information can be important to study the regulatory function of genomic sequences. Finally, I will focus on the model plant *Arabidopsis thaliana* (hereafter Arabidopsis) and on flowering as a system to study gene regulation, and I will conclude by focusing on one TF in particular, LEAFY (LFY), a master TF with a central role in flowering, as it will be at the core of chapter one and two.

1.2 Transcription factors in gene regulation

Transcription factors (TFs) are proteins that can bind DNA at specific positions to regulate gene expression, either positively or negatively. They constitute about 5% of protein-coding genes in *Arabidopsis* and around 8% in the human genome (Lambert et al., 2018; Riechmann et al., 2000).

TFs fulfill their regulatory function through the physical interaction of their DNA-binding domain (DBD) with short DNA motifs called TF binding sites (TFBSs). The recognition of such sequences on the genome is influenced by several factors that depend either on the nature of the DNA sequence itself, or on the presence of other proteins (Figure 1.2-1).

DNA sequence is an important determinant of TF-DNA binding, and it is directly linked to the three-dimensional (3D) structure of the DBD (Figure 1.2-1A). TFs with similar DBDs tend to recognize similar motifs, and such similarities have led to the use of DBD structure as a basis for TFs classification in mammals and plants (Blanc-Mathieu et al., 2023; Wingender et al., 2015).

Regulatory regions are enriched in TFBSs bound by different TFs, and their arrangement and nature can also influence gene expression (Figure 1.2-1B) (Spitz & Furlong, 2012). This is the so-called combinatorial nature of the *cis*-regulatory code, whereby combinations of different TFs co-occur on related target regions at preferential distances from each other, establishing additional regulatory interactions. Such interactions can lead to cooperativity (i.e. binding of one TF increases the binding affinity of another TF) or competition (i.e. binding of a TF prevents binding of a second one) between TFs, which influence DNA binding and downstream gene expression (Figure 1.2-1B). This topic will be further developed in another section of the introduction: Combinatorial nature of the *cis-regulatory* code, p. 29.

DNA shape has also been shown to influence TF-DNA binding (Figure 1.2-1C) (Rohs et al., 2009). TFs can recognize specific DNA secondary structures (Spiegel et al., 2021), or their DBD can specifically allow binding on the minor or major groove of the DNA. For instance, the human AT-hook factor HMGA1, like other AT-hook TFs, binds to the minor groove (Fonfría-Subirós et al., 2012), while the *Arabidopsis* AP2 family TEMPRANILLO 1 TF binds the major groove (Hu et al., 2021). LEAFY (LFY), a plant-specific TF, is an example of a transcription factor (TF) that can establish contacts with both the major and minor grooves

of DNA. Its helix-turn-helix-like DBD engages the major groove through its alpha helices and the minor groove through its N-terminal loop (Hamès et al., 2008). Such DNA shape constraints contribute to the binding specificity of TFs.

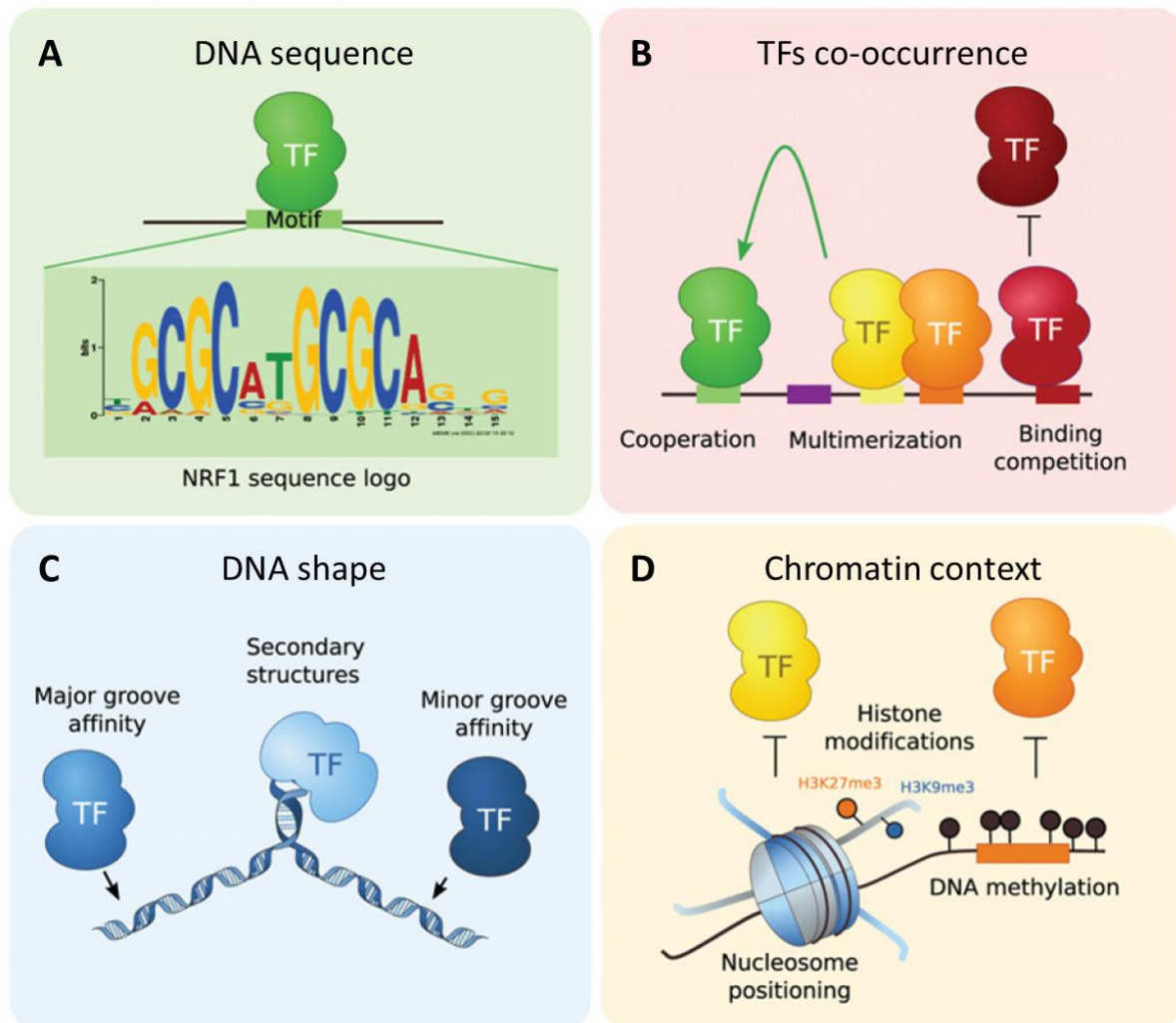


Figure 1.2-1 Factors influencing TF-DNA binding, adapted from (Héberlé & Bardet, 2019). A: Recognition of a specific DNA sequence (Motif) by a TF, where the sequence logo of NRF1 (nuclear respiratory factor 1) represents its motif binding preference, is shown. B: Co-occurrence of multiple TFs on the same regulatory region, and how they influence each other's binding through multimerization, i.e. the formation of multi-protein complexes, cooperative binding or competition. C: TFs can have DNA shape preferences, like an affinity for the major or minor groove of double-stranded DNA or binding to secondary DNA structures. D: Chromatin context influences TF binding through nucleosome positioning, specific histone modifications (with activating, such as H3K9me3, or repressive functions, like H3K27me3) and DNA methylation (i.e. the presence of methylated cytosines at specific genomic regions).

The ability of a TF to recognize and bind to its target regions is not only influenced by the presence of its binding partners, DNA shape or the formation of oligomeric complexes – DNA modifications and accessibility also impact TF binding (Figure 1.2-1D) (Spitz & Furlong, 2012).

Both these processes are highly dynamic and undergo major rearrangements throughout development (Lloyd & Lister, 2021). DNA methylation is the chemical modification of a cytosine by the addition of a methyl group to the fifth carbon atom, forming a 5-methylcytosine, and it has been shown to influence, either positively or negatively, the binding of the majority of human and plant TFs (O'Malley et al., 2016; Yin et al., 2017). While in vertebrates this modification happens in a CpG context, in plants it also occurs in CHG or CHH contexts (Law & Jacobsen, 2010). DNA accessibility depends on a combination of nucleosome occupancy and specific histone modifications. While most TFs preferentially bind to nucleosome-depleted regions (F. Zhu et al., 2018), chromatin marks can also differentiate active enhancers (i.e. regulatory regions located far from the genes that they regulate) from inactive ones. Histone H3 lysine 27 acetylation (H3K27ac), and more recently histone H2B N-terminus multisite lysine acetylation, have been shown to mark active enhancers in human and mouse models, while H3K27ac does not seem indicative of active enhancers in *Arabidopsis* (Creyghton et al., 2010; Narita et al., 2023; Yan et al., 2019).

There is a category of TFs, called pioneer TFs, that can recognize their DNA target sequence despite nucleosome occupancy, and trigger chromatin opening and nucleosome displacement to further facilitate the binding of other TFs. In mammals, examples of pioneer TFs include OCT4, SOX2 and SOX11, which bind exposed nucleosomal DNA and induce a distortion of the DNA structure that widens the minor groove (Kagawa & Kurumizaka, 2021; Soufi et al., 2012). In plants, LFY, APETALA1 (AP1) and SEPALLATA3 (SEP3) have been shown to have pioneer activities (Jin et al., 2021; Lai, Blanc-Mathieu, et al., 2021; Pajoro et al., 2014).

As the regulatory action of TFs is fulfilled upon DNA binding, it is essential to characterize how and where they are bound on the genome. Numerous techniques were developed over the years to detect and probe TF-bound regions and TF binding specificity, initially relying on binding assays on regulatory regions based on prior genetic insights. As high-throughput sequencing became more widely available and affordable, techniques to identify TF binding genome-wide quickly became state-of-the-art approaches to study TF-dependent gene regulation. Compared to techniques such as electrophoretic mobility shift assay (EMSA), which tests protein binding *in vitro* to a DNA probe, genome-wide techniques give a wider picture of the action of TFs in gene regulation. Techniques for genome-wide binding profiling

can provide information about TF specificity and genomic targets at once, and they have contributed to the characterization of binding specificity for thousands of TFs (Castro-Mondragon et al., 2022; Matys et al., 2006). I will present some of these key techniques, their strengths and their weaknesses, in the next section.

1.2.1 Experimental strategies to map TF-DNA binding genome-wide

Knowing where a TF binds on the genome is an important step to find out how it controls gene expression. This section will present in more detail some of the most widely used techniques to capture TF-DNA binding genome-wide *in vivo* and *in vitro* (Figure 1.2-2). Since binding profiling *in vivo* (i.e. in the native chromatin context of the probed tissue at a given developmental stage, and in the presence of potential cofactors) can yield substantially different results from what is observed *in vitro* (i.e. with purified proteins and naked DNA), I deemed it necessary to introduce such techniques in different sub-sections.

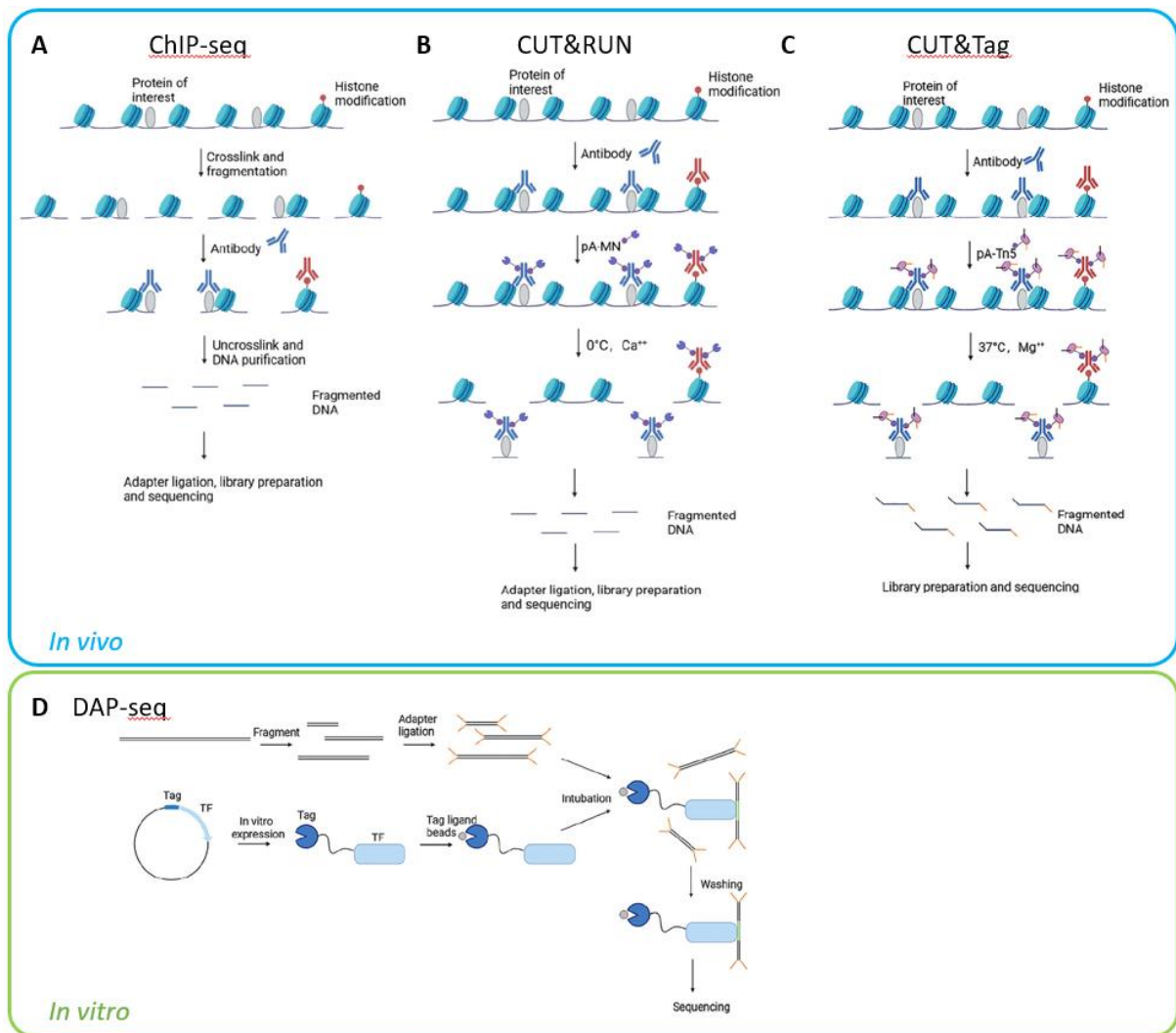


Figure 1.2-2 Schematic representation of experimental techniques to capture genome-wide TF-DNA binding, adapted from (Wang et al., 2023). Blue box: techniques for genome-binding profiling *in vivo*; Green box: binding profiling *in vitro*. A: in ChIP-seq, chromatin and DNA are crosslinked and fragmented, followed by immunoprecipitation of the protein of interest with specific antibodies. Then, crosslinking is disrupted to separate chromatin and DNA, and the latter is purified to obtain the genomic regions bound by the protein of interest, which will eventually be sequenced. B: In CUT&RUN crosslinking is not required, and cells are first incubated with an antibody for the protein of interest and with pA-MN. Upon the addition of Ca⁺⁺, the MNase cleaves DNA on the sides of the protein of interest, and fragmented DNA is prepared for sequencing. C: In CUT&Tag, an antibody targets the protein of interest on non-crosslinked chromatin, followed by the addition of pA-Tn5 enzyme. Upon the addition of Mg⁺⁺, DNA cleavage and adapter ligation are performed simultaneously on the sides of the protein of interest, and samples are sent to sequencing. D: In DAP-seq, fragmented DNA is ligated to sequencing adapters, while the TF of interest, fused to a tag, is expressed *in vitro* and purified through binding to ligand-coupled beads. Then, DNA and TF-tag are incubated, unbound DNA fragments are washed away and DNA fragments bound by the TF of interest are sequenced.

1.2.1.1 *In vivo* genome-wide TF-DNA binding assays

1.2.1.1.1 Chromatin immunoprecipitation assay followed by sequencing (ChIP-seq)

To this day, ChIP-seq is one of the most widely used techniques to study genome-wide protein occupancy (Figure 1.2-2A). This technique was first developed in 2007 for mammalian cells, and has since become increasingly used in plants as well (Robertson et al., 2007). ChIP-seq involves cross-linking of DNA-bound proteins and sonication, followed by the use of protein-specific antibodies to purify DNA through immuno-precipitation. This allows the isolation of DNA fragments specifically bound by the protein of interest, for instance a TF, *in vivo*. Sequencing of these fragments produces reads that one can map on a reference genome, highlighting enriched DNA regions, called “peaks”, which contain the sequences bound by the protein. In the case of TF ChIP-seq experiments, the peak maximum, i.e. the region where the maximum amount of immuno-precipitated sequenced reads align, should correspond to the DNA region where the protein is bound *in vivo* (Johnson et al., 2007; Robertson et al., 2007).

A high-quality control experiment is important to exclude biases associated with different cross-linking, sonication and immunoprecipitation propensity in different genomic regions. The use of replicate samples increases the reliability of the obtained protein-DNA binding landscape.

Over 15 years of intensive application of this technique have shown its inherent biases, as well as possible solutions. Namely, it was shown that some genomic regions are bound in ChIP-seq experiments across different and supposedly unrelated experiments, which has prompted the development of so-called blacklisted regions that are commonly removed in standard ChIP-seq analysis to enrich for protein-specific DNA-binding signal (Amemiya et al., 2019; D. Park et al., 2013; Teytelman et al., 2013). Such bias has been observed in plants as well, leading to the recent construction of an Arabidopsis-specific blacklist (Klasfeld et al., 2022).

While ChIP-seq is still the experiment of choice for *in vivo* TF-DNA and chromatin mark profiling, some inherent limitations hinder its application in specific circumstances (P. J. Park, 2009). First, as the technique is based on immuno-precipitation, antibody quality is crucial for its success. In the absence of a specific antibody for the protein of interest, a tagged

version can be used, but it is essential to run thorough controls to ensure it retains its native DNA-binding capacities. Second, the amount of starting material available is fundamental to obtain a high-quality ChIP-seq, which means that tissue-specific experiments on restricted cell populations can prove challenging and outcomes can vary significantly. Third, due to the relatively large size of the DNA fragments produced during sonication, ChIP-seq has a rather low resolution: the immuno-precipitated protein is bound within the ~200-500 bp of the recovered peak, supposedly with an enrichment at the peak maximum, but detailed TFBS determination and identification within the bound region requires additional motif-search analysis. Finally, as in every technique relying on high-throughput sequencing, sequencing depth is just as important as the technical caveats mentioned above.

Over the years, a plethora of techniques to identify genome-wide protein-DNA binding has been developed to improve upon some of the limitations of ChIP-seq. ChIP-exo (from ChIP-exonuclease) and ChIP-nexus (for ChIP experiments with nucleotide resolution through exonuclease, unique barcode and single ligation) were developed to obtain nucleotide-level resolution of *in vivo* protein binding. In ChIP-exo, immuno-precipitated DNA is treated with a 5' to 3' exonuclease, which digests accessible DNA flanking the protein of interest. As a result, only a short stretch of DNA, made inaccessible by the presence of the bound protein, is left for sequencing, resulting in nucleotide-resolution mapping of genome-wide protein occupancy (Rhee & Pugh, 2011). A subsequent improvement of ChIP-exo, ChIP-nexus was developed to improve the efficiency of library preparation, and with the additional advantage of retaining strand information (He et al., 2015). ChIP-exo and ChIP-nexus data have been extensively used for deep learning-based modeling of TF binding for human TFs (Avsec, Weilert, et al., 2021; Y. Zhang et al., 2021), but they have not been implemented in plants. A more detailed overview of genome-wide binding techniques can be found in (Hajheidari & Huang, 2022; Lai, Stigliani, et al., 2019).

1.2.1.1.2 CUT&RUN and CUT&Tag

More recently, Cleavage Under Targets and Release Using Nuclease (CUT&RUN) and the closely related Cleavage Under Targets and Tagmentation (CUT&Tag) were developed to overcome the limitations of ChIP-seq in terms of required starting material and DNA-binding resolution for *in vivo* protein-DNA binding (Kaya-Okur et al., 2019; Skene & Henikoff, 2017).

CUT&RUN relies on the initial binding of a protein-specific antibody followed by the tethering of a Protein A/Micrococcal Nuclease (pA-MNase) to the antibody (Figure 1.2-2B). Addition of Ca^{2+} activates the pA-MNase, which cleaves accessible DNA around the protein of interest, and the protein-bound sequences in the supernatant will be used for library preparation and sequencing. In CUT&Tag, a protein-specific antibody first targets the protein of interest, and a secondary antibody binds to the primary one (Figure 1.2-2C). Then, a hyperactive Tn5 transposase-protein A (pA-Tn5) fusion protein, previously loaded with adapter sequences, binds to the antibodies and is activated through the addition of Mg^{2+} . The addition of Mg^{2+} results in the immediate tagmentation (i.e. DNA fragmentation and inclusion of adapter sequences in a single step) of the accessible sites around the protein of interest and the production of library-level sequences ready for amplification and sequencing. However, it was reported that methods based on Tn5 transposase such as CUT&Tag are biased towards open chromatin regions (Wang & Zhang, 2021). While the impact of this bias seems rather limited in CUT&Tag, at least compared to other Tn5-based techniques, it should be taken into account.

1.2.1.2 *In vitro* genome-wide TF-DNA binding assays

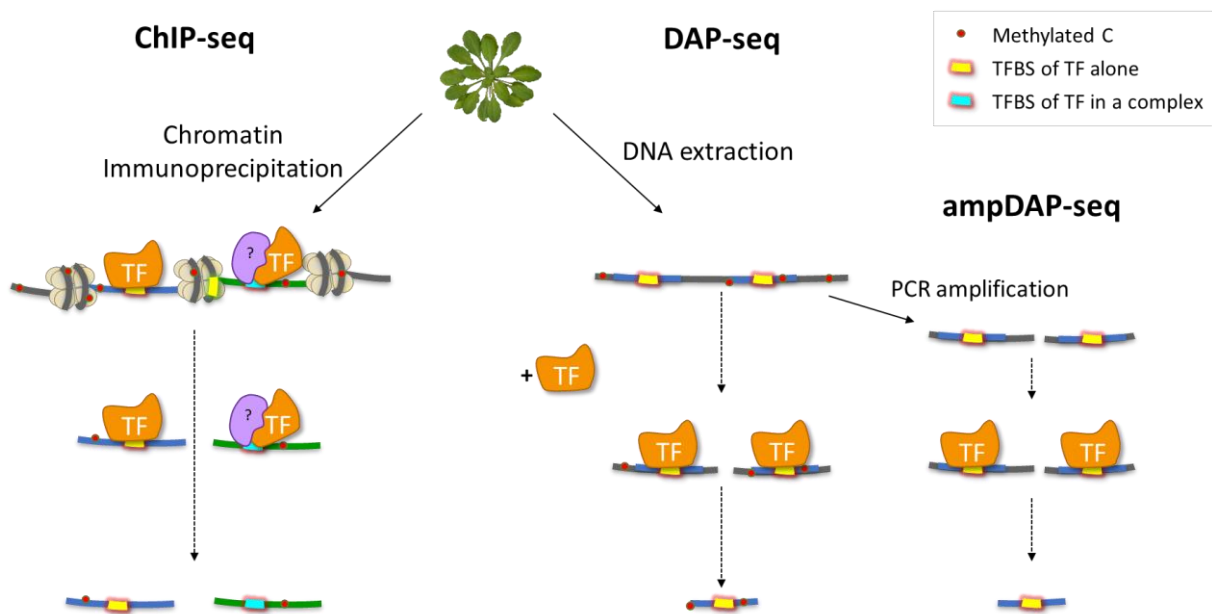
1.2.1.2.1 DNA affinity purification followed by sequencing (DAP-seq)

In addition to *in vivo* systems, some techniques have focused on finding sequences bound by a protein of interest *in vitro*. DAP-seq and its variations ampDAP-seq and seq-DAP-seq, provide fast and efficient ways to obtain direct, genome-wide binding of a protein (or, with some adjustments, protein complex) of interest *in vitro*.

DAP-seq was obtained through the adaptation of an existing microarray-based technique to high-throughput sequencing (Figure 1.2-2D) (O'Malley et al., 2016; Rajeev et al., 2014). Like ChIP-seq, DAP-seq aims at finding DNA sequences bound by a (tagged) protein of interest, but in this case on genomic DNA devoid of its chromatin context. ampDAP-seq is proposed as a DAP-seq variation that includes an additional step of PCR amplification which gets rid of DNA methylation (O'Malley et al., 2016). The publication of the DAP-seq method by O'Malley et al. came with public datasets of genomic regions bound *in vitro* by over 500

Arabidopsis TFs, which represented a significant contribution to the advancement of the plant regulatory landscape (O'Malley et al., 2016).

While ampDAP-seq represents the ensemble of all potential target sites of a TF on the genome *in vitro*, ChIP-seq contains additional information about which targets are bound in a certain tissue and at a given developmental stage *in vivo* (Figure 1.2-3). As a result, not all ChIP-seq-bound regions are also bound in ampDAP-seq, and vice versa, due to differences in the presence/absence of cofactors and overall chromatin accessibility (O'Malley et al., 2016). Moreover, comparing a TF's binding profile *in vivo* and *in vitro* or its genomic targets *in vitro* with (DAP-seq) or without (ampDAP-seq) DNA methylation can provide important insights about its binding preferences, if and where they require an intact chromatin context and its sensitivity to DNA methylation (Figure 1.2-3) (O'Malley et al., 2016; Lai, Blanc-Mathieu et al. 2021).



*Figure 1.2-3 Schematic representation of the differences in TF-bound sequences retrieved with ChIP-seq, DAP-seq and ampDAP-seq techniques. In ChIP-seq (left), immunoprecipitation of the protein of interest (TF) in its chromatin context indistinctly retrieves genomic sequences bound by the protein alone or in a bigger complex. In DAP-seq (middle), DNA is devoid of chromatin but it retains the DNA methylation marks (red dots) of the developmental stage and tissue that the DNA was extracted from. In ampDAP-seq (right), an additional step of PCR amplification gets rid of DNA methylation marks and allows the target protein to bind to its target sequences. In both DAP-seq and ampDAP-seq, the experimental setting *in vitro* and the use of purified protein ensures that the retrieved sequences are bound by the protein of interest by itself.*

1.2.2 Statistical modeling of TFBSs

Whether it is through step-by-step TF-DNA binding assays or at the genome-wide level, knowing a TF's binding specificity is crucial to pinpoint its TFBSs on regulatory regions, and several strategies have been developed to this end.

1.2.2.1 Consensus sequences

The most immediate way to represent a TFBS is through a consensus sequence, i.e. a short sequence of nucleotides typically recognized by a TF, which usually allows little to no nucleotide variation at each position (Boeva, 2016).

The main limitation of consensus-based strategies is that TF-DNA binding can tolerate a certain degree of variation at given positions. This is not accounted for when using consensus sequences to define the presence of a TFBS. An example to illustrate this limitation is the binding of the Arabidopsis LEAFY (LFY) master TF, that has been shown to bind DNA motifs with up to three mismatches *in vitro* but that does not always bind consensus-matching sequences (Moyroud et al., 2011; Winter et al., 2011). Moreover, as consensus sequences are typically short (≤ 10 bp), their detection beyond specific (and experimentally tested) sets of *cis*-regulatory sequences is limited, as they can occur randomly on the genome. Therefore, some predicted sites are not actually recognized, while experimentally-bound, high-affinity sites that do not strictly display the canonical consensus sequence are not detected (Stigliani et al., 2019; Winter et al., 2011).

1.2.2.2 Position Weight Matrices (PWMs)

Position Weight Matrices (PWMs) are among the most widely adopted and flexible ways to model TF specificity. A PWM is typically 4-8 nucleotides long and it contains, at every position of the TFBS, a so-called 'score' for each nucleotide (A, C, T, G) which represents how beneficial or detrimental it is for TF binding, based on its position in the TFBS (Figure 1.2-4C) (Stormo, 2015).

PWMs can be computed from the alignment of TF-bound regions identified from genome-wide DNA-binding experiments such as ChIP-seq or DAP-seq (Figure 1.2-4A). By counting the occurrences of nucleotides A, C, T, and G at each position in the alignment, a position

frequency matrix can be constructed (Figure 1.2-4B). A position frequency matrix can be converted to a PWM (Figure 1.2-4C) based on the observed frequency of each nucleotide at each position relative to the expected genomic frequency of A, C, T and G. Any DNA sequence can then be assessed for its TFBS potential with a PWM by retrieving the score of each nucleotide at its corresponding position in the sequence (Figure 1.2-4D). Depending on the chosen implementation, the best possible TFBS has either a score of zero or a positive one; the more negative the sequence score, the worse the match between TF and DNA (Lai, Stigliani, et al., 2019; Stormo, 2015; Wasserman & Sandelin, 2004). TF binding specificity can be visually represented as a sequence logo (Figure 1.2-4E), where letter size represents the importance of each base for binding, i.e. the so-called Information Content (IC), expressed in bits (Schneider & Stephens, 1990). Maximum IC equals 2 bits, which represents a nucleotide that is always found at a certain position in all aligned sequences. The more variable the position, the lower its overall IC, so that, in a sequence logo, bigger letters imply a strict nucleotide requirement, while smaller letters reflect higher tolerance to nucleotide variability.

Several databases collect PWMs computed from published ChIP-seq, DAP-seq and protein binding microarray data, in an ever-growing number of organisms (Matys et al., 2006; Weirauch et al., 2014). Among them, JASPAR, created in 2004 and since then regularly updated, provides a manually curated collection of PWMs that is widely used by the scientific community (Sandelin, 2004). As TFs belonging to the same family share a similar binding motif, the latest version of JASPAR also features binding archetypes where redundancy was removed to represent the binding specificity of TF groups (Castro-Mondragon et al., 2022).

One of the drawbacks of PWMs is that they quantify the binding affinity of a TF for a given DNA sequence, which ranges from sequences with a perfect match and high affinity to no affinity at all, passing through many sequences with suboptimal affinity but that are still bound in experimental settings. However, many applications require a threshold to identify which potential TFBS are considered as true TFBS, i.e. what score is good enough for the sequence to be reasonably bound. A strategy to overcome this is to use genome-wide distributions to set a percentile-based threshold, or to associate a p-value to each site (Ambrosini et al., 2018; Boeva, 2016). Moreover, PWMs are based on the assumption that

each nucleotide position contributes independently to protein binding, which may not always be the case (Benos et al., 2002; Jolma et al., 2013). Therefore, more complex models have been developed over the years to calculate and include position interdependencies within matrices, to detect changes in DNA shape or to combine PWM score with experimentally-determined biophysical properties of the TF itself (Mathelier et al., 2016; Mathelier & Wasserman, 2013; Moyroud et al., 2011; Roeder et al., 2007; Workman et al., 2005).

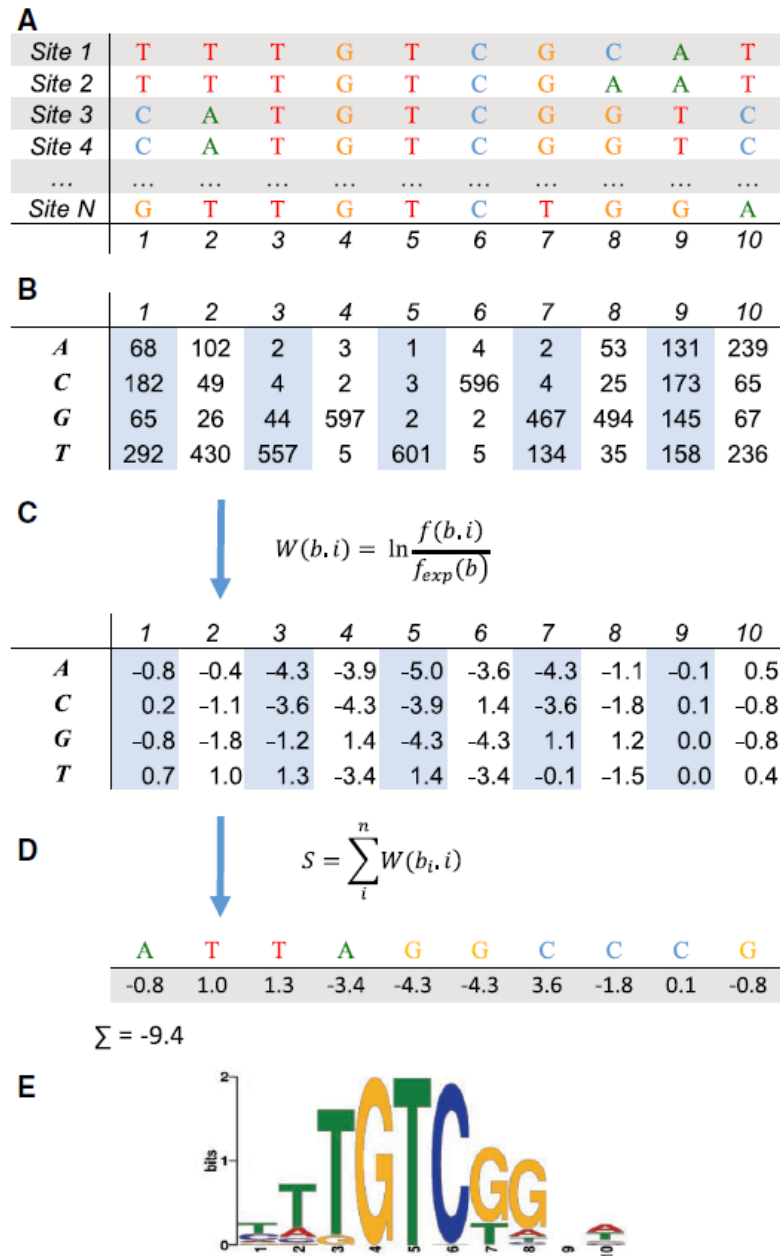


Figure 1.2-4 How to build a Position Weight Matrix (PWM), from (Wasserman & Sandelin, 2004). A: DNA sequences bound by the protein of interest are aligned. B: A Position Frequency Matrix (PFM) is computed, containing the number of sequences where each nucleotide (A, C, G, T) is found at each position of the binding site. C: The PFM is converted to a PWM, where each position represents the influence of each nucleotide on binding at each position, based on the expected frequency of each nucleotide on the genome. Negative values indicate a negative effect of the nucleotide to the binding of the protein to a DNA sequence. $W(b,i)$ = weight (PWM value) of base b in position i ; $f(b,i)$ = frequency of nucleotide b in position i ; $f_{exp}(b)$ = background probability of nucleotide b . D: The PWM can be used to probe the affinity of the TF to a new DNA sequence by summing the weight of each nt in the DNA sequence to its corresponding weight in the PWM. E: A PWM can be represented as a sequence logo, where taller letters indicate nucleotides that are important for binding at specific positions.

1.2.2.3 Predicted occupancy (POcc)

The PWM score quantifies the affinity of a given TF for a DNA sequence of the same length as the PWM itself. Therefore, the presence of multiple TFBSs on the same regulatory region, and the TF's properties of association and dissociation from DNA are not taken into account. A more sophisticated method was proposed to quantify the predicted occupancy (POcc) of a DNA sequence by a TF of interest, based on the TF's biophysical properties in addition to its binding motif. This method was developed in yeast and subsequently adapted to the LFY TF in Arabidopsis (Moyroud et al., 2011; Roeder et al., 2007).

POcc is calculated as shown in Equation 1 and Equation 2: it is the sum of a TF's predicted occupancy at all sites detected by a PWM of length W on a sequence of length L , given the TF's experimentally-determined equilibrium association constant at each site ($K_{A,s}$) and its concentration ($[TF]$). $K_{A,s}$ is the inverse of the dissociation constant $K_{D,s}$, which depends on the relationship between binding affinity quantified through the PWM score and the TF's binding affinity (Equation 2), as measured experimentally through quantitative multifluorescence relative affinity (QuMFRA) assay (Moyroud et al., 2011). For example, LFY's association constant was established through QuMFRA assay by quantifying the affinity between LFY's DBD and 48 oligonucleotides with known PWM score (Moyroud et al., 2011).

Put in simpler terms, POcc quantifies the occupancy of a TF on a DNA sequence by taking into account all the TF's PWM-detected sites within that sequence, and the DNA association and dissociation properties of the TF itself. Crucially, this means that, unlike the PWM score, POcc can be calculated for DNA sequences of any size, provided that their length is at least greater than the length of their binding motif.

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

Equation 1 Predicted occupancy (POcc) computation, from (Moyroud et al., 2011). L = length of the tested sequence; W = PWM length; $K_{A,s}$ = relative equilibrium association (A) constant for sites (s), determined experimentally as $K_{A,s} = 1/K_{D,s}$; $[TF]$ = TF concentration.

$$\text{score}_s = -\ln(K_{D,s}) a + b \rightarrow K_{D,s} = e^{\frac{(b - \text{score}_s)}{a}}$$

Equation 2 Calculation of $K_{D,s}$, dissociation constant, based on site's PWM score and experimental data. $K_{D,s}$ is the inverse of the association constant, $K_{A,s}$, shown in Equation 1.

One of the limitations of POcc is that, as it relies on PWM score to define binding sites, it requires setting a PWM score threshold, as mentioned for PWMs before. Moreover, POcc is sensitive to input sequence length: longer sequences will be more likely to contain a higher amount of TFBSs, which will all contribute to the final POcc value. This becomes particularly important if one wants to compare POcc values on regulatory regions of different length, or in different species (Minguet et al., 2015).

Nevertheless, POcc was instrumental to investigate the conservation of two important targets of LFY, *AP1* and *AGAMOUS* (*AG*), during plant evolution through its application to the regulatory regions of multiple plant species (Minguet et al., 2015; Moyroud et al., 2011). Moreover, POcc accounts for LFY's cooperative binding, as it calculates a global value of occupancy that takes into account the presence of multiple TFBSs within the given sequence (Sayou et al., 2016).

1.2.2.4 Machine-learning algorithms to model TF binding

Machine-learning algorithms can learn patterns from vast amounts of data and make new predictions based on those patterns. This can be particularly useful when handling sequencing data about thousands of genomic regions with different characteristics where several processes are at play. I would like to note that some of the methods described in the previous sections, e.g. computing a PWM from a set of bound sequences, also represent a form of machine learning, but in this new section I will focus on algorithms that are more suitable for more complex tasks.

In so-called supervised machine learning algorithms, data are assigned a label so that the model can learn patterns that help it distinguish between different categories, and make predictions on new sets of unlabeled data (Libbrecht & Noble, 2015; van Dijk et al., 2021). Figure 1.2-5 shows an example of how a supervised algorithm can be trained on a set of sequences, which are labeled as either 'TSS' or 'Not TSS' and are associated with a series of features that describe them. The trained model can be tested on a new series of sequences

for which the label is not given and, based on the patterns learned during training on labeled sequences, it can now predict whether the new, unlabeled sequences belong to the 'TSS' or 'Not TSS' category (Libbrecht & Noble, 2015).

The advantage of supervised models is that it is possible to estimate their prediction capacities through cross-validation, a procedure whereby an algorithm is trained on a subset of the labeled data and is tested on the remaining part of the labeled dataset. This provides a way to evaluate whether the model can correctly separate input classes, before making predictions on new sequences.

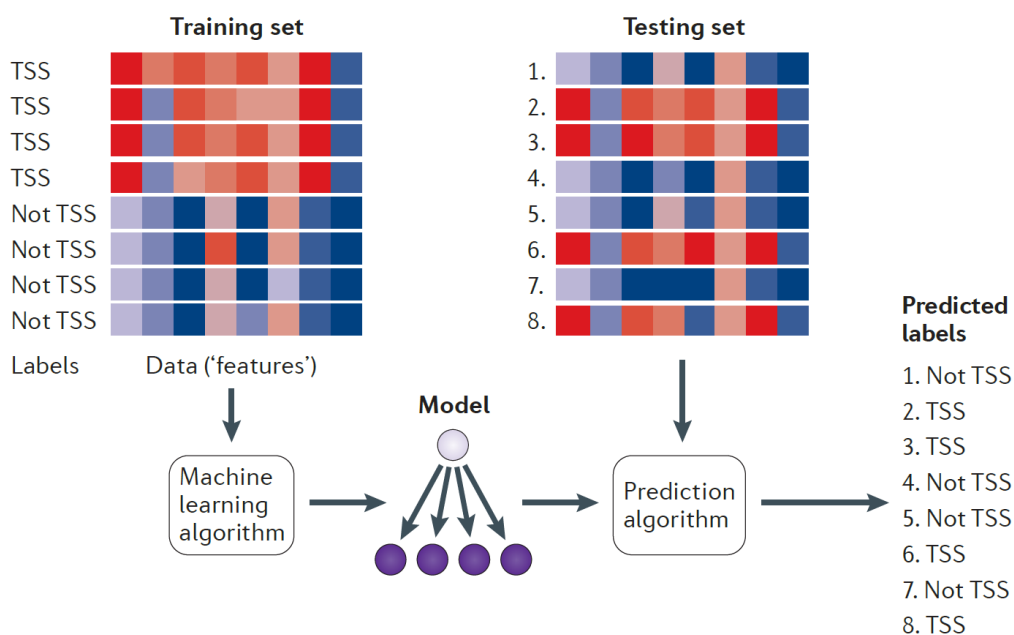


Figure 1.2-5 Example of supervised machine learning algorithm to predict whether a given sequence is centered on a transcription start site ('TSS') or not ('Not TSS'), from (Libbrecht & Noble, 2015). The algorithm is given a 'Training test' from which it learns to distinguish labels ('TSS' or 'Not TSS') based on the information given by the features in each column (represented here as a red-blue gradient). When presented with a new series of sequences without labels ('Testing set'), it can predict them based on the previously learned information.

Another important distinction is between generative and discriminative models. The former will learn from the input data and generate new entries based on the learned pattern, while a discriminative model will learn how to divide input data into different classes (Libbrecht & Noble, 2015).

As I mentioned before, computing a PWM from a set of bound sequences also represents a form of machine learning. However, PWMs only allow to predict binding affinity of a TF for a DNA sequence of the same length as the matrix itself. This means that they do not take into

account the broader genomic context in the bound sequences, which typically span several hundreds of bp. The availability of an ever-increasing amount of data has allowed the application of more complex machine learning algorithms that can take advantage of more features and generally a wider genomic context. Such models have been shown to outperform PWM models in predicting TFBS in mammals and plants (Avsec, Weilert, et al., 2021; Liu et al., 2021; Shen et al., 2021).

An example of a supervised model is DeepBind, which aims at predicting the binding specificity of DNA- and RNA-binding proteins from different types of binding data (Alipanahi et al., 2015). DeepBind correctly recognizes TF motifs from larger (~100 bp) sequences and it outperforms PWM-based methods, and it can be used to predict the effects of single mutations on binding affinity.

DeepBind was initially developed on human and mouse data, but the same approach was used to build trans-species prediction of TFBS (TSPTFBS), which models TF binding in other species from Arabidopsis DAP-seq data (Liu et al., 2021). Like DeepBind, TSPTFBS reaches a higher performance than MEME, a commonly used tool to compute PWMs from input sequences. While DeepBind performed well on both *in vitro* and *in vivo* data, on both human and mouse sequences (Alipanahi et al., 2015), TSPTFBS gave underwhelming results on ChIP-seq data from three other plant species, performing well only on a subset of rice TFs and not on maize or soybean data. This result may seem unsurprising, as the evolutionary scales between the plant species are wider than those between human and mouse. However, it highlights the fact that, despite the promising performances obtained by these algorithms, training them on model species in the hope of transferring the knowledge on other species where less data are available (“transfer learning”) is still challenging.

Generative models have been developed to predict binding signals at nucleotide resolution, like in the case of BPNet, which was trained with ChIP-nexus data (see page for more details about ChIP-nexus) (Avsec, Weilert, et al., 2021). BPNet can quantify strand-specific signal of TF binding on any DNA sequence, and predictions mirror experimentally determined occupancy with remarkable accuracy. Moreover, this model could recover TF binding syntax (i.e. TF-TF distance and orientation preferences) for multiple TFs simultaneously (Avsec, Weilert, et al., 2021).

Generative models have also been used to predict gene expression directly from DNA sequences, by indirectly learning binding syntax in regulatory regions. An example is Enformer, which is currently the state-of-the-art for deep learning-based generative models constructed for this task in humans (Avsec, Agarwal, et al., 2021). The key feature of Enformer is that it supposedly integrates long-range context information through the use of exceptionally long DNA sequences (up to 100 kb around the TSS), allowing it to account for distal regulatory regions and binding syntax. The Enformer model is a so-called “black-box” that does not allow direct investigation of the information used for predictions, and long-range interactions were thought to contribute significantly to its high performance. However, it was recently reported that a sequence input window of 39 kb (20% of the ~200 kb full sequence) was already sufficient to account for most of the binding signal (Karollus et al., 2023).

Deep-learning strategies to predict DNA accessibility or gene expression output from sequence were also developed in plants. A first example is PlantDeepSEA, which leveraged DNA accessibility data from six plant species and can be used to (i) predict the effects of sequence variants on chromatin accessibility as well as to (ii) reveal important cis-regulatory elements on DNA sequences (Zhao et al., 2021). More recently, Akagi et al. combined public (amp)DAP-seq data for TF binding in *Arabidopsis* (O’Malley et al., 2016) and RNA-seq data at multiple tomato ripening stages to build another model that can correctly predict gene expression at specific developmental stages in tomato (Akagi et al., 2022). Once trained, the model can also be used to design synthetic promoters to drive gene expression in tomato (Akagi et al., 2022). These examples show that approaches to predict DNA accessibility and gene expression from sequence are not only available for mammalian systems but are becoming more and more widespread in plants as well.

1.2.3 Combinatorial nature of the *cis*-regulatory code

TFs rarely act alone and, in most cases, several TFs bind together to common regulatory regions. From a mechanistic and structural point of view, TFs can work together through direct protein-protein interactions, but also in the absence of direct protein-protein contacts (Morgunova & Taipale, 2017). The latter can be explained by changes in DNA structure induced by the binding of one protein, or by nucleosome depletion as a result of the binding of one TF. Either way, TFs can facilitate the binding of additional proteins with or without direct protein-protein interactions.

Protein-protein interactions can be involved in cooperative binding of multiple copies of the same TF, as TFs often bind the DNA as dimers or tetramers, and sometimes through higher-order oligomeric structures (Amoutzias et al., 2008). The interactions stabilizing such oligomeric structures can occur at the level of the DBD, as it happens for basic helix-turn-helix (bHLH) dimers (Brownlie et al., 1997), or through a different domain, as in plant Type II MADS box factors tetramers (Lai, Daher, et al., 2019). The nature of the binding partners participating in the oligomerization can contribute to the binding specificity of the complex, determining the recognition of different regulatory regions and therefore the activation of distinct genes (Smaczniak et al., 2012).

TF-TF interactions between different TFs can also strengthen DNA binding. For example, it has been shown that PHYTOCHROME-INTERACTING 4 (PIF4), a bHLH TF involved in light and temperature responses, can directly interact with CYCLING DOF FACTOR 2 (CDF2), a DOF TF that is temporally regulated by the circadian clock, to promote hypocotyl cell elongation (H. Gao et al., 2022). PIF4 dimers can form tetramers that increase their DNA-binding affinity and can interact with CDF2 to enhance its binding strength on DNA, securing its access to genes involved in hypocotyl elongation in response to light.

Transcriptional complexes can also involve non-TF proteins acting as cofactors, which can in some cases modify the binding specificity of the TF itself. This is what happens in the protein complex involving LFY, a master plant TF involved in flowering, and UNUSUAL FLORAL ORGANS (UFO), an F-box protein (Rieu et al., 2023). This recent and particular case will be treated in more detail in Chapter 2: The LFY-UFO complex regulates distinct genes from LFY, p. 67.

If multiple proteins bind to the same regulatory regions, it should be detectable when comparing their individual ChIP-seq binding profiles in the same tissue or cell line. Indeed, when integrating over 100 ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) project, Gerstein et al. found that different TF pairs associate on regulatory regions, and that those combinations differ between proximal (i.e. close to the TSS) and distal (i.e. far from the TSS) regulatory regions (Gerstein et al., 2012).

A different approach relies on finding DNA-bound proteins by looking at accessible chromatin regions at high resolution. Techniques such as DNase-seq allow this approach at the genome-wide level. DNase I is an endonuclease that preferentially cleaves accessible DNA regions, and it has been extensively used to map gene regulatory regions (Sullivan et al., 2015). Combining DNase I digestion of isolated nuclei with high-throughput sequencing allows the identification of accessible regions genome-wide, and within them, the stretches of DNA that are protected by DNA-bound proteins and thus not cleaved. Such information can be used to find the (combinations of) TFs bound to accessible regulatory regions (Song & Crawford, 2010).

Vierstra et al. leveraged high-resolution mapping of DNase I hypersensitive sites (DHS) to dissect the TF combinations bound to accessible regulatory regions in hundreds of cell and tissue types (Vierstra et al., 2020). They were able to distinguish the occupancy of separate TFs as opposed to that of TF complexes based on the size of the identified TF footprint (i.e. the DNA stretch bound by the TF and thus protected from DNase I cleavage). Based on their results it appears that, in humans, accessible regulatory regions are, on average, bound by 5 TFs at the same time, at a given distance from each other, within a ~200 bp region (Vierstra et al., 2020).

While I focused on TF-DNA binding so far, it is crucial to note that total number and diversity of TFBS in regulatory regions can also influence downstream gene expression. TFBS diversity is intended here as the number of *different* TFs binding to a certain sequence. Synthetic enhancers tested in mouse and human embryonic stem cells revealed that, for a comparable number of TFBSs, more diverse regulatory regions showed higher enhancer activity than same-TF sequences (Singh et al., 2021). Moreover, enhancer activity was lost in cells with less than 10 different TFBSs, indicating that the (i) amount of TFBSs in regulatory sequences and (ii) their diversity are both important to regulate gene expression (Singh et al., 2021).

In addition to the overall amount and diversity of TFBSs in regulatory regions, their order has also been shown to be important for gene expression, with some TF-TF combinations in a precise order leading to a stronger transcriptional response than others (Georgakopoulos-Soares et al., 2023).

Not only are certain TF-TF combinations preferred compared to others, but their distance can also be informative. In plants, auxin response factor (ARF) TFs have been shown to bind at preferential distances and orientations on auxin response elements (Freire-Rios et al., 2020; Stigliani et al., 2019). Such preferential distances can also be due to steric and conformational constraints, as in the case of tetrameric binding of MADS TFs (Lai, Stigliani, et al., 2019; Lai, Vega-Leon, et al., 2021).

For most TFs, such steric constraints imposed by the interaction of a TF with other proteins contribute to specificity by restricting the pool of potential binding sites on the genome. This results in TFs from the same family sharing some target sites but not others. Indeed, experiments looking at genome-wide binding of DBDs without the rest of the protein showed that, for many TFs, potential target sites are a somewhat broader pool compared to those that are bound *in vivo* when the full protein is present (Brodsky et al., 2020).

TF combinations driving the expression of specific genes or processes can be conserved across multiple species, although the relationship between conservation and regulatory function is not always straightforward. In the next section, I will try to highlight some key aspects of this relationship and of the relevance of evolutionary conservation evidence for the study of regulatory regions.

1.2.4 Role of evolutionary conservation in the study of gene regulation

Comparing the genomes of different species can give insights about which genomic regions are more or less prone to variation, and this can give information about their functional importance. As mutations in protein-coding regions can compromise protein function (and ultimately survival), such genomic regions are more conserved even between distantly related species; on the other hand, genomic regions that do not code for proteins,

commonly referred to as “noncoding regions”, generally display poor conservation (Thomas et al., 2003).

Whole-genome alignments between different species can be used to estimate conservation scores at single-nt resolution. This is the objective of methods such as PhastCons and PhyloP, although the two rely on slightly different assumptions. PhastCons aims at estimating the probability that a nucleotide is found in a conserved element as a result of negative selection, taking into account the level of conservation of its flanking regions (Siepel et al., 2005). On the other hand, PhyloP aims at estimating changes in substitution rates at individual nucleotides compared to expected levels under neutral drift, so that lower substitution rates than expected imply conservation while higher than expected indicate rapidly evolving regions (K. S. Pollard et al., 2010). Both PhastCons and PhyloP conservation scores have initially been computed on the human and mouse genome at multiple evolutionary scales (K. S. Pollard et al., 2010). More recently, the same methods have been applied to over 60 plant species (Tian et al., 2020). These resources are extremely valuable to estimate the conservation level of genomic regions and to give insights about their biological function.

Detection of conservation in noncoding regions, similarly to conservation of gene-coding regions, is often interpreted as a sign of functional importance, as their mutation can impact the action of important regulators resulting in deregulation of downstream genes (D. A. Pollard et al., 2004; Siepel et al., 2005; Woolfe et al., 2004). Therefore, many comparative genomics methods aim specifically at identifying conserved noncoding sequences (CNSs, i.e. stretches of conserved nucleotides in noncoding regions). Such approaches revealed that CNSs are enriched in TFBSs and are associated with developmental genes (Bejerano et al., 2004; Berthelot et al., 2018; Burgess & Freeling, 2014; Woolfe et al., 2004).

CNSs have been detected in both animals and plants, but important differences in genome structure, evolutionary history and more recent whole genome duplication events in plants make it difficult to directly compare findings in the two kingdoms (Murat et al., 2012). When the same CNS detection approach was applied both on plant and animal genomes, it revealed important differences in the amount and features of CNSs, with animal CNSs being more abundant and longer, as well as being more syntenically conserved than those

detected in plants (Reneker et al., 2012). Therefore, as this section aims at providing examples of the presence, features and functional importance of highly conserved regions in gene regulation, evidence coming from animal and plant systems will be treated separately.

1.2.4.1 Animal systems

An important study published in 2004 showed that the human genome contains over 400 loci of at least 200 bp that can also be found, identical, in the mice genome, despite their divergence nearly 100 million years ago (Mya) (Bejerano et al., 2004; Nei et al., 2001). Most of these highly conserved sequences were found, almost identical, in about 100 other vertebrate genomes with divergence times up to 400 Mya (Bejerano et al., 2004). The majority of these ultra-conserved loci, although not all of them, are found in noncoding regulatory regions and are associated with developmental genes, supporting their functional importance (Bejerano et al., 2004).

Distal regulatory regions such as enhancers and their surrounding genes have also been reported to be evolutionarily conserved and, in some cases, to drive similar expression patterns in distant species (Dickel et al., 2018; Snetkova et al., 2021; Wong et al., 2020). While the high level of conservation in these regions may suggest that any mutation in their sequence leads to severe phenotypic defects, it was reported that this is not always the case (Dickel et al., 2018; Snetkova et al., 2021). In particular, a systematic mutational study of early embryonic development in mice showed that the majority of the tested highly conserved enhancers can tolerate the mutation of 2 to 5% of their most-conserved nucleotides (Snetkova et al., 2021). The fact that low rates of mutations are not lethal simultaneously highlights the functional importance of their conserved sequence (which leads to severe defects beyond 5% sequence mutation), and that highly conserved regions still tolerate low levels of mutation.

1.2.4.2 *Plant systems*

As previously mentioned, whole-genome duplications and a complex evolutionary history can make the detection of CNSs more challenging in plant genomes compared to vertebrate genomes (Murat et al., 2012; Reineke et al., 2011). One of the most important CNS studies in plants was published in 2013, and found over 90,000 conserved sequences among a set of Brassicaceae species (Haudry et al., 2013). These sequences were rather short (median length less than 40 bp) and the majority of them was found in regulatory regions upstream of the TSS, which explains their enrichment in TFBS and transcription-associated motifs (e.g. TATA box). Interestingly, detection of these CNSs outside of Brassicaceae was rather low, ranging from less than 1% in rice to 3.4% in papaya (more recently diverged, less than 100 Mya (Ming et al., 2008)), and it mostly concerned CNSs overlapping small noncoding RNAs. These findings are in stark contrast with the evidence previously presented for vertebrates, where longer elements were detected, mostly identical, at even higher evolutionary distances (Bejerano et al., 2004).

The data from Haudry et al. were recently used to study conservation of Brassicaceae CNSs in different *Arabidopsis* accessions (Yocca et al., 2021). While most CNSs found in the reference Col-0 are also found in the other *Arabidopsis* accessions, a few hundreds of them are actually missing. Moreover, nearly 1000 CNSs exhibit positional variation in other *Arabidopsis* accessions, i.e. they are present in another accession but in a different locus compared to the Col-0 reference. Interestingly, CNSs showing positional variation are on average shorter (<20 bp) than CNSs retaining their position in multiple accessions (40 bp). These results highlight how CNSs can vary even within the same species.

A different approach applied on 12 dicotyledonous species (not only Brassicaceae) also identified over 90k CNSs, which only partly overlapped with those found by Haudry et al. (Velde et al., 2014). Over three quarters of these CNSs were less than 20 bp long, and the majority was found up to 1 kb upstream of the TSS. By further increasing the evolutionary distance to include monocotyledonous species, another study found over 1M CNSs, of which >70k in *Arabidopsis* (Velde et al., 2016). Once again, the majority of these CNSs were found in regulatory regions, in this case within 500 bp upstream of the translation start site. CNSs

found in both of these studies were significantly enriched in ChIP-seq peaks, highlighting their biological and regulatory functions.

More recently, over 1M *Arabidopsis* genomic regions were identified as conserved among over 60 genomes of flowering plants (both mono- and dicotyledonous species) diverging ~160 Mya (Tian et al., 2020). Combining conservation information with binding and expression data for 21 TFs, the authors showed that TFBSs with high conservation scores and high binding affinity were more likely to have a functional role. It is worth mentioning that this study provided the community with the first plant resource of genome-wide conservation scores after multiple genome alignment of flowering plant species (Tian et al., 2020).

This evidence stresses the difference between sequence conservation, which can be detected by searching multiple genomes for CNSs to infer the functional importance of genomic regions, and gene regulation conservation, which requires experimental evidence of TF binding and expression profiling in multiple species. In distantly related species, conservation of regulatory modules seems to be the key to the conservation of gene regulatory networks, and it allows TFBS displacement in evolving regulatory sequences while retaining key TF combinations to drive gene expression (Maher et al., 2018; Ravel et al., 2014; Taher et al., 2011; Wong et al., 2020).

1.3 Flowering in Arabidopsis: an ideal system to study gene regulation

Arabidopsis is one of the most important model plants. It is a dicotyledonous species belonging to the group of angiosperm (flowering) plants. Its short life cycle, which lasts 8-12 weeks from germination to seed harvesting, and its ability to self-fertilize, make it a suitable candidate for genetic studies as well as plant physiology and development (Figure 1.3-1) (Krämer, 2015).

Arabidopsis was the first plant for which a reference genome was sequenced and published (The Arabidopsis Genome Initiative, 2000). The Arabidopsis nuclear genome is around 120 Mb in size, it is diploid and it is organized in 5 chromosomes. It is, to this day, one of the best-annotated plant genomes available, and numerous resources exist to mine it (Berardini et al., 2015; Krishnakumar et al., 2015). In addition to the availability of a reference Arabidopsis genome, over a thousand different Arabidopsis accessions, native from all over the world, have been sequenced over the years, and they revealed a complex history of migration and a wide spectrum of variation within this species (Alonso-Blanco et al., 2016). All these features surely contributed to Arabidopsis becoming the plant species with the greatest amount of published genomic datasets (Fu et al., 2022).

Arabidopsis is also an established model for plant development. After a first vegetative phase during which the SAM produces rosette leaves, there is a switch from vegetative to reproductive development. This change is visually apparent as Arabidopsis plants go through bolting, that is rapid stem elongation (Figure 1.3-1).

The onset of flowering is a crucial step in plant development, as its timely initiation determines species survival. It is a tightly regulated process and it requires the integration of multiple environmental and endogenous signals, and therefore it provides an ideal system to study gene regulation in all its complexity.

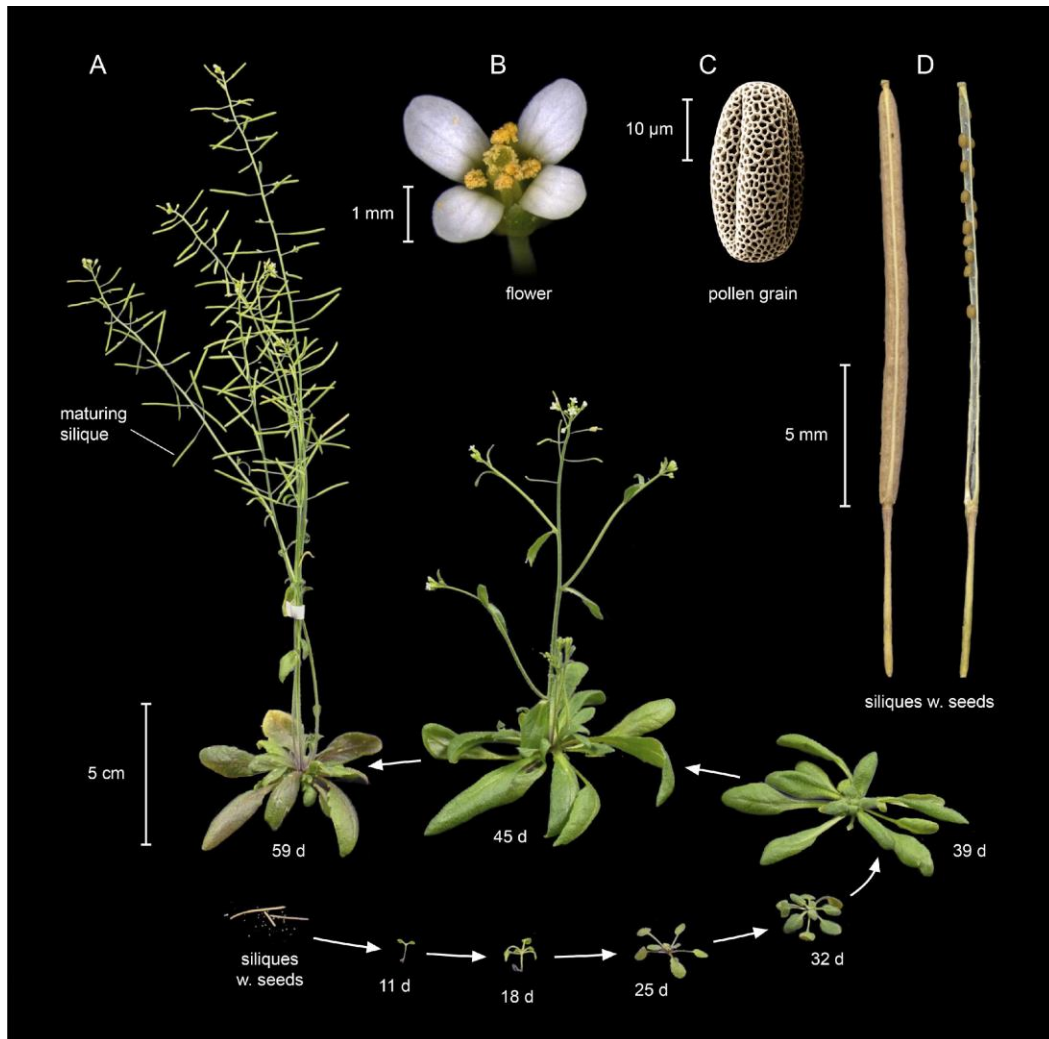


Figure 1.3-1 *Arabidopsis* life cycle, from (Krämer, 2015). A: Adult *Arabidopsis* plant with developing siliques on its fully developed stems. Siliques contain the seeds (see panel D), each giving rise to seedlings with increasing numbers of rosette leaves during the vegetative stage (bottom part of the figure). Upon the transition to flowering (reproductive stage), the plant will undergo rapid stem elongation (see the difference in plant height between 39 d plants and 45 d plants) and start developing flowers. B: *Arabidopsis* flower, with four petals, six stamens and two fused carpels visible. C: Pollen grain of *Arabidopsis*, which contains the male gamete. D: *Arabidopsis* siliques, containing seeds.

1.3.1 Flowering in *Arabidopsis* and the ABCDE model

For flowering to happen in favorable conditions, plants perceive and integrate several environmental and endogenous signals. Such signals include day length, temperature and nutrient availability as well as phytohormones such as gibberellins and auxin. Several interconnected pathways, made of highly complex gene regulatory networks, participate in its regulation and eventually converge on a few so-called floral integrators that promote

flowering. I will not cover this topic in detail, but recent reviews of the main pathways involved are available about light signaling and photoperiod (Freytes et al., 2021), vernalization (Costa & Dean, 2019), ambient temperature and other environmental stimuli (Cho et al., 2017), age (Hyun et al., 2017), hormonal control (Izawa, 2021) and the autonomous pathway (Cheng et al., 2017).

The floral transition leads to the conversion of the shoot apical meristem into an inflorescence meristem. Then, floral meristems are produced on the flanks of the inflorescence meristem, which will start developing floral organs.

Arabidopsis flowers are composed of four whorls, which harbor, from the most external to the most internal one, four sepals, four petals, six stamens and two fused carpels (Figure 1.3-2). Genetic studies in *Arabidopsis* identified five classes of floral homeotic genes governing the development of these structures: class A, containing *APETALA1* (*AP1*) and *APETALA2* (*AP2*); class B, represented by *APETALA3* (*AP3*) and *PISTILLATA* (*PI*); class C, fulfilled by *AGAMOUS* (*AG*); class D, with *SHATTERPROOF* (*SHP*) and *SEEDSTICK* (*STK*), and class E, with *SEPALLATA1-4* (*SEP1-4*) (Wellmer et al., 2014). A-class genes control sepal formation, petals are specified by a combination of A- and B-class genes, stamen development relies on B- and C-class combined activities and finally carpel formation is regulated by C-class gene *AG* (Figure 1.3-2). D-class genes control ovule development, while E-class genes are partially redundant and they are required for the specification of all floral organs in *Arabidopsis* (Figure 1.3-2). A-class TF *AP2* also represses C-class *AG* expression in the outer whorls, to effectively separate their different action territories (Krogan et al., 2012). All the floral organ identity genes mentioned code for MADS-box TFs with the exception of *AP2*, which encodes an AP2 family TF. The four classes of MADS TFs form tetrameric protein complexes that bind pairs of CArG-box motifs to regulate their target genes (Figure 1.3-2) (Mendes et al., 2013).

Expression of ABCDE genes is regulated by *LFY* and *AP1*, in some cases through the interaction with cofactors. The role of *LFY* in flowering will be detailed in the following section.

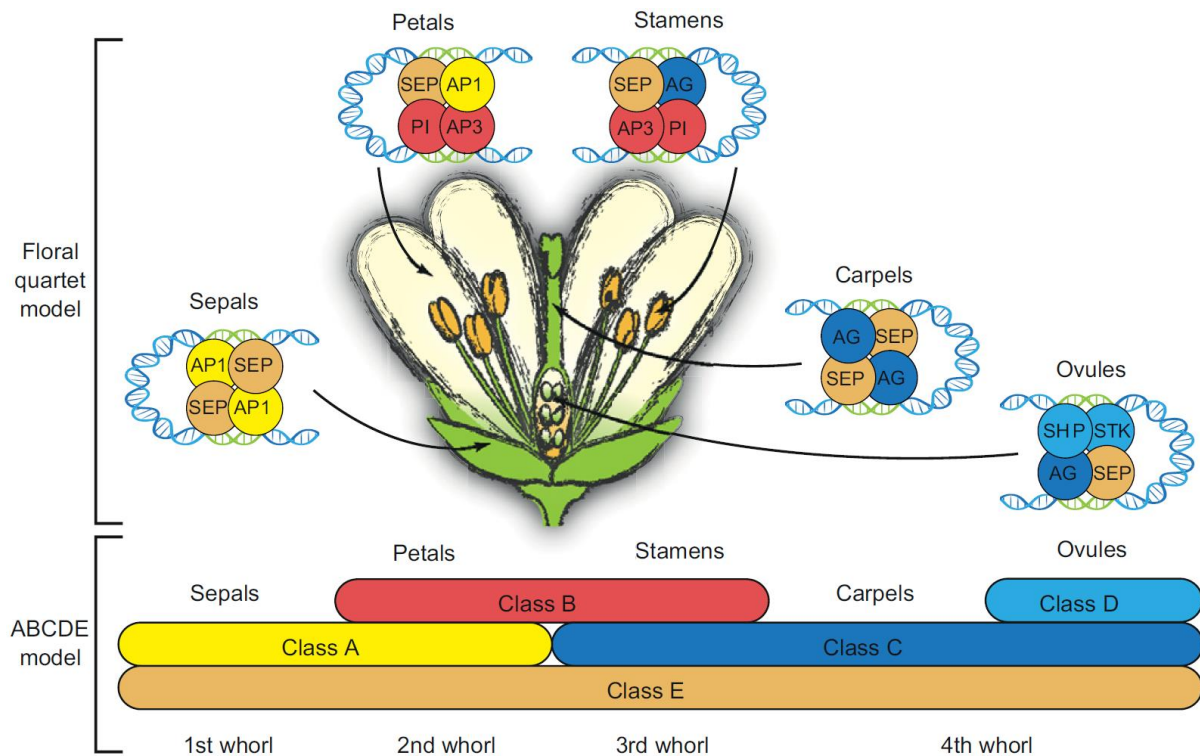


Figure 1.3-2 *Arabidopsis* flower and the ABCDE model, from (Theißen et al., 2016). The floral quartet model and the underlying ABCDE model of organ identity determination in *Arabidopsis thaliana*. From top to bottom: floral quartet model, where the five floral organ identities (sepals, petals, stamens, carpels and ovules) are specified by the formation of tetrameric complexes of MADS-domain transcription factors by binding to cis-regulatory TFBs (in green). Sepal identity is determined by a complex of two AP1 proteins and two SEP proteins; petals are specified by AP1, a SEP protein and B-class TFs PI and AP3; stamens require AG, a SEP TF, and PI and AP3; two SEP proteins and two AG proteins specify carpel identity, and a combination of one SEP protein, AG, and SHP and/or STK control ovule identity. At the bottom, the ABCDE model with the corresponding gene classes and the whorls that they specify is displayed.

1.3.2 LEAFY, a master floral regulator

LFY is the master regulator of flower development and, in *Arabidopsis*, it controls both floral meristem fate and patterning (Parcy et al., 1998; Schultz & Haughn, 1991; Weigel et al., 1992). Strong *lfy* mutants in *Arabidopsis* display delayed flowering and conversion of flowers to inflorescence- or leaf-like structures, often subtended by floral bracts, which are male sterile (Figure 1.3-3B, D, F) (Parcy et al., 1998; Schultz & Haughn, 1991; Weigel et al., 1992). Conversely, LFY overexpression induces early flowering, production of ectopic flowers in the axils of rosette leaves and conversion of meristem into solitary flowers (Figure 1.3-3G, H) (Chahtane et al., 2018; Kardailsky et al., 1999; Kobayashi et al., 1999; Sayou et al., 2016).

LFY is expressed since the early stages of flower development (Blázquez et al., 1997; Weigel et al., 1992; Yamaguchi et al., 2016). Because of its crucial involvement in flowering, its expression is tightly regulated to ensure its correct spatiotemporal induction. The photoperiod pathway regulates LFY levels and its temporal activation through SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1), AGAMOUS-LIKE24 (AGL24) and SHORT VEGETATIVE PHASE (SVP) TFs as follows. SOC1, a floral integrator, forms a complex with AGL24 to induce LFY (J. Lee et al., 2008), and both AGL24 and SVP have been shown to bind the LFY promoter and induce its expression (Grandi et al., 2012). Auxin increase in the floral anlage prompts LFY induction by MONOPTEROS (MP), an auxin-response factor (Yamaguchi et al., 2013), while the repressor and shoot identity TF TERMINAL FLOWER 1 (TFL1) prevents LFY expression in the SAM, ensuring spatial regulation (Bradley et al., 1997; Hanano & Goto, 2011). LFY regulation by TFL1 is also linked to the photoperiodic pathway through the floral integrator FLOWERING LOCUS T (FT): TFL1 and FT seem to compete for complex formation with FD at target FD sites on LFY, with the TFL1-FD complex repressing and FT-FD inducing LFY expression, respectively (Y. Zhu et al., 2020).

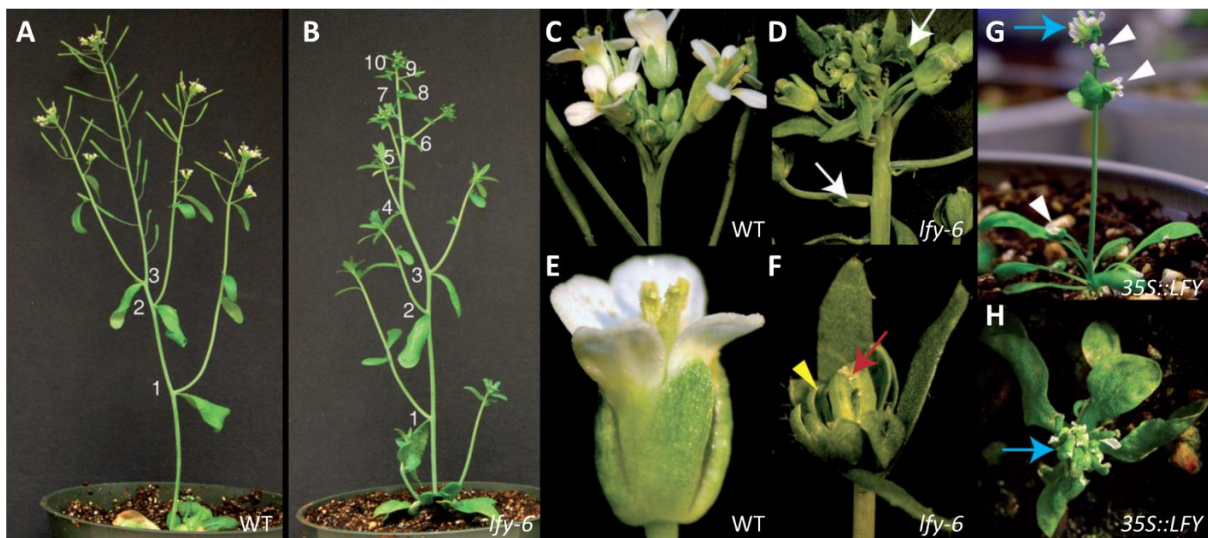


Figure 1.3-3 Effects of *Lfy* knock-out mutation and overexpression in *Arabidopsis*, adapted from (Siriwardana & Lamb, 2012). *Arabidopsis* plants (A, B, G), inflorescences (C, D, H) and flowers (E, F). Genotypes are indicated on the picture. In B, numbers indicate cauline leaves with subtending branches. Bracts (white arrows), partially fused carpel-like organs (red arrow) and leaf-like sepals (yellow arrowhead) are indicated on *lfy* plants. Terminal flowers (blue arrows) and single flowers formed at the base of the stem in place of branches (white arrowheads) are indicated on LFY overexpressor plants.

LFY induces the expression of ABCE flowering genes to control floral patterning. LFY can directly induce A-class *AP1* expression through binding to its target sites on the *AP1* promoter (Parcy et al., 1998; Wagner et al., 1999), but its proper induction additionally requires FT and FD (Wigge, 2005) as well as BLADE ON PETIOLE1 and 2 (BOP1 and 2) (Chahtane et al., 2018). Moreover, B- and C-class gene expression requires additional LFY cofactors. For correct *AP3* and *PI* expression both LFY and UNUSUAL FLORAL ORGANS (UFO), an F-box cofactor, are required (Honma & Goto, 2000; Lamb et al., 2002). More details on how LFY works with UFO to regulate petal and stamen development will be given in Chapter 2: The LFY-UFO complex regulates distinct genes from LFY, p. 67. Finally, C-class gene (*AG*) expression, which drives carpel development as well as flower termination, requires LFY and WUSCHEL (*WUS*), a homeodomain TF controlling shoot meristematic activity (Jha et al., 2020; Lohmann et al., 2001). While these LFY co-factors are known, the exact mechanism is not.

LFY's involvement in the early stages of flower development is also made possible by its pioneer TF properties. LFY can bind nucleosome-occupied DNA in closed chromatin regions, and recruit chromatin remodelers at its *AP1* binding site (Jin et al., 2021; Lai, Blanc-Mathieu, et al., 2021). Access to closed chromatin regions is possible thanks to the N-terminal SAM domain of LFY (Sayou et al., 2016). LFY also physically interacts with SPLAYED (*SYD*) and BRAHMA (*BRM*), two closely related SWI2/SNF2 chromatin remodelers, on *AG* and *AP3* regulatory regions (M.-F. Wu et al., 2012).

In addition to the ABCE floral homeotic genes, LFY has also been shown to regulate genes involved in flowering time and meristem identity as well as organ polarity, hormone responses and resistance to pathogens (Winter et al., 2011).

1.3.3 LFY is a highly conserved TF

The structure of LFY's N-terminal SAM domain, required for LFY-LFY oligomerization and cooperative binding, and its C-terminal DBD have been solved (Hamès et al., 2008; Sayou et al., 2016). LFY was initially indicated to bind a 7-bp CCANTG[G/T] consensus sequence (Busch et al., 2022; Lamb et al., 2002), but further studies have refined the LFY target motif to a 19-bp palindromic site (Moyroud et al., 2011). Interestingly, the LFY motif presents trinucleotide

dependencies on the flanks of the core motif and between the three bases separating the two LFY monomers (Moyroud et al., 2011).

The availability of LFY sequences from other plant species and the structure of the DBD allowed a thorough exploration of LFY's functional evolution. LFY's protein sequence is highly conserved throughout plant evolution, especially its SAM domain and DBD (B. Gao et al., 2019; Maizel, 2005; Sayou et al., 2014). Unlike most TFs, which are part of bigger and partially redundant TF families, LFY makes up a TF family of its own. Moreover, it is found in single copy in most plants, although some notable exceptions are some gymnosperm or moss species and crops such as maize and soybean, which contain two copies of LFY (B. Gao et al., 2019).

The binding specificity of LFY is also highly conserved in the green lineage. LFY has three possible binding conformations, which are linked to few but crucial variations in the residues of its DBD. In embryophytes, LFY specificity is conserved and the TF binds as an obligate dimer, with two DBDs facing each other on the same side of the DNA and separated by 3 bp (Sayou et al., 2016).

While LFY's sequence and binding specificity are highly conserved, its importance for floral meristem initiation and patterning can vary in other plant species. LFY homologs have been shown to be involved in inflorescence and overall plant architecture in rice and maize, and additional roles in leaf dissection have also been reported in pea and petunia (Moyroud et al., 2009).

1.3.4 LFY binding and expression resources in Arabidopsis

Besides the central role of LFY in genetic, evolutionary and structural studies, this TF has been the subject of many binding and expression studies.

Several datasets of *in vivo* genome-wide binding assays have been published over the years. The first ChIP-seq experiment to assess genome-wide LFY binding on DNA *in vivo* was performed on LFY-overexpressing seedlings and published in 2011 (Moyroud et al., 2011). A second ChIP-seq experiment in the same overexpression setting was published a few years later (Sayou et al., 2016). In 2017, a ChIP-seq on Arabidopsis inflorescences expressing 35S:LFY-GR in the *ap1 cal* background was published (Goslin et al., 2017), and the latest

ChIP-seq experiment was performed on callus expressing the same 35S:*LFY-GR* transgene (Jin et al., 2021). In addition to *in vivo* binding profiling, *in vitro* binding was assayed more recently through DAP-seq and ampDAP-seq experiments (Lai, Blanc-Mathieu, et al., 2021).

A variety of genome-wide expression datasets is also available for LFY, even though the majority is represented by microarray experiments (Schmid et al., 2003, 2004; William et al., 2004). Only one LFY RNA-seq experiment has been published so far, performed on callus expressing 35S::LFY-GR recombinant protein (Jin et al., 2021).

All the features of LFY mentioned above make it an excellent TF model to study transcriptional regulation in Arabidopsis.

2 Objectives

The overall objective of my PhD work was to study transcriptional regulation by TFs in Arabidopsis, by taking LFY as a model TF. This objective has taken the shape of two main projects, which I will explain in further detail in the two following chapters of this manuscript:

1. A machine-learning approach to characterize LFY's transcriptional regulation genome-wide, distinguishing between transcriptionally active and inactive LFY sites on the genome. I built a classifier that can distinguish between the two classes of LFY sites based on site quality, genomic context and evolutionary conservation. This will be the focus of Chapter 1: A machine-learning model to predict transcriptional regulation of LFY sites genome-wide based on genomic context and evolutionary conservation, p. 45.
2. An approach combining biochemistry, genomics and structural biology to elucidate the role of the LFY-UFO complex in the development of Arabidopsis petals and stamens. While I was not directly involved in the biochemical and structural parts of the project, the computational component I contributed to was instrumental to uncover the different specificity of the LFY-UFO complex compared to LFY's canonical binding, as well as the role of UFO in allowing LFY to target new sites through its involvement in a transcriptional complex. This second project will be introduced in Chapter 2: The LFY-UFO complex regulates distinct genes from LFY, p. 67.

The first chapter will contain a specific introduction, as well as results and discussion sections. In the second chapter, a short introduction will detail my involvement in the recently published study that constitutes the core of the chapter itself, and that will be followed by a section containing additional results and discussion integrating new recent findings on the same topic. Finally, an overall conclusion will end this manuscript.

3 Chapter 1: A machine-learning model to predict transcriptional regulation of LFY sites genome-wide based on genomic context and evolutionary conservation

3.1 Introduction

LFY is a master regulator of floral development, and it triggers major changes in this important process (Kaufmann & Airoidi, 2018; Parcy et al., 1998, p. 199; Wagner et al., 1999; Weigel et al., 1992; Weigel & Nilsson, 1995). Decades of research have elucidated its key role in multiple plant species, as well its structural and biophysical properties (Hamès et al., 2008; Moyroud et al., 2009, 2010, 2011; Sayou et al., 2014, 2016).

As LFY holds a central role in flower development, its action is tightly regulated in space and time, making it an excellent TF model to study the complexity of gene regulatory processes. LFY's binding profile has been extensively studied, both *in vivo* and *in vitro*, making it one of the plant TFs with the most binding data available (Goslin et al., 2017; Jin et al., 2021; Lai, Blanc-Mathieu, et al., 2021; Moyroud et al., 2011; Sayou et al., 2016; Winter et al., 2011). Moreover, the study of LFY's DNA binding characteristics has led to the development of state-of-the-art TFBS models specifically tailored to predict its binding to DNA sequences. First, a PWM with nucleotide dependencies, which dramatically increases prediction accuracy compared to a classical PWM (Moyroud et al., 2011); second, POcc, a biophysical model outperforming the previous ones (Minguet et al., 2015; Moyroud et al., 2011). However, such models alone do not provide any information about whether the binding really happens *in vivo*, or whether it has a measurable effect on target gene expression, and this hinders their application.

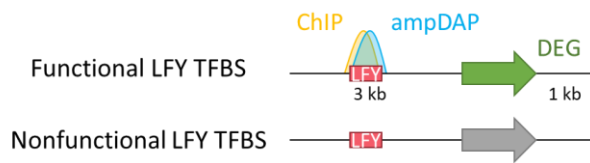
Upon TFBS recognition, TFs can recruit the transcriptional machinery to initiate gene expression (H. Chen & Pugh, 2021). Transcriptional profiling with microarrays or RNA-seq techniques can capture gene expression levels at the tissue or single-cell level, and this information could be integrated with binding profiles to better characterize the panorama of TFBSs with an active transcriptional role. Multiple studies have investigated LFY's target genes, providing information about LFY-dependent gene expression at various developmental stages and through different experimental designs (Jin et al., 2021; Schmid et al., 2003, 2004; William et al., 2004).

TF-DNA binding does not determine effective transcriptional regulation of target genes, as gene expression can be influenced by other factors including the effect of multiple TFs on regulatory regions (Spitz & Furlong, 2012). Therefore, we sought to make a comprehensive model of genome-wide transcriptional regulation by TFs in Arabidopsis, and all the information stated above makes LFY an ideal candidate for this task. In particular, we wanted to know why only a subset of LFY-bound regions genome-wide are regulated by the protein *in vivo*, and whether the information present in their genomic context could help explain these preferences.

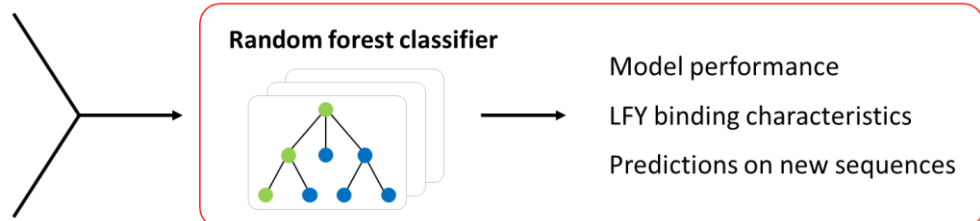
To this end, I identified LFY sites genome-wide based on their score using LFY's PWM with dependencies (Moyroud et al., 2011). Then, I analyzed and integrated LFY binding and expression data to define whether each PWM-identified LFY site was 'functional' (i.e. transcriptionally active, based on evidence of binding *in vitro* and *in vivo*, and having a significant effect on gene expression) or 'nonfunctional' (i.e. transcriptionally inactive, lacking any binding or gene expression evidence) (Figure 3.1-1). For each LFY site, I computed state-of-the-art PWM and POcc scores, and I investigated the presence of TFBSs for other TFs surrounding LFY sites, their density and diversity. I also looked at the distance between each LFY site and the closest TSS, as some TFs have been shown to bind at preferential distances from the TSS in Arabidopsis and maize, and we did not know whether this could also influence transcriptional regulation by LFY (Bernard et al., 2010; Rozière et al., 2022). Moreover, I computed the level of conservation of each LFY site in flowering plants, under the hypothesis that conserved regulatory elements would be more likely to be transcriptionally active. Finally, I trained Random Forest algorithms to distinguish functional and nonfunctional LFY sites based on all of the information stated above (Figure 3.1-1).

Once the model was trained, I used it to investigate the rules underlying transcriptional regulation by LFY, as well as to make predictions on LFY sites that were not included in training, revealing new potential targets of this TF (Figure 3.1-1). For this purpose, I took advantage of a set of LFY sites that I labeled as 'unknown' as they had some evidence of transcriptional regulation, but they did not fulfill all requirements to be confidently labeled as 'functional'. These sites were not included in model training but were used to discover new LFY-regulated sites on the Arabidopsis genome for which the available evidence was inconclusive.

① Labels: functional and nonfunctional sites



ChIP: ChIP-seq peak
 ampDAP: ampDAP-seq peak
 DEG: Differentially Expressed Gene
 LFY: LEAFY TFBS
 TF1: TFBS of TF1; etc.



② Features: Genomic context + evolutionary conservation



Figure 3.1-1 Graphical abstract for Chapter 1: A machine-learning model to predict transcriptional regulation of LFY sites genome-wide based on genomic context and evolutionary conservation. (1) LFY binding sites (pink boxes) are predicted genome-wide using a PWM model with nucleotide dependencies (Moyroud et al. 2011), and then they are labeled as ‘functional’ or ‘nonfunctional’ based on binding and expression assays. These labeled data are used to train a Random Forest model, which aims to accurately classify LFY sites as functional or not. (2) Genomic features describing each predicted LFY BS are fed into the Random Forest model. These features include the genomic context, which measures the presence of the binding site of all *A. thaliana*’s TFs nearby the LFY BS, and a measure of evolutionary conservation. Once the model achieves high performance, it can be used to predict the state of new sequences, providing valuable biological insights into TF biology. Moreover, we can identify which genomic features contribute significantly to the model’s performance, shedding light on the molecular mechanics that govern gene regulation by LFY.

3.2 Results

3.2.1 Genomic context features better predict functional LFY sites than features based on state-of-the-art LFY-DNA binding models

As mentioned above, the first step consisted in defining LFY TFBSs on the Arabidopsis genome, and labeling them as either functional, nonfunctional or ‘unknown’ based on experimental data (more details in Definition of LFY sites genome-wide, p. 115 and Integration of binding and expression data and definition of nonfunctional/functional/‘unknown’ LFY , p. 115).

Based on the amount of LFY TFBSs labeled as functional at increasing PWM score percentiles (Figure S3.4-1A and C), we chose the 99.9th percentile threshold, which corresponds to a PWM score of -18.45, to define genome-wide LFY TFBSs. Threshold choice was the result of a compromise between site quality, which increases at higher percentiles as the score gets close to 0, and the amount of functional sites (positive entries) that would be required to successfully train our algorithm. At 99.9th percentile threshold, 1204 LFY sites are labeled as functional (Figure S3.4-1, p. 65), 57699 as nonfunctional and 39488 as ‘unknown’. For the analyses presented next, I only considered functional and nonfunctional sites.

We decided to build Random Forest classifiers to distinguish functional and nonfunctional LFY sites based on genomic context information and evolutionary conservation. We chose Random Forest algorithms because they have been widely applied in biology and genomics (Back & Walther, 2021; X. Chen & Ishwaran, 2012; Smet et al., 2023), and they accommodate multiple types of features, allowing us to use both continuous and categorical features. Moreover, they can account for nonlinear interactions between features.

We first sought to compare the performance of Random Forest models in predicting functional LFY sites when trained with state-of-the-art LFY-DNA binding models (PWM with dependencies or POcc) and when trained on genomic context information. Genomic context was described by the following properties:

- Distance, in bp, between each LFY site and the closest site of other TFs, called ‘co-occurrence’ hereafter. To avoid redundancy between TFs with similar DNA-binding motifs, I used clusters of TF DNA-binding models available on JASPAR 2022 and computed based on motif similarity (Castro-Mondragon et al., 2022), and I applied

the same procedure as for LFY to define their binding sites on the Arabidopsis genome (see Computing co-occurrence and LFY-LFY distances, p. 117).

- Distance, in bp, between each LFY site and the next closest LFY TFBS; called 'LFYdist' in Figure 3.2-1A (see Computing co-occurrence and LFY-LFY distances, p. 117).
- Distance, in bp, between each LFY site and the closest TSS; called 'TSS' in Figure 3.2-1A (see Computing LFY-TSS distances, p. 118).
- Non-LFY TFBS density (i.e. total number of TFBSs) and diversity (through an adaptation of Shannon's entropy (Shannon, 1948)) around each LFY site (± 500 bp). These features are named 'density' and 'diversity' in Figure 3.2-1A, respectively. More details about how I calculated TFBS density and diversity, and how we chose these two features to include this information, can be found in section Computing TFBS density and diversity around LFY TFBSs, p. 118).
- Sequence type, i.e. whether each site was found within a promoter region, coding sequence, intron, downstream regulatory region, 5' or 3' UTR (see Encoding sequence type, p. 118). This feature is named 'seq_type' in Figure 3.2-1A.

To evaluate the performance of Random Forest algorithms in classifying functional and nonfunctional LFY sites, I used cross-validation. Briefly, this strategy consists in dividing the data into a group that will be used to train a model and another one that will be given to the trained model to evaluate its prediction power on sequences with hidden labels. The splitting procedure can be repeated multiple times and each time a new model will be trained and evaluated. Once context features were computed, I used different combinations of them to train Random Forest algorithms to classify functional and nonfunctional LFY sites. As our data is unbalanced (few positives, i.e. functional sites, and many negatives, i.e. nonfunctional sites), I used Precision-Recall (PR) curves to evaluate and compare models built with different sets of features (more details about the proposed cross-validation strategy in Training and testing Random Forest models, p. 124).

Random Forest models exclusively built with co-occurrence information of 46 TF clusters performed better than those built with LFY PWM or POcc alone, as indicated by a higher median PR AUC over 100 models (median PR AUC = 0.22 for co-occurrence, green boxplot in Figure 3.2-1A, compared to 0.05 for PWM and 0.14 for POcc alone) (more details on the

cross-validation strategy and why I obtained 100 models can be found in Training and testing Random Forest models, p. 124). Distance between two LFY sites ('LFYdist', Figure 3.2-1A) strongly improved predictions (median PR AUC = 0.29), while the addition of LFY sites to TSS distance, sequence type and TFBS density and diversity ('TSS', 'density', 'diversity', 'seq_type', respectively) seemed to only marginally improve the PR curve's AUC (Figure 3.2-1A). Including PWM scores and POcc as features in addition to all genomic context information, for a total of 58 features in the model, further increased mean PR AUC to 0.37 (Figure 3.2-1A and B), suggesting that, while their impact alone is limited, their integration within the genomic context of LFY sites provides new crucial information. Taken together, these results suggest that the genomic context of LFY sites is informative to distinguish functional and nonfunctional sites, and that context information leads to better predictions than with the exclusive use of previous state-of-the-art LFY models, PWM and POcc.

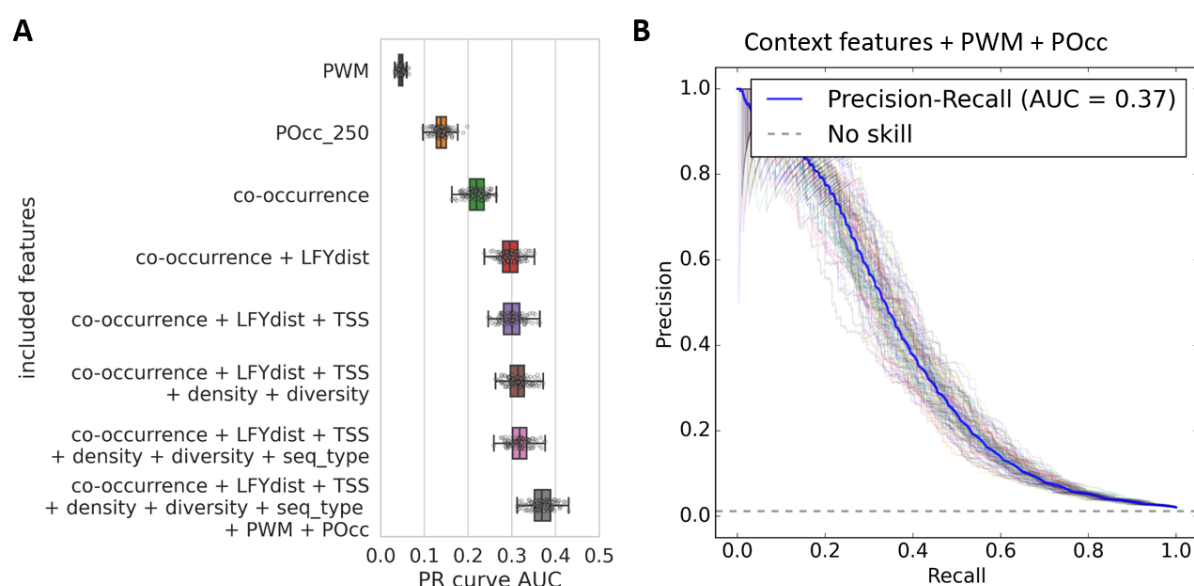


Figure 3.2-1 Performance of random forest classifiers built with state-of-the-art LFY TFBS models (PWM with dependencies and POcc) and genomic context information. A: boxplot of PR AUC in 100 models trained with different features as indicated. 'PWM' = score obtained with LFY PWM with dependencies; POcc = POcc score computed ± 250 bp around each LFY site (more details on POcc calculation and the choice of the interval around LFY can be found in Computing POcc around LFY TFBS, p. 116); all other features named here are mentioned in the text just above. B: Detailed PR curves for the model trained with all features (bottom boxplot in A, in grey). Each line corresponds to a model built at each cross-validation iteration. Blue line is the average. Horizontal dashed line in grey corresponds to the expectation of a random model. More details on the cross-validation strategy I followed can be found in Training and testing Random Forest models, p. 124.

3.2.2 Inclusion of evolutionary conservation does not improve predictions

Next, we decided to integrate evolutionary conservation to the previous model containing genomic context and LFY models features, under the hypothesis that conserved sites would be more likely to have a functional role. After trying several ways to encode conservation at LFY sites, I decided to use the average PhastCons and PhyloP scores detected at LFY sites, as computed genome-wide by (Tian et al., 2020) on Arabidopsis after the multiple genome alignment of over 60 flowering plant genomes. More information on how I chose the average conservation score over LFY sites can be found in Computing average conservation at LFY sites, p. 121. Average PhyloP and PhastCons scores at each LFY site were included as two separate features, but were used together in the model when conservation is indicated ('avg conservation' in Figure 3.2-2A). Surprisingly, conservation alone was virtually uninformative for predictions, and its integration to our previous model with genomic context and LFY models even seemed to slightly decrease its overall PR AUC (Figure 3.2-2A).

One possible explanation for this surprising result is that the evolutionary distance included in the model is too broad, and that site conservation may have been lost: LFY may no longer target the same sites, or the sites themselves may not be detected at the genome alignment stage. The range of species used by Tian et al. to compute PhyloP and PhastCons scores spanned over 100 million years of plant evolution, and included both mono- and dicotyledons (Tian et al., 2020). Therefore, I decided to use additional data sources to see whether decreasing the evolutionary distance of the species included to compute conservation could help understand the poor impact of conservation on predictions. As it was not possible to modify the species range used to compute PhyloP and PhastCons scores, I relied on four published datasets of conserved noncoding sequences (CNSs) at increasing evolutionary distances (Figure 3.2-2B), and I looked at the proportion of LFY sites overlapping with CNSs as a proxy of site conservation (Figure 3.2-2C).

The first dataset I used comprises over 90,000 CNSs computed for Brassicaceae species, and it is the one with the shortest evolutionary distance from Arabidopsis (Haudry et al., 2013). The second dataset was computed on noncoding regions of 10 dicotyledon species (Velde et al., 2014), while the third one included 12 species of both mono- and dicotyledons (Velde et al., 2016). The fourth dataset was published along with the nucleotide-resolution

conservation scores that we had already used to encode conservation in the model, but in this case it represents conserved genomic regions among 63 mono- and dicotyledon species instead of continuous nucleotide-level conservation scores (Tian et al., 2020). As the first three datasets were specifically computed on noncoding regions, I focused on LFY sites in noncoding regions.

Our results show that the proportion of functional LFY sites overlapping with CNSs is higher than that of nonfunctional sites in all the CNS datasets used (Figure 3.2-2C). Moreover, the ratio between the fraction of functional and that of nonfunctional sites overlapping with CNSs is highest when using Brassicaceae CNSs from the Haudry2013 dataset (Haudry et al., 2013), while it seems to decrease as more species are taken into account and evolutionary distances increase (Figure 3.2-2D). It should be noted, however, that this decrease is not constant, as the ratio in the CNSs from the second dataset is lower than in the third one, despite the fact that the second one is limited to dicotyledons (Figure 3.2-2D). I also checked CNS enrichment in ‘unknown’ sites, and they tend to have a CNS-overlapping rate that is lower than that of functional sites but higher than the nonfunctional ones (Figure 3.2-2C). This is due to the fact that ‘unknown’ sites represent a mixture of functional and nonfunctional sites.

Overall, these results suggest that functional sites in noncoding regions tend to be more evolutionarily conserved than nonfunctional ones, even at broad evolutionary distances. However, it remains unclear why conservation information has no impact on predictions (Figure 3.2-2A), even though the CNS dataset issued from the same publication as the data we previously used in our model (Tian2020 in Figure 3.2-2B, from (Tian et al., 2020)), still shows strong proportions of conserved sites among functional LFY sites.

Another possibility is that the high conservation levels generally observed in coding sequences cannot be distinguished from high conservation levels linked to regulatory functions. This seems to be supported by Figure S3.4-1D, which was generated following the same procedure as Figure 3.2-2C but without excluding LFY sites in coding sequences, and which shows that in the fourth dataset the proportion of conserved functional LFY sites is similar to that of nonfunctional sites. Either way, we decided not to include conservation

features in the analyses that will be presented next, in consideration of the slight decrease in median PR AUC that they produced (Figure 3.2-2A).

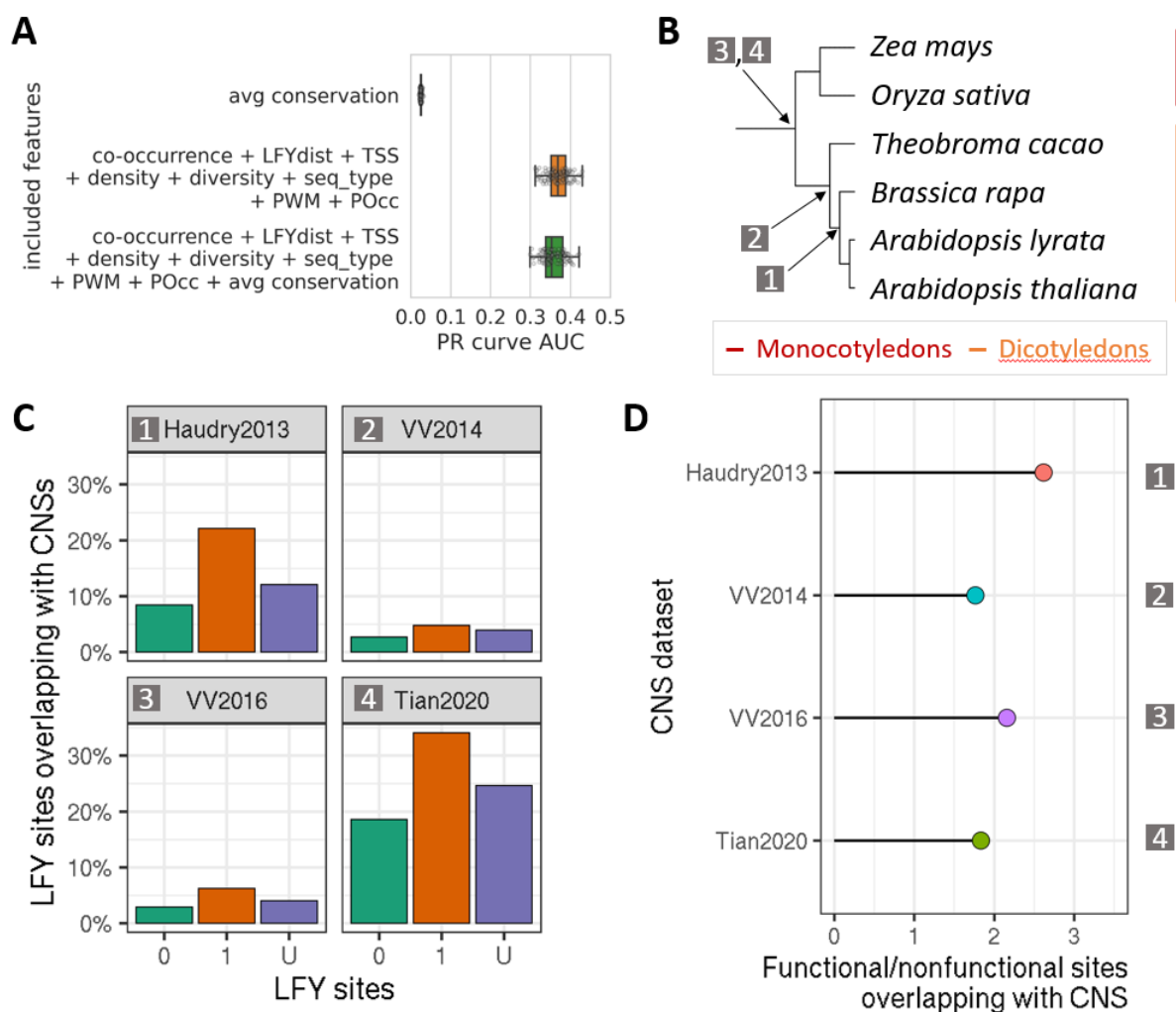


Figure 3.2-2 Including average conservation level of LFY sites does not improve predictions. A: Performance of random forest classifiers built with conservation information and/or genomic context and state-of-the-art LFY models. Each boxplot represents the PR AUC obtained from 100 models trained with different features as indicated (see Training and testing Random Forest models, p. 124, for details on cross-validation). Avg conservation = models built exclusively with average conservation scores (PhyloP and PhastCons) on LFY sites (see Computing average conservation at LFY sites, p. 121). Please note that the orange boxplot is the same as the one shown in Figure 3.2-1A in grey, and the green one corresponds to a model with the same features as the orange one plus two conservation features, for a total of 60 features in the Random Forest model. B: Species tree displaying the plant species used to compute different public CNS datasets, indicated with white numbers in boxes. As each dataset was computed with a different set of species, the ones displayed here are examples used in at least one dataset. The tree was built with TimeTree.org (Kumar et al., 2022). C: Proportion of noncoding LFY sites at least partially overlapping with CNSs at increasing evolutionary distances, based on their functionality label. 1 = functional LFY sites; 0 = nonfunctional LFY sites; U = 'unknown' LFY sites. Haudry2013 = CNSs computed by (Haudry et al., 2013); VV2014 = CNSs computed by (Velde et al., 2014); VV2016 = CNSs computed by (Velde et al., 2016); Tian2020 = CNSs computed by (Tian et al., 2020). More details on this analysis can be found in Calculating CNS enrichment at LFY sites, p. 123. D: Ratio between the proportion of functional sites and the proportion of nonfunctional sites overlapping with CNSs in panel C, for each CNS dataset.

3.2.3 The model reveals important information about how LFY binds the DNA

As I showed before, genomic context features combined with state-of-the-art LFY models lead to good predictions in cross-validation (Figure 3.2-1B, p. 50). However, the provided features (58 in total in this case) are not all equally important to separate functional and nonfunctional LFY sites. Therefore, I investigated which features were the most important for classification, as this information can be useful to understand more about the differences between the two site classes (functional and nonfunctional sites), and ultimately about how LFY fulfills its regulatory activity.

I decided to look at the Gini importance given to each feature, which measures the total decrease in node impurity, i.e. in our case, how efficiently a given feature separates a group of sites into two separate classes of functional vs nonfunctional sites. I extracted the Gini importance of all features from each random forest model in Figure 3.2-1B (p. 50) and identified the most important ones (Figure 3.2-3A, see Extracting feature importance from Random Forest models, p. 126). POcc around LFY sites ('pocc_250') was consistently the most important feature, followed by the PWM score ('PWM') and the distance of each LFY site from the next LFY site ('LFYdist') (Figure 3.2-3A). Therefore, although POcc and PWM resulted in a lower PR AUC when separately used to build random forest classifiers, these two features, all directly related to LFY sites, have higher importance scores than any genomic context features individually.

Besides LFY-related features, another important one was the level of TFBS diversity around LFY sites ('Sindex_500' in Figure 3.2-3A), highlighting that, while the position of other LFY sites is very informative, the overall presence of diverse TFBSs is also important to distinguish functional LFY sites from nonfunctional ones.

Finally, the distance of LFY from specific TF clusters, which represent variable amounts of TFs, turned out to be important as well (Figure 3.2-3A and B). These relevant clusters and the TFs they represent are shown in Table 3.2-1. Cluster_1 contains 25 TFs belonging to the plant-specific Teosinte-branched 1/Cycloidea/Proliferating (TCP) family, which have been shown to be involved in many processes including plant development and flowering (D. Li et al., 2019; S. Li, 2015). Cluster_46 contains one TF, CELL DIVISION CYCLE5 (CDC5), which has been reported to be involved in cell cycle regulation in Arabidopsis (Lin et al., 2007).

Cluster_4 contains 41 TFs predominantly belonging to the bZIP TF family, as well as some NAC TFs. bZIP and NAC are among the largest TF families in Arabidopsis, and are involved a large spectrum of processes (Blanc-Mathieu et al., 2023). Clusters 38, 32 and 18 contain TFs from different families, including several growth-regulating factor (GRF), RELATED TO ABI3 AND VP1 (RAV) and MADS TFs, which regulate a plethora of plant processes including flower development (Hugouvieux & Zubieta, 2018; Matías-Hernández et al., 2014; Omidbakhshfard et al., 2015). I want to stress the fact that the high importance of these clusters for predictions does not necessarily mean that LFY is interacting with any of them specifically; rather, it suggests that their distance from LFY differs between functional and nonfunctional LFY sites, and that this information helps classification. I will further discuss the relevance of these TF clusters in the appropriate discussion section.

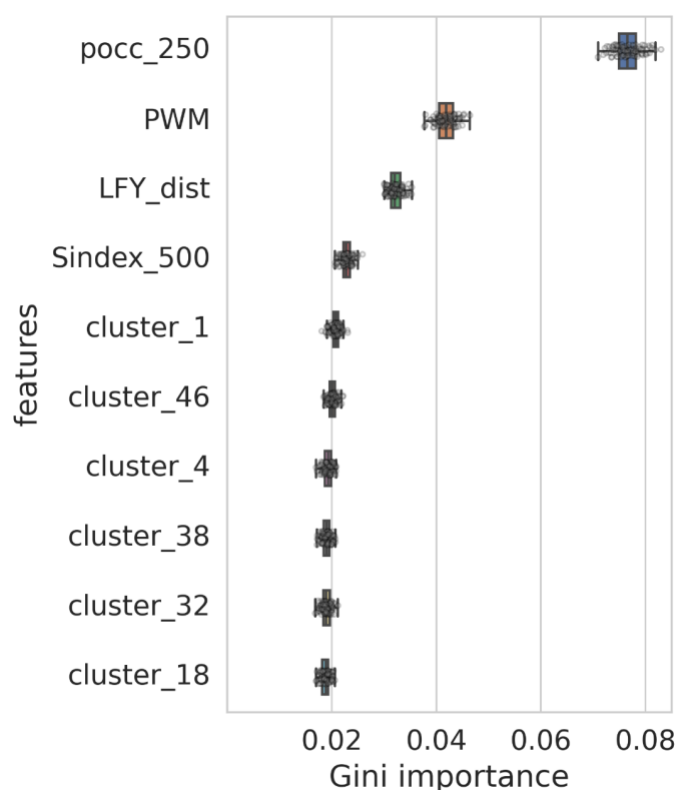
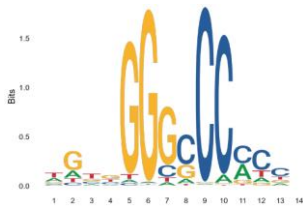
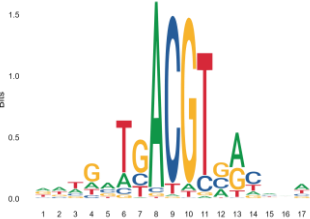
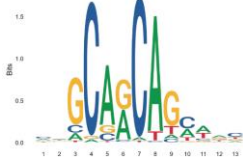
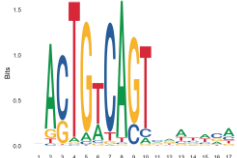


Figure 3.2-3 Most important features included in our Random Forest models. Ten features with the highest Gini importance (highest median value over 100 iterations) in our Random Forest models. 'pocc_250' = POcc score computed within ± 250 bp around LFY sites (see Computing POcc around LFY TFBS, p. 116). PWM = LFY site score obtained with LFY's PWM with dependencies (Moyroud et al., 2011) (see Definition of LFY sites genome-wide, p. 115). LFY_dist = distance between LFY and the next closest LFY site (see Computing co-occurrence and LFY-LFY distances, p. 117). Sindex_500 = non-LFY TFBS diversity within ± 500 bp around LFY sites (see Computing TFBS density and diversity around LFY TFBSs, p. 118). Clusters = TF clusters from JASPAR 2022 (Castro-Mondragon et al., 2022).

Table 3.2-1 Six most important JASPAR 2022 clusters (Castro-Mondragon et al., 2022) shown in Figure 3.2-3A. The table shows correspondence between JASPAR cluster names and the TFs they contain, as well as the TF families they belong to.

JASPAR Cluster name	Cluster motif	TF(s) included in the cluster	Function
cluster_1		25 TFs: TCP2, TCP3, TCP4, TCP5, TCP7, TCP8, TCP13, TCP16, TCP17, TCP19, TCP21, TCP22, TCP23, TCP24, StBRC1, TB1, OJ1581_H09_2, ARALYDRAFT_897773, ARALYDRAFT_496250, ARALYDRAFT_495258, ARALYDRAFT_484486, ARALYDRAFT_493022, Osl_08196, Glyma19g26560_1, Zm00001d038683	TFs, TCP family
cluster_46		CDC5	TF, Myb-like family
cluster_4		41 TFs: NAC047, NAC035, O2, bZIP911, HYH, ABF4, BZIP28, BZIP68, BZIP16, BZIP48, BZIP44, BZIP11, BZIP53, BZIP3, BZIP2, BZIP63, BZIP43, BZIP42, TGA6, TGA5, TGA2, TGA7, TGA3, TGA10, TGA9, TGA4, TGA1, bZIP910, TGA1A, BZIP60, NAC043, NTL9, NAC029, NAC002, NAC055, NAC046, NAC079, NAC019, NAC025, NAC083, NAC018	TFs, bZIP and NAC families
cluster_38		GRF4 RAV1	TFs, GRF and RAV families
cluster_32		GRF6, GRF9, SIZF2, GLYMA-07G038400	TFs, GRF and C2H2 ZF families
cluster_18		AGL55, AGL42, RAV2, TEM1	TFs, MADS and RAV families

Taken together, these results indicate that transcriptionally active LFY sites display a different genomic context from nonfunctional ones, with features related to the presence of other LFY sites having a greater impact on predictions. Although to a lesser extent, differences in non-LFY TFBS diversity and the distance from specific TF groups also contribute to predicting functional LFY sites.

3.2.4 The model can be used to make predictions of LFY functional sites on ‘unknown’ sites

Once a model is trained, it can be used to make predictions on any site for which the same information (genomic context, LFY-related features) is available. This can be done either with a model trained on the entirety of LFY sites, or by following a strategy more similar to cross-validation, where a subset of the data is used to train a new model, repeating the process many times. The advantage of the latter is that each trained model can be used to make predictions on new sites, revealing which ones display high probability of being functional over many iterations. Following this principle, I decided to train the random forest models shown in Figure 3.2-1B (p. 50) on functional and nonfunctional sites, and to make predictions with each one on ‘unknown’ sites, i.e. LFY sites with evidence of either binding or differential expression but which could not be confidently labeled as functional nor as nonfunctional (more details in Using trained Random Forest models to make predictions on ‘unknown’ sites, p. 126).

Among the ‘unknown’ sites with the highest median probability of being functional (over 100 models), a site located on chromosome 3, in the promoter of the *BOP1* gene (chr3:21146118-21146137), looks particularly promising (Figure 3.2-4A). It is the site with the third highest median probability value, and it is surrounded by additional LFY sites (Figure 3.2-4B). This site was labeled as ‘unknown’ because it is bound in ChIP-seq and ampDAP-seq experiments, but there was no evidence of differential expression. When looking at the other four sites with the highest median probability of being functional (Figure 3.2-4A), none is close to a gene for which the link to LFY regulation is as evident as *BOP1* based on existing literature (Figure S3.4-2). In absence of prior evidence of their potential as LFY functional sites, and of experimental validation, I decided not to explore these sites any further, and to focus on the site in the *BOP1* promoter as an example.

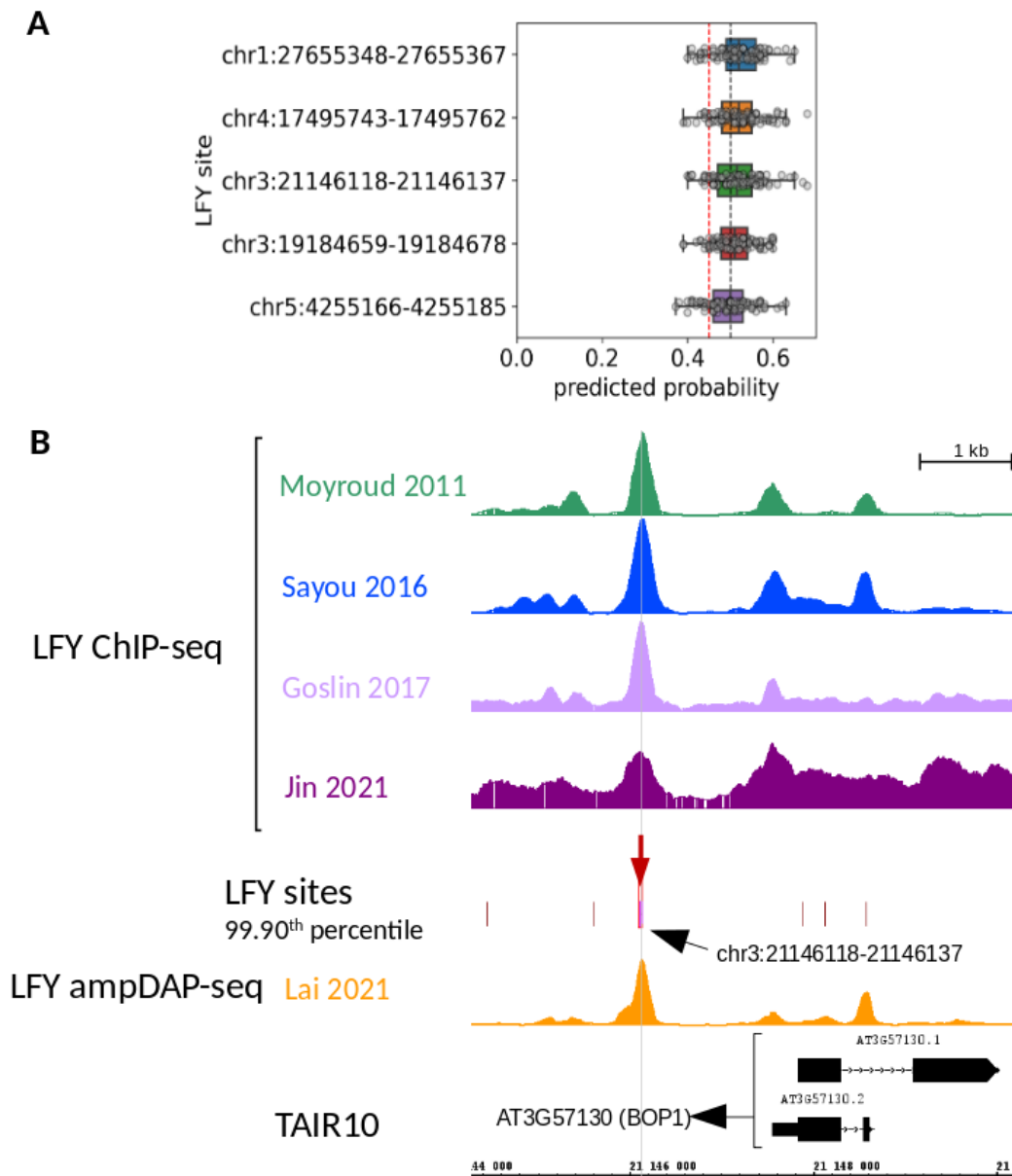


Figure 3.2-4 The model can be used to predict whether ‘unknown’ LFY sites (which could not be confidently labeled as functional nor as nonfunctional) are functional or nonfunctional. A: Five ‘unknown’ LFY sites with the highest median predicted probability of being functional, based on 100 random forest models (see Training and testing Random Forest models, p. 124). Red dashed line marks a predicted probability value of 0.45, while the black dashed line marks 0.5, for a better visual assessment of the probability values displayed. B: Screenshot taken with Integrated Genome Browser (Freese et al., 2016) over the promoter of the BOP1 gene. From top to bottom: LFY binding signal in ChIP-seq experiments from (Goslin et al., 2017; Jin et al., 2021; Moyroud et al., 2011; Sayou et al., 2016); LFY sites computed with LFY’s PWM with dependencies with score above the 99.90th threshold, where the site with the highest median probability shown in panel A is indicated with an arrow and its name; LFY ampDAP-seq signal from (Lai, Blanc-Mathieu, et al., 2021); two isoforms of the BOP1 gene as included in the TAIR10 annotation.

3.3 Discussion

3.3.1 Genomic context carries key regulatory information

We built a classifier that can distinguish between transcriptionally active and inactive LFY sites identified based on their PWM score. Instead of using ROC curves to evaluate and compare models built with different sets of features, I chose to focus on the metric of PR curve AUC because the data are strongly unbalanced, and in such cases ROC curves can lead to high AUC values inflated by an excessive prediction of the majority class. While the model itself, with an average PR AUC of 0.37, is far from being perfect (AUC = 1), it still suggests that the features we included can indeed help distinguish the two sites' categories.

Our results show that co-occurrence, encoded as the distance between each LFY site and the closest site of each of over 40 groups of TFs, is more informative than state-of-the-art LFY DNA-binding models characterizing site quality (PWM score) or LFY occupancy in the ± 250 bp surrounding each reference LFY site (POcc) (Figure 3.2-1, p. 50). Nevertheless, combining context information and LFY models leads to the best results, indicating that, ultimately, LFY binding *in vivo* happens on regulatory regions with particular characteristics.

Such characteristics are revealed by the importance given by the model to different features: the most important features are all related to LFY binding sites, with POcc having an especially strong impact on predictions. The importance of POcc over PWM highlights the fact that, beyond binding affinity, the biochemical properties of LFY and the presence of multiple LFY sites in the same region are signals of LFY active sites. This is in line with previous reports of the importance of the N-terminal SAM domain of the protein, which is crucial for cooperative binding of LFY oligomers on regulatory regions (Sayou et al., 2016). Using the same approach with other TFs that do not oligomerize will be crucial to determine the importance of this feature beyond LFY. Furthermore, it would be interesting to assess the importance of cooperative binding by looking at whether functional and nonfunctional LFY sites are more or less likely to lose binding signal when LFY's cooperativity is disrupted, as it happens in a previously published mutated version of LFY (LFY_{TERE}, where T75E, R112E substitutions alter the protein's oligomerization properties) (Sayou et al., 2016).

In addition to LFY-related features, TFBS diversity, which we encoded by adapting Shannon's entropy measure (Shannon, 1948), also seems to be important for predictions. Previous

reports using synthetic promoters had shown the importance of diverse regulatory regions, containing TFBSs belonging to different TFs, for reporter gene expression, as opposed to a suite of the same amount of sites from the same TF (Singh et al., 2021). It is also possible that this measure simply helps setting apart regulatory regions from non-regulatory ones, although total amount of TFBSs around LFY sites (TFBS density), which is generally higher in regulatory regions, was less important than diversity.

Distance between LFY sites and TFBSs belonging to some key TF clusters also seems to have an impact on predictions (Figure 3.2-3, p. 55). Cluster_1, which contains TCP TFs, is among the most important ones. While TCPs are involved in a variety of processes in Arabidopsis, TCP5, TCP13 and TCP17 have been shown to interact with the *AP1* promoter, a major LFY target (D. Li et al., 2019). As cluster_1 is the cluster with the highest median importance for predictions, LFY could be sharing other targets with TCP proteins.

Next, cluster_46 only contains one TF, CELL DIVISION CYCLE 5 (CDC5), a Myb-like TF. CDC5 is expressed in the shoot apical meristem and in inflorescences, and it has been shown to be involved in development and bacterial infection response in Arabidopsis (Hirayama & Shinozaki, 1996; Lin et al., 2007; Palma et al., 2007; S. Zhang et al., 2013). Moreover, *cdc5* mutants display phenotypic defects such as delayed flowering and sterility (Lin et al., 2007). While the importance of the distance between LFY and CDC5 TFBSs does not necessarily imply that the two proteins interact, the evidence reported above supports the fact that the two proteins are coexpressed, and that they could be regulating some common targets, leading to their sites being close to each other.

As cluster_4, cluster_38, cluster_32 and cluster_18 represent a variety of proteins and TF families for which a direct implication with LFY is not evident, I do not want to over-speculate on their importance in the context of LFY binding. In order to truly determine the role of the other clusters as potential TFs co-occurring with LFY, additional analyses and experimental validation will be required. Techniques such as TARGET, transiently expressing TFs in Arabidopsis protoplasts, could be instrumental to study the involvement of such clusters in co-regulating genes with LFY (Bargmann et al., 2013; Brooks et al., 2023).

More broadly, I explicitly chose to focus on sites where LFY binds by itself and without strict requirement for cofactors, as I focused on LFY sites bound in ChIP-seq *and* ampDAP-seq

experiments. This means that clusters that are important for predictions are important because of the distance of their TFBSs from LFY's, but not necessarily as direct interactors of LFY. If the objective were to look for potential interactors, for example, it would be possible to modify the classification labels in order to look for LFY sites that are bound *in vivo* and not *in vitro*, which could require binding partners and/or a chromatin context that is missing in ampDAP-seq experiments. While I have not tried this strategy so far, it could be interesting for exploratory analyses with LFY as for other TFs.

3.3.2 Several possible reasons to why conservation of LFY sites does not help predictions

Functionally important elements in regulatory regions have been reported to be evolutionarily conserved, and this information can be used to infer regulation (Vandepoele et al., 2006). We hypothesized that including evolutionary conservation of LFY sites in a predictive model would positively impact predictions as sites with a functional role in transcriptional regulation would tend to be more conserved than nonfunctional sites randomly arising in the genome. However, when I included the average conservation score (PhyloP and PhastCons) of LFY sites in the model, which already contained genomic context as well LFY PWM score and POcc, it did not have any significant effect on PR AUC, and even slightly decreased its median value (Figure 3.2-2, p. 53).

We initially thought that the reason behind conservation not improving our classification could be that the evolutionary interval included to compute conservation scores was too broad, and that some transcriptionally important LFY sites were lost. Therefore, I investigated whether LFY sites were overlapping with conserved regions computed at increasing evolutionary distances. I first excluded LFY sites found in coding regions because three out of four datasets were specifically computed on noncoding regions. My results show that functional LFY sites are consistently more enriched in CNSs than nonfunctional ones, and that the ratio between the fraction of functional sites overlapping with CNSs and the proportion found for nonfunctional sites decreases as the evolutionary distance increases (Figure 3.2-2C and D, p. 53).

However, when I used a dataset including conserved regions in coding and noncoding sequences (Tian et al., 2020), the proportion of LFY sites that overlap with conserved regions was very similar between functional and nonfunctional sites (Figure S3.4-1D, p. 65). It is important to note that this dataset was published along with the PhastCons and PhyloP conservation scores that I included in the model (Tian et al., 2020). This difference could be amplified by the differences in sequence type distributions between the three classes of LFY sites (Figure S3.4-1B, p. 65). Therefore, it is possible that high conservation observed for sites found within coding regions cannot be distinguished from high conservation in noncoding regions, although the latter could be an indicator of functional importance of regulatory regions. While I did not try to build a model including conservation as a binary feature representing overlap (or lack thereof) of each LFY site with CNSs, it would be interesting to know whether this kind of information has any impacts on predictions, and whether the impact changes when using CNS datasets at increasing evolutionary distances. This approach, based on the results shown in Figure 3.2-2C (p. 53), has the potential to reveal a higher influence of conservation on predictions, but it should be exclusively applied to noncoding LFY sites.

It is important to note that both strategies, the averaged PhyloP and PhastCons conservation per site and the overlap of at least a portion of LFY sites with conserved regions, have a caveat: neither of them take into account the importance of the conserved nucleotides (or portions of the LFY TFBS) for binding. As not all positions are equally important for binding, conservation of key positions could be more important than conservation of more variable positions within the TFBS. My previous attempts at using a weighted average to encode overall conservation of LFY sites, where the information content per position represents the weight, did not improve results, although I did not try to do the same when looking at overlap of LFY sites with CNSs and this remains to be tested.

Another possibility that could be explored is to switch the focus from sequence conservation to the conservation of regulatory modules, which has been shown to be important for gene expression conservation over broad evolutionary distances (Nitta et al., 2015; Ravel et al., 2014; Taher et al., 2011; Wong et al., 2020). As the model presented in this chapter is built to include features describing the genomic context of functional and nonfunctional LFY sites, the focus for encoding conservation could switch from single LFY sites towards the

conservation of the regulators in that context. This would require more complex comparative genomics approaches, but it holds great potential for functional site predictions.

Bearing in mind all the considerations above, it is still possible that a low impact of site conservation on predictions is due to the fact that, while LFY's sequence and specificity are highly conserved (Sayou et al., 2014), its targets may not be as conserved, and neither its sites. Previous reports of *cis*-regulation being rather restricted support the possibility that *cis*-regulatory targets could be scarcely conserved (Ballester et al., 2014; Tu et al., 2022). A particularly relevant example in plants is a study where only about 13% of Arabidopsis GOLDEN2-LIKE (GLK) TF targets were also found to be targeted by GLK orthologs in tomato, tobacco, maize and rice (Tu et al., 2022). From this point of view, despite evidence of LFY still targeting some important genes such as *AP1* and *AG* outside of Brassicaceae (Minguet et al., 2015; Moyroud et al., 2011), it is possible that its other targets, which have not been studied in such detail, may not be conserved. *AP3* regulation, which requires both LFY and its cofactor UFO, is also conserved outside of Brassicaceae (Ingram & Coena, 1995.; S. McKim & Hay, 2010; Rieu et al., 2023; Souer et al., 2008). LFY has also been shown to be involved in additional developmental processes in other species than Arabidopsis (Moyroud et al., 2009, 2010), suggesting that the gene regulatory networks where it is involved may have changed, and LFY target sites with them. Nevertheless, at the moment, we do not have enough evidence to establish whether this is the reason behind the poor results we obtained when adding LFY sites conservation scores to our model.

3.3.3 New potential targets of LFY revealed by the model

I used the model presented in Figure 3.2-1B to predict whether 'unknown' LFY sites, i.e. sites that only partially met the requirements to be labeled as functional, were transcriptionally active or not. My results indicate that an 'unknown' LFY site likely to be functional is found in the promoter of *BOP1*, a TF that has been shown to negatively regulate LFY protein (Chahtane et al., 2018). Like LFY's, *BOP1* expression is detected since the floral anlagen, and the BOP1 protein is detected since stage 5/6 and at the base of developing organs, which overlaps with LFY's expression domain (Karim et al., 2009; S. M. McKim et al., 2008; Weigel

et al., 1992; Yamaguchi et al., 2016). The site within the *BOP1* promoter predicted to be functional is bound in ChIP-seq and ampDAP-seq experiments, but the lack of *BOP1* differential expression led to its ‘unknown’ label. Based on the evidence presented above, LFY could be targeting the *BOP1* promoter. Dual luciferase reporter assays in Arabidopsis protoplasts with the full *BOP1* promoter sequence would be a good start to determine whether LFY can induce *BOP1* expression. Furthermore, sequentially mutating the LFY site predicted to be functional as well as the other LFY sites on the *BOP1* promoter would help determine the importance of that site in particular, as predicted by our model. None of the other sites with the highest probability of being functional (Figure 3.2-4A, p. 58) maps close to genes with apparent link to flowering or show enough compelling prior evidence of being targeted by LFY (Figure S3.4-2, p. 66).

In our approach, we focused on LFY sites found within or around genes, up to 3 kb upstream of the TSS and 1 kb downstream of the TTS. However, trained models could be used to make predictions on LFY sites found in intergenic regions, i.e. even farther away from genes, following the same approach used for ‘unknown’ LFY sites. While in Arabidopsis the mean size of intergenic regions is about 1 kb (Lamesch et al., 2012), gene density can vary and it could reveal transcriptional activity at longer distances from genes.

Finally, as our approach relies on learning the genomic context of transcriptionally active sites rather than direct TF-target links, a model trained on Arabidopsis data could be used to make predictions on TFBSs in other species. This could reveal if and to what extent the cis-regulatory code that applies to a TF is, itself, conserved in other species. Binding and expression data in five plant species as published by (Tu et al., 2022) for GLK TFs could be instrumental to further develop and validate this approach.

3.4 Supplementary figures

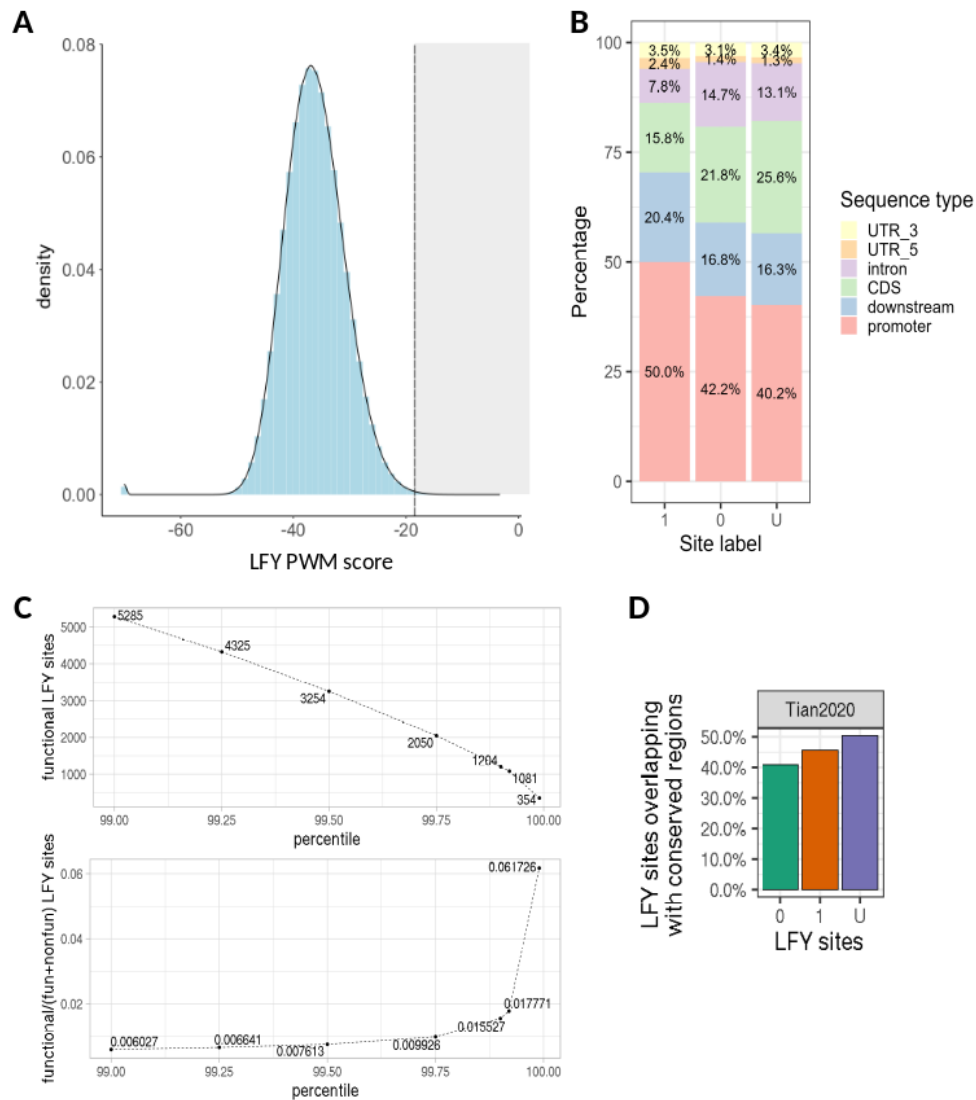


Figure S3.4-1 Supplementary information about LFY sites genome-wide. **A:** Distribution of LFY PWM scores as calculated on the whole *Arabidopsis* genome with LFY's PWM with dependencies. The vertical line represents the 99.90th percentile threshold chosen for subsequent analyses; grey shaded part represents the LFY sites with a PWM score above the 99.90th threshold that were used in subsequent analyses. More details in Definition of LFY sites genome-wide, p. 115. **B:** Type of sequence to which functional (1), nonfunctional (0) and 'unknown' (U) LFY sites map on the *Arabidopsis* genome. UTR_5 = 5' UTR, UTR_3 = 3' UTR, CDS = coding sequence, downstream = 1 kb downstream of the 3' UTR, promoter = 3 kb upstream of the 5' UTR. LFY sites mapping at multiple sequences types (e.g. promoter and CDS) are shown multiple times. **C:** Total number of functional LFY sites (top panel) and ratio of functional/total functional + nonfunctional sites (bottom panel) at increasing PWM score percentile thresholds. More details in Definition of LFY sites genome-wide, p. 115. **D:** Proportion of LFY sites at least partially overlapping with conserved regions, LFY sites mapping to coding and noncoding regions confounded. The analysis was conducted as in Figure 3.2-2C (p. 53), but this time with all LFY sites in coding and noncoding regions confounded. 1 = functional LFY sites; 0 = nonfunctional LFY sites; U = 'unknown' LFY sites. Haudry2013 = CNSs computed by (Haudry et al., 2013); VV2014 = CNSs computed by (Velde et al., 2014); VV2016 = CNSs computed by (Velde et al., 2016); Tian2020 = CNSs computed by (Tian et al., 2020). See Calculating CNS enrichment at LFY sites, p. 123, for more details.

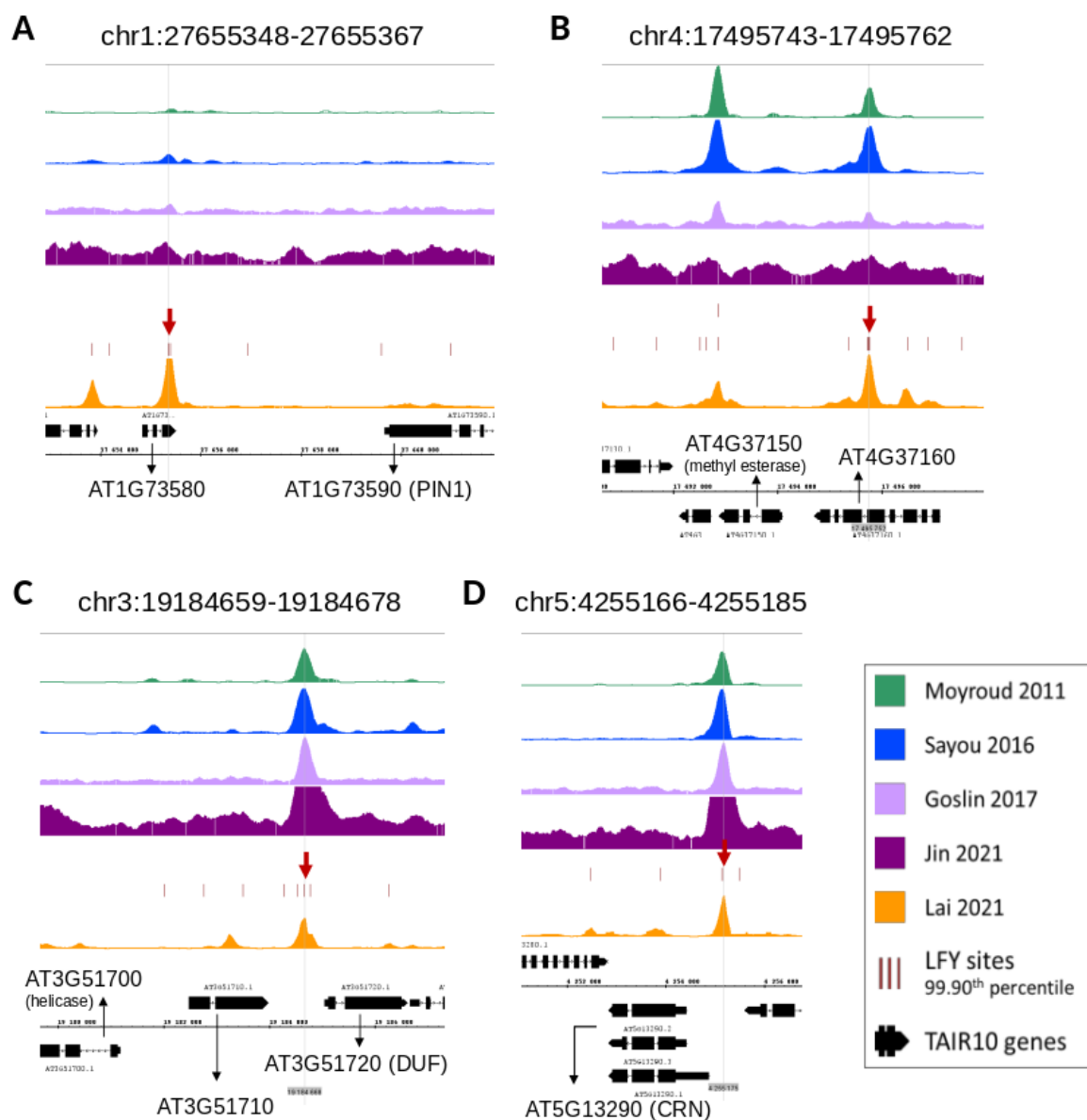


Figure S3.4-2 Snapshots of LFY sites labeled as ‘unknown’ and predicted to be functional in Figure 3.2-4A, as indicated by their name at the top of each panel. Order and colors correspond to those in Figure 3.2-4A and are indicated in the legend on the right.

4 Chapter 2: The LFY-UFO complex regulates distinct genes from LFY

4.1 Introduction

4.1.1 LFY and UFO are involved in petal and stamen development

As explained in the introduction of this manuscript, *Arabidopsis* flowers are made of four concentric whorls of organs: sepals, petals, stamens and carpels. The correct formation of these organs relies on effective spatiotemporal control of homeotic gene expression. Petal and stamen development is regulated by LFY and UFO through the activation of *AP3* (I. Lee et al., 1997).

This chapter will focus on LFY and UFO, and on their role in petal and stamen development. First, I will explain the proposed role of the LFY-UFO interaction based on previous findings, then I will explain the main findings of our article, published early in 2023, highlighting my contribution. I will include the article in this manuscript and, finally, I will show some additional results and discuss the possibility that LFY and UFO have a broader role in plant development.

4.1.2 UFO has been proposed to interact with LFY to promote its ubiquitination and degradation

UFO is an F-box protein that is part of an SCF E3 ubiquitin ligase complex, and it directly interacts with LFY to regulate *AP3* expression (Chae et al., 2008). Unlike LFY, which is expressed throughout the floral meristem since the earliest stages of floral development (Weigel et al., 1992), UFO is expressed in the peripheral zone of the shoot apical and inflorescence meristems (Reddy, 2008). Therefore, the interaction between LFY and UFO has been proposed to allow the spatiotemporal control of *AP3* expression in the appropriate domains for the development of petals and stamens (Parcy et al., 1998).

Moreover, LFY can also be ubiquitinated in a UFO-dependent manner in inflorescences, as its ubiquitination level is lower in a *ufo-2* background (Chae et al., 2008). This information pointed towards UFO interacting with LFY to promote its ubiquitination and subsequent proteasome-dependent degradation. An additional hint in this direction was provided by

Risseeuw et al., who showed that overexpression of UFO without its F-box domain in a WT background resulted in a phenotype similar to *ufo* weak alleles, suggesting that the involvement of UFO in protein ubiquitination and degradation constitutes an important part in its flower development role (Risseeuw et al., 2013).

Despite this body of evidence, the exact molecular mechanism remained unclear. Is UFO recruited to *AP3* to promote LFY's degradation, or is it there to contribute to LFY's transcriptional function? Does the presence of UFO alter LFY's properties allowing its binding at the *AP3* promoter? We addressed these questions through a combination of molecular, computational and structural approaches.

4.1.3 LFY and UFO form a transcriptional complex that recognizes new genomic regions

Our results suggest that UFO forms a transcriptional complex with LFY to control the expression of developmental genes such as *AP3* (Figure 4.1-1). Interestingly, the F-box domain of UFO is partially dispensable for its transcriptional role, unlike what was previously suggested (Risseeuw et al., 2013), indicating that this action is independent of its involvement in an SCF ubiquitination complex. It is likely that the weak *ufo* phenotype observed in previous reports of plants expressing UFO without its F-box domain was due to protein folding issues of the truncated protein.

Instead, we found that the LFY-UFO interaction changes LFY's binding specificity to what we named a LFY-UFO Binding Site (LUBS), which is a variation of a LFY canonical motif with the addition of a UFO-recruiting motif at the 5' end. I detected this motif thanks to a genome-wide LFY-UFO binding assay *in vitro* (ampDAP-seq), performed by P. Rieu, after designing a procedure to retrieve complex-specific regions as opposed to regions bound by LFY alone in another ampDAP-seq experiment. Crucially, the UFO-recruiting motif at the 5' end of LFY's canonical motif was detected in LFY *in vivo* binding data. These findings were supported by evidence of the LFY-UFO complex binding to a LUBS *in vitro* in electrophoretic mobility shift assays, and by a strong reduction in *AP3* expression upon mutation of the strongest LUBS sites in the *AP3* promoter.

LFY-UFO and LFY targeting different genomic regions allows an additional layer of transcriptional control, by restricting the expression of a subset of genes to the tissues where both LFY and UFO proteins are present. Indeed, a LUBS is also present in a set of genomic regions that LFY can bind with UFO but not on its own, and that are distinct from LFY's canonical targets.

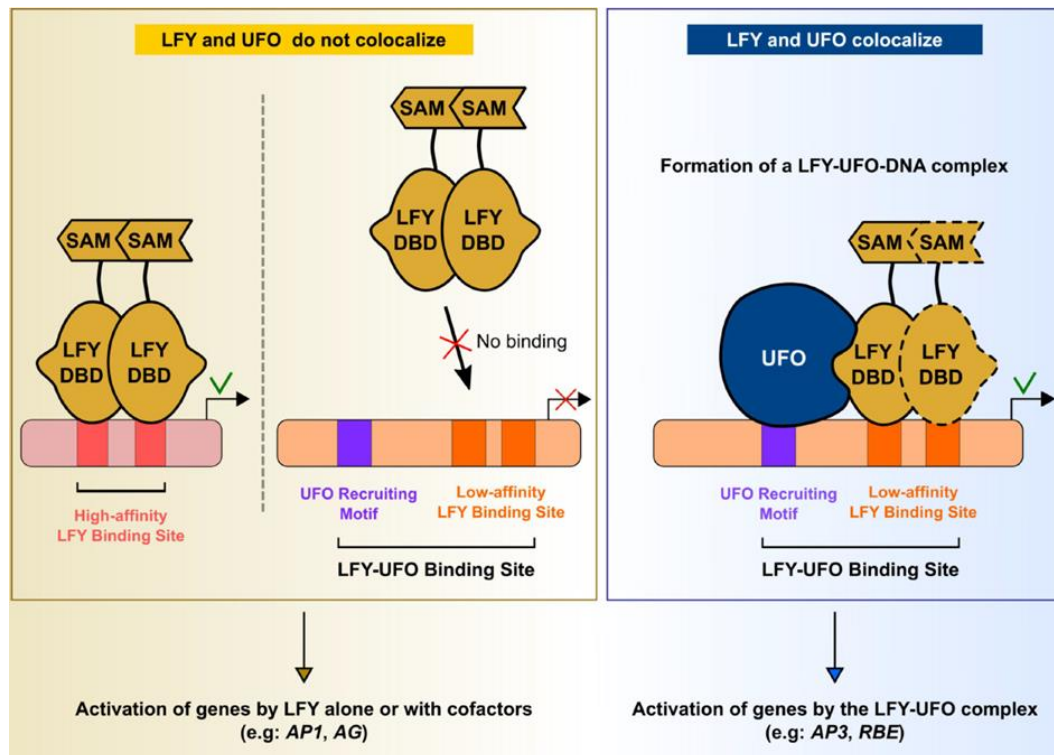


Figure 4.1-1 LFY and UFO form a transcriptional complex to regulate flower developmental genes. Credits: Philippe Rieu.

Solving the cryo-EM structure of the LFY-UFO complex revealed that UFO directly participates in the binding, and that the bound DNA is slightly bent. Finally, yeast-two-hybrid assays using LFY proteins from several species at increasing evolutionary distances and UFO from *Arabidopsis* suggest this interaction to be conserved in angiosperms, gymnosperms and ferns, but not in mosses and algae.

I contributed to the computational side of this project, by analyzing LFY and LFY-UFO ampDAP-seq data and comparing the read coverage (i.e. signal strength) of the regions bound in the two experiments. I also identified the new LUBS in LFY-UFO-enriched regions, and I analyzed public microarray data to identify new potential targets of the complex. I assembled bioinformatics-related figures, as well as the materials and methods and a first

draft of the main text relative to these parts. I received help from Jérémy Lucas on the analysis of the quality of LFY sites on LFY-UFO- vs LFY-specific regions, and Romain Blanc-Mathieu was responsible for finding the UFO-recruiting motif in LFY ChIP-seq data.

The full paper will be included in the next section. It will be followed by a section showing some additional results and discussing the role of the LFY-UFO complex in plant development in the light of recent findings.

4.2 [Article](#)

The F-box protein UFO controls flower development by redirecting the master transcription factor LEAFY to new *cis*-elements

Received: 18 August 2022

Accepted: 20 December 2022

Published online: 2 February 2023



Philippe Rieu¹, Laura Turchi^{1,2}, Emmanuel Thévenon¹, Eleftherios Zarkadas^{3,4}, Max Nanao⁵, Hicham Chahtane^{1,6}, Gabrielle Tichtinsky¹, Jérémy Lucas¹, Romain Blanc-Mathieu¹, Chloe Zubieta¹, Guy Schoehn^{1,3} & François Parcy¹✉

In angiosperms, flower development requires the combined action of the transcription factor LEAFY (LFY) and the ubiquitin ligase adaptor F-box protein, UNUSUAL FLORAL ORGANS (UFO), but the molecular mechanism underlying this synergy has remained unknown. Here we show in transient assays and stable transgenic plants that the connection to ubiquitination pathways suggested by the UFO F-box domain is mostly dispensable. On the basis of biochemical and genome-wide studies, we establish that UFO instead acts by forming an active transcriptional complex with LFY at newly discovered regulatory elements. Structural characterization of the LFY–UFO–DNA complex by cryo-electron microscopy further demonstrates that UFO performs this function by directly interacting with both LFY and DNA. Finally, we propose that this complex might have a deep evolutionary origin, largely predating flowering plants. This work reveals a unique mechanism of an F-box protein directly modulating the DNA binding specificity of a master transcription factor.

The formation of flowers is key to the reproductive success of angiosperms. Flowers are made of four types of organs (sepals, petals, stamens and carpels) arranged in concentric whorls. The patterning of flower meristems requires the localized induction of the ABCE floral homeotic genes that determine specific floral organ identities. In *Arabidopsis thaliana*, this developmental step is largely controlled by the master transcription factor (TF) LEAFY (LFY) that activates the ABCE genes^{1,2}. LFY directly activates the A class gene *APETALA1* (*API*) uniformly in the early flower meristem^{3,4}, while activations of B and C

genes are local and require the activity of cofactors. For instance, LFY regulates the C class gene *AGAMOUS* (*AG*) in conjunction with the TF WUSCHEL to specify third whorl (stamen) and fourth whorl (carpel) identities⁵. The activation of the B class gene *APETALA3* (*AP3*), necessary to specify the identity of the second (petal) and third whorls of the flower, requires the combined activity of LFY and the spatially delineated cofactor UNUSUAL FLORAL ORGANS (UFO)^{6–8}. In *Arabidopsis*, the main function of LFY and UFO is to activate *AP3* (ref. ⁹), but in numerous species (such as rice, wheat, tomato and petunia), their joint role

¹Laboratoire Physiologie Cellulaire et Végétale, IRIG-DBSCI-LPCV, Université Grenoble Alpes, CEA, CNRS, INRAE, Grenoble, France. ²Translational Innovation in Medicine and Complexity, Université Grenoble Alpes, CNRS, Grenoble, France. ³IBS, Université Grenoble Alpes, CNRS, CEA, Grenoble, France. ⁴EMBL, ISBG, Université Grenoble Alpes, CNRS, CEA, Grenoble, France. ⁵Structural Biology Group, European Synchrotron Radiation Facility, Grenoble, France. ⁶Present address: Green Mission Pierre Fabre, Conservatoire Botanique Pierre Fabre, Institut de Recherche Pierre Fabre, Soual, France. ✉e-mail: francois.parcy@cea.fr

goes well beyond B gene activation and is key to floral meristem and inflorescence development^{10–13}.

At the molecular level, little is known on the nature of LFY–UFO synergy. Unlike most floral regulators, *UFO* encodes not for a TF but for an F-box protein, one of the first to be described in plants^{14–16}. *UFO* is part of a SKP1–Cullin1–F-box (SCF) E3 ubiquitin ligase complex through the interaction of its F-box domain with ARABIDOPSIS SKP1-LIKE (ASK) proteins^{15,17}. In addition, its predicted carboxy-terminal Kelch-type β -propeller domain physically interacts with LFY DNA Binding Domain (DBD)¹⁸. As the control of TF activity through proteolytic and non-proteolytic ubiquitination is a well-described mechanism¹⁹, it has been suggested that LFY is targeted for ubiquitination and possibly degradation by the SCF^{UFO} complex. Other data have shown that adding a repression or activation domain to *UFO* changes its activity and that *UFO* is recruited at the *AP3* promoter in a LFY-dependent manner, suggesting a more direct role of *UFO* in gene regulation^{18,20}. However, direct evidence explaining how *UFO* regulates a specific subset of LFY targets was still missing, and the molecular mechanism underlying LFY–UFO synergistic action remained elusive.

Here we show that the connection of *UFO* to the SCF complex is largely dispensable for this protein's activity and that an important role of *UFO* is to form a transcriptional complex with LFY at genomic sites devoid of canonical high-affinity LFY binding sites (LFYBS). Our study presents a unique mechanism by which an F-box protein acts as an integral part of a transcriptional complex.

The *UFO* F-box domain is partially dispensable for its floral role

A dual luciferase reporter assay (DLRA) in *Arabidopsis* protoplasts was used to study floral promoter activation by LFY and *UFO*. We used promoter versions known to allow full complementation of mutants or to be able to recapitulate a wild-type (WT) expression pattern (Methods). We found that the *AP3* promoter (*pAP3*) was more strongly activated when LFY (or LFY–VP16, a fusion of LFY with the VP16 activation domain) was co-expressed with *UFO* (or *UFO*–VP16) than by either effector alone (Fig. 1a,e). Similar results were obtained with the promoter of *RABBIT EARS* (*RBE*), another *UFO* target (Fig. 1b)²¹. We also analysed the promoters of *API* (*pAPI*) and *AG* (*pAG*), two LFY targets regulated independently of *UFO*^{3,4,22} that are required for organ identity of the first and second (*API*) or third and fourth (*AG*) floral whorls. We found that their activation by LFY and LFY–VP16 was insensitive to *UFO* (Fig. 1c,d). Thus, the protoplast assay accurately reproduced several floral promoter activation patterns.

We next investigated the involvement of an SCF^{UFO}-dependent ubiquitination pathway in *pAP3* activation by LFY–UFO. We found that, when co-expressed with LFY, amino-terminally truncated *UFO* versions lacking the F-box domain (*UFO* Δ Fbox and *UFO* Δ Fbox–VP16) activated *pAP3* similarly to the full-length (FL) *UFO* (Fig. 1e). The connection of *UFO* to an SCF complex thus appears dispensable for *pAP3* activation in transient protoplast assays. The previously reported inactivity of *UFO* with an internal deletion of its F-box probably reflects the poor folding of this protein variant rather than the functional importance of the F-box domain (Extended Data Fig. 1a–c)²⁰.

We also constitutively expressed tagged versions of *UFO* and *UFO* Δ Fbox in *Arabidopsis*. Irrespective of the presence of the F-box, plants displaying a detectable *UFO* or *UFO* Δ Fbox expression (Extended Data Fig. 1d) showed a typical *UFO* gain-of-function phenotype (Fig. 1f,g). In addition, both *UFO* versions complemented the strong *ufo-1* mutant and induced gain-of-function phenotypes (Fig. 1h and Extended Data Fig. 1e,f)⁸. Still, minor defects (such as some missing or misshapen petals and disorganized flowers) were specifically observed in the absence of the F-box, suggesting that this conserved domain might be important for a subset of *UFO* functions (Fig. 1h and Extended Data Fig. 1g). Overall, *UFO* and *UFO* Δ Fbox have very similar activities, showing that the role of the F-box domain is largely dispensable

and that a ubiquitination-independent mechanism determines the LFY–UFO synergy.

LFY and *UFO* form a transcriptional complex on a new DNA motif

Protoplast assays established that the *AP3* and *RBE* promoter sequences contain the information that dictates their specific activation by LFY–UFO. Several regulatory regions driving *AP3* regulation in early floral meristem have been identified, including the distal and proximal early elements (DEE and PEE; Fig. 2a)^{23,24}. The DEE contains a predicted canonical LFYBS, but in protoplasts, like in plants²⁴, this site is not sufficient to explain *pAP3* activation (Extended Data Fig. 2). By systematically testing *AP3* promoter variants in the transient assay, we identified a 20-base-pair (bp) DNA element around the PEE important for LFY–UFO-dependent activation but devoid of canonical LFYBS (Fig. 2b and Extended Data Fig. 3a–c). We investigated the possibility that LFY and *UFO* form a complex on this DNA element using electrophoretic mobility shift assays (EMSAs). For this, we mixed either recombinant LFY–DBD (the LFY domain interacting with *UFO*)¹⁸ or in-vitro-produced FL LFY with a reconstituted ASK1–UFO complex. None of the proteins bound the DNA probe alone, but a shift was observed when LFY–DBD or FL LFY was mixed with ASK1–UFO (Fig. 2c). Thus, a presumptive ASK1–UFO–LFY complex was formed on a *pAP3* DNA element (hereafter named LFY–UFO Binding Site 0 (LUBS0)) that each partner did not bind on its own. We did note that *UFO* had a weak affinity for DNA, as ASK1–UFO shifted the DNA probe when we performed EMSAs with low competitor DNA concentrations (Extended Data Fig. 3d). EMSAs performed with LUBS0 mutated at various bases provided evidence that the formation of the complex is sequence-specific and suggests a bipartite DNA motif (Extended Data Fig. 3f).

To identify all genome regions possibly targeted by the ASK1–UFO–LFY complex, we performed amplified DNA affinity purification sequencing (ampDAP-seq) with a reconstituted ASK1–UFO–LFY complex (Extended Data Fig. 4a,b). We identified numerous genomic regions where LFY binding was strongly enhanced by the presence of ASK1–UFO. For each bound region, we computed the ratio (the coverage fold change (CFC)) between the coverage of peaks in the presence or absence of ASK1–UFO (Fig. 2d). Searches for enriched DNA motifs in the 600 regions with the highest CFC (>4.7) identified two bipartite motifs made of a 6-bp RRNRCA (N indicates A/C/G/T; R indicates A/G) sequence, four bases of variable sequence and either a monomeric or a dimeric site resembling canonical LFYBS but with more variability (Fig. 2e). Consistent with the presence of a sequence resembling LFYBS, we found that *pAP3* activation in protoplasts required the LFY amino acid residues involved in binding to canonical LFYBS (Extended Data Fig. 4c,d).

We named the identified motifs mLUBS and dLUBS for monomeric and dimeric LFY–UFO Binding Sites, respectively (Fig. 2e). Since it is observed specifically with ASK1–UFO, the RRNRCA element was named the *UFO* Recruiting Motif (URM). Position weight matrices (PWMs) for dLUBS and to a lesser extent mLUBS outperformed the canonical PWM for LFY, showing that they reliably predicted the binding of ASK1–UFO–LFY (Extended Data Fig. 4e). The LFYBS present within the LUBS of high-CFC regions tended to have a lower predicted affinity than those present in regions bound by LFY alone (Extended Data Fig. 4f), explaining why LFY binding to those sequences occurs only with *UFO* and the URM. Remarkably, we also identified the URM de novo from published LFY chromatin immunoprecipitation sequencing (ChIP-seq) data (Extended Data Fig. 4g)²⁵. Moreover, we found that the LFY ChIP-seq performed in inflorescences²⁵ correlates better with the ASK1–UFO–LFY ampDAP-seq than with the LFY ampDAP-seq (Spearman rank correlation of 0.481 versus 0.338 for the first 1,000 ChIP-seq peaks), strongly suggesting that many regions are bound in vivo by *UFO* (see examples of such regions in Extended Data Fig. 7b,c).

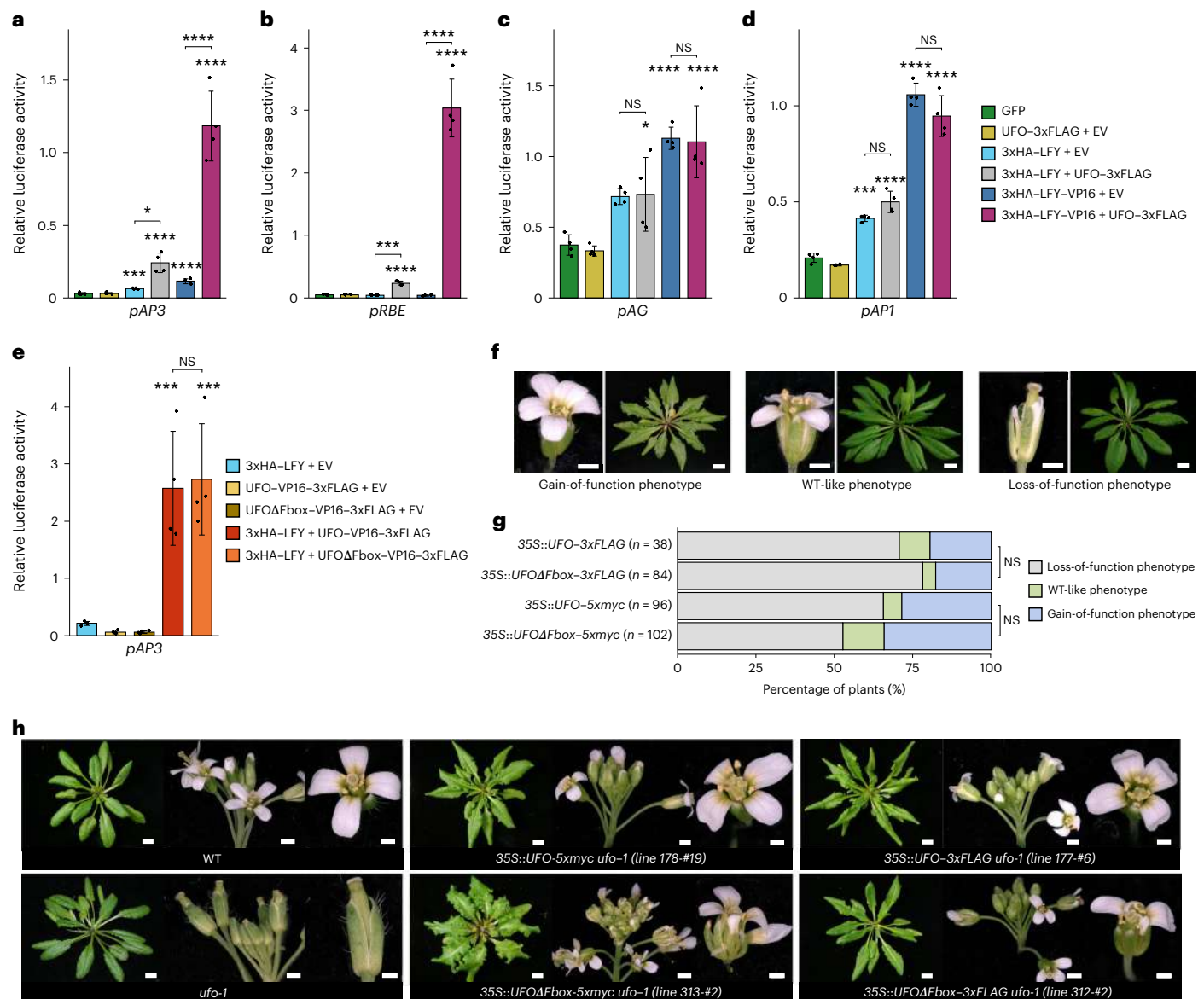


Fig. 1 | UFO action is largely independent of its F-box domain. **a–e**, Promoter activation in *Arabidopsis* protoplasts, with the indicated effectors (right) and promoters (below each graph). EV, empty vector. The data are mean $\pm$ s.d. ($n = 4$ biological replicates). We used one-way analysis of variance (ANOVA) with Tukey's multiple comparisons test (**a, c–e**) or Welch's ANOVA with Games–Howell post hoc tests (**b**). For **a** and **b**, the ANOVAs were performed on log-transformed data (Methods). The asterisks represent a significant statistical difference compared with GFP (**a–d**) or 3xHA-LFY + EV (**e**), non-significant (NS) otherwise. Other comparisons are indicated with brackets. NS, $P > 0.05$; * $P < 0.05$; *** $P < 0.001$;

**** $P < 0.0001$. **f**, Representative pictures of the different phenotypic classes obtained in the *T*₁ population of the indicated transgenic plants (scale bars, 1 mm for flowers and 1 cm for rosettes). **g**, Distribution of *T*₁ plants in phenotypic classes as described in **f**. The distribution of 35S::UFO and 35S::UFOΔFbox lines within phenotypic classes is not significantly different (χ^2 tests; NS, $P > 0.05$). *n*, number of independent lines. **h**, *ufo-1* complementation assay by the 35S::UFO and 35S::UFOΔFbox transgenes. Rosettes (scale bars, 1 cm), inflorescences (scale bars, 1 mm) and flowers (scale bars, 0.5 mm) are shown. Source data are available in Supplementary Data 4.

The ampDAP-seq findings were validated by EMSAs with DNA probes corresponding to optimal mLUBS and dLUBS motifs (Fig. 2f and Extended Data Fig. 4h). We observed a complex of slower mobility with dLUBS than with mLUBS, consistent with the presence of two LFY molecules on dLUBS. ASK1–UFO also supershifted LFY bound to canonical LFYBS from *pAP1* and *pAP3* DEE (Extended Data Fig. 4i), sometimes (but not systematically) increasing apparent LFY binding.

LUBS are functional regulatory elements

Examination of the *pAP3* genomic region in ASK1–UFO–LFY ampDAP-seq revealed a peak that is absent in the experiment performed with LFY alone (Fig. 3a). This peak is roughly located on the PEE and is

consistent with LFY ChIP-seq peaks^{25,26}. We searched for LUBS under this peak, and, to our surprise, we identified several sites predicted to be better than LUBS0 (Fig. 3a). In EMSAs, the two highest-scoring sites, LUBS1 and LUBS2, were specifically bound by LFY in the presence of ASK1–UFO (Fig. 3b and Extended Data Fig. 5a). EMSAs performed with a LFY mutant version affected in its ability to dimerize further confirmed the stoichiometry of LFY–UFO complexes on LUBS1 and LUBS2 (Extended Data Fig. 5b). A similar binding was also observed when combining LFY and UFOΔFbox (Extended Data Fig. 5c,d), consistent with the F-box being facultative for LFY–UFO transcriptional activity (Fig. 1). In the protoplast assay, altering LUBS1 or LUBS2 (or both) significantly reduced *pAP3* activation (Fig. 3c); the LUBS1 alteration had a stronger

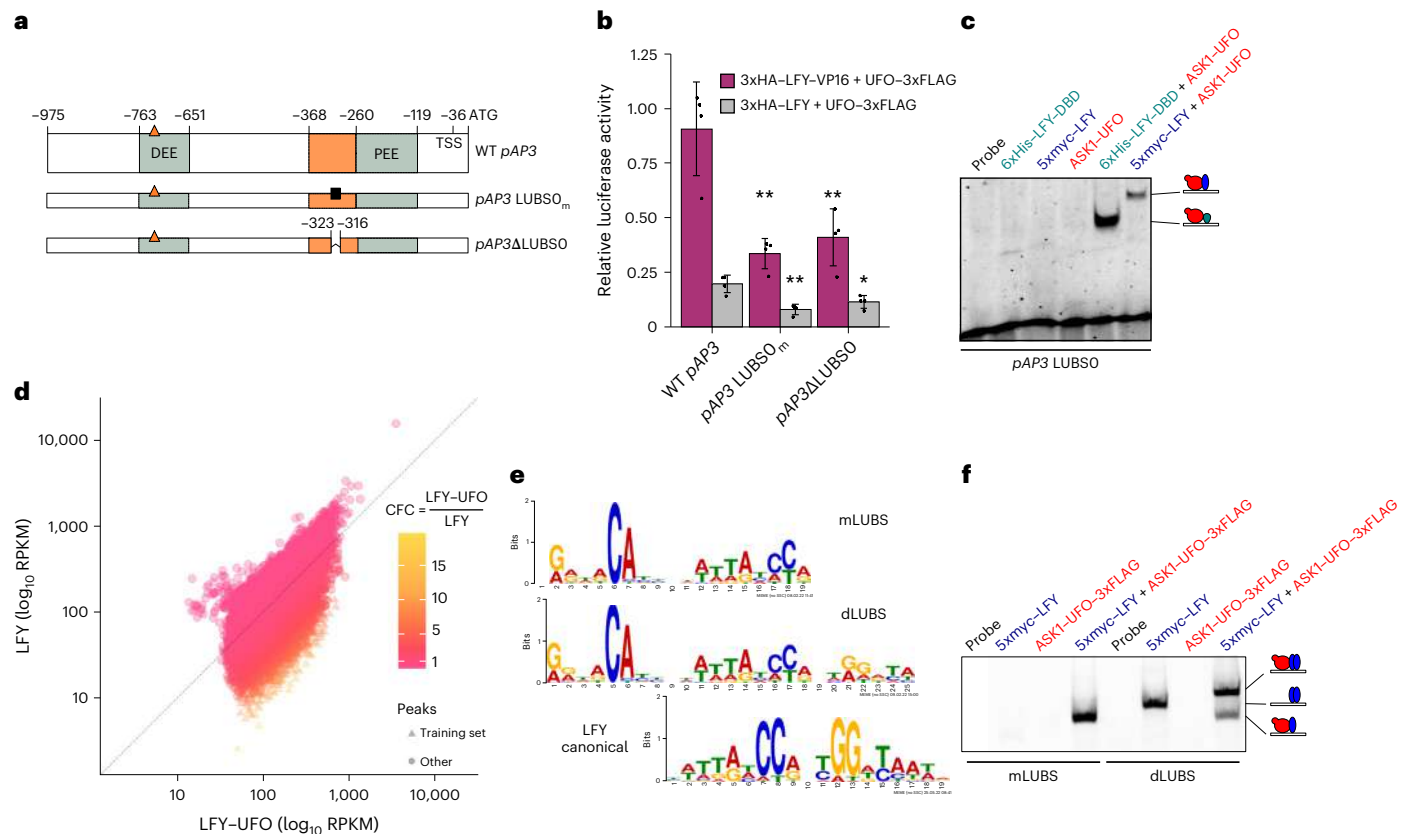


Fig. 2 | LFY and UFO together bind a new DNA motif. a, WT *pAP3* with regulatory regions and *cis*-elements (top line). The coordinates are relative to the *AP3* start codon. TSS, transcription start site. The orange triangles represent canonical LFYBS. The orange rectangle represents the 107-bp region and the black square represents the mutation introduced in *pAP3* LUBSO (*LUBSO_m*). The detailed functional dissection of the 107-bp region and the LUBSO mutation (*LUBSO_m*) are described in Extended Data Fig. 3. The other rows show the promoter versions used in **b**. **b**, *pAP3* activation in *Arabidopsis* protoplasts. The data are mean $\pm$ s.d. ($n = 4$ biological replicates). We used one-way ANOVA with data from the same effector and Tukey's multiple comparisons tests. The asterisks represent a significant statistical difference compared with WT *pAP3* (* $P < 0.05$; ** $P < 0.01$). **c**, EMSA with LUBSO DNA probe and the indicated proteins. Size exclusion chromatography (SEC) coupled to multi-angle laser light scattering

(MALLS) established a mass of 102 ± 3.3 kDa for the ASK1-UFO-LFY-DBD-LUBSO complex, consistent with a 1:1:1:1 stoichiometry (Extended Data Fig. 3e). The drawings represent the different complexes with FL LFY (blue), LFY-DBD (pale blue) and ASK1-UFO (red) on DNA. **d**, Comparison of peak coverage in LFY and LFY-UFO ampDAP-seq experiments, coloured by CFC. The LFY-UFO-specific peaks used to build the mLUBS and dLUBS motifs in **e** are triangle-shaped. RPKM, reads per kilobase per million. **e**, Logos for mLUBS, dLUBS and LFYBS. The LFY logo was generated using the 600 peaks with the strongest LFY ampDAP-seq signal. **f**, EMSAs with the mLUBS and dLUBS DNA probes that had the highest-scoring sequences. The drawings represent the different complexes with LFY (blue) and ASK1-UFO (red) on DNA. Source data are available in Supplementary Data 4.

effect. Specifically altering the URM of *pAP3* LUBS1 and LUBS2, which abolished LFY-UFO binding on individual sites in EMSAs (Extended Data Fig. 5e), also reduced *pAP3* activation, albeit less effectively than altering the whole LUBS (Extended Data Fig. 5f). Finally, the previously described *pAP3::GUS* staining pattern in the second and third whorls of early floral meristems in *Arabidopsis* was severely reduced when LUBS1 and LUBS2 were altered, demonstrating the importance of these sites in vivo (Fig. 3d and Extended Data Fig. 5g). Similarly, the *RBE* promoter contains an ASK1-UFO-LFY ampDAP-seq peak that is absent with LFY alone (Extended Data Fig. 6a), and the functional importance of the single LUBS identified under this peak was confirmed using EMSAs, transient assays in protoplasts and stable reporter constructs in plants (Extended Data Fig. 6b–e).

In addition to *AP3* and *RBE*, LFY and UFO together probably regulate many other genes in *Arabidopsis*. To identify such potential LFY-UFO targets, we established a list of genes bound (in ampDAP and ChIP) and regulated by LFY-UFO (Extended Data Fig. 7a). This procedure identified the other B gene *PISTILLATA* (*PI*), previously proposed as a LFY-UFO target but through an unknown regulatory element that the LUBS model precisely localized (Extended Data Fig. 7b). We also found floral regulators such as *SQUAMOSA PROMOTER BINDING PROTEIN*-

LIKE 5 and *FD* as well as additional candidates probably regulated by LFY and UFO (Extended Data Fig. 7a,c).

The LFY K249R substitution specifically affects UFO-dependent LFY functions

In *Arabidopsis*, LFY performs UFO-dependent and independent functions³, and we wondered whether they could be uncoupled by introducing specific alterations in LFY. As we were initially looking for LFY ubiquitination mutants, we substituted exposed lysines of LFY-DBD with arginines and tested the effect of such alterations on LFY-UFO-dependent *pAP3* activation in protoplasts. We found one substitution (LFYK249R; Extended Data Fig. 8a) that strongly reduced *pAP3* activation by LFY-UFO (Fig. 4a) or LFY-VP16-UFO (Extended Data Fig. 8b) without affecting the UFO-independent *pAG* activation (Extended Data Fig. 8c) or the LFY-UFO interaction (Extended Data Fig. 8d). AmpDAP-seq experiments showed that the LFY K249R substitution specifically impaired the binding of LFY-UFO but not that of LFY alone (Fig. 4b,c and Extended Data Fig. 8e–i), revealing that Lys249 plays a key role in LFY-UFO interaction with the LUBS DNA.

The importance of LFY Lys249 for UFO-dependent LFY functions was also confirmed using complementation assays of the *Arabidopsis*

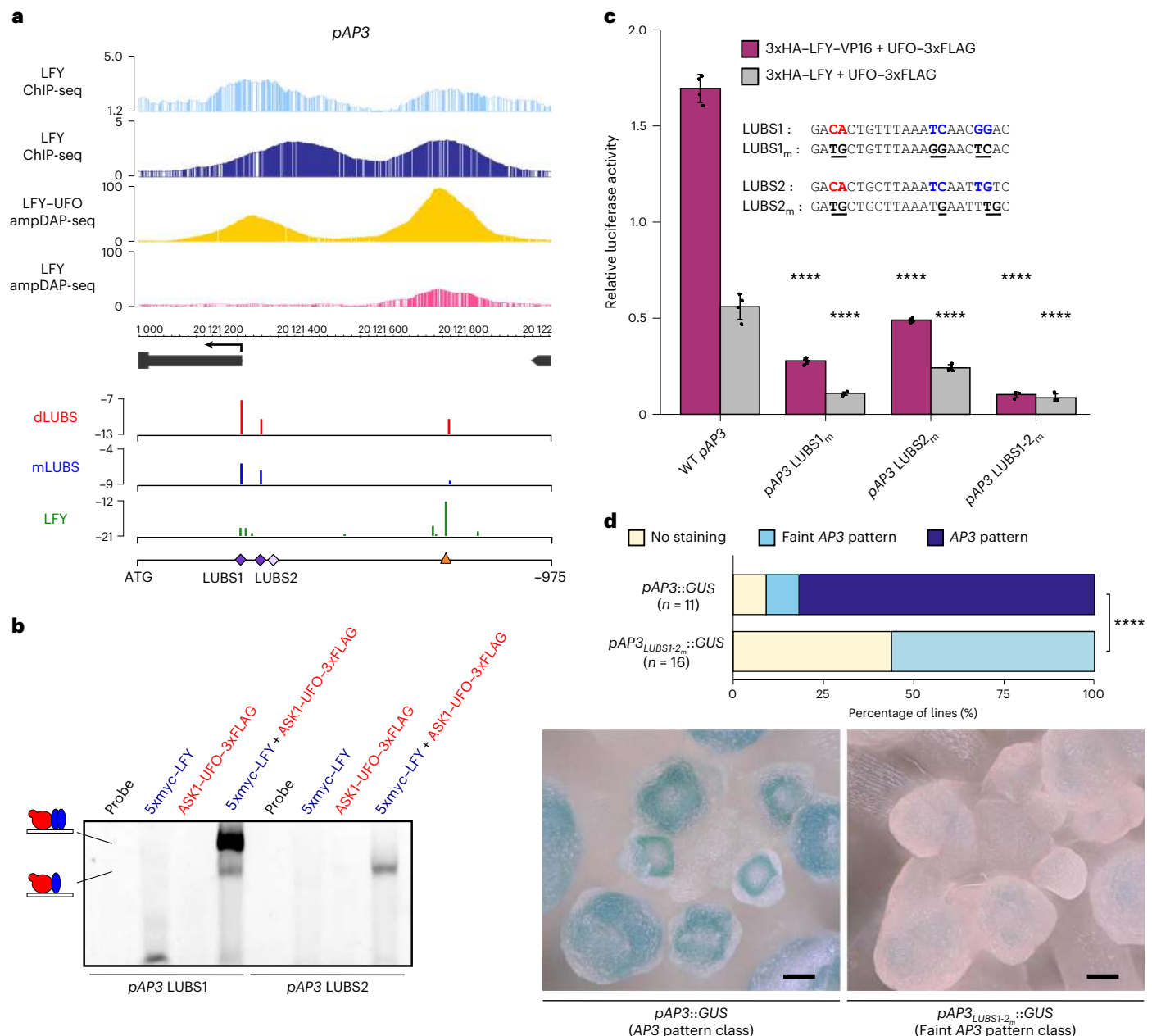


Fig. 3 | Functional validation of LUBS. a, Top, Integrated Genome Browser view of *pAP3* showing LFY ChIP-seq in inflorescences (light blue)²⁵ and seedlings (dark blue)²⁶, LFY-UFO ampDAP-seq (yellow)⁴⁸, and LFY ampDAP-seq (pink)⁴⁸. The y axis indicates the read number range. Bottom, identification of LUBS in *pAP3*. The predicted binding sites using dLUBS and mLUBS models and the LFY PWM are shown; the y axis represents score values. LUBS1 and LUBS2 are indicated with purple diamonds, and canonical LFYBS is shown as an orange triangle. LUBS0 (light purple diamond) is not visible because of its low score. **b**, EMSAs with *pAP3* LUBS probes. The drawings represent the different complexes involving LFY (blue) and ASK1-UFO (red) on DNA. **c**, *pAP3* activation in *Arabidopsis* protoplasts. The effects of alterations (underlined) in URM (red) and LFYBS (blue) bases of *pAP3* LUBS were assayed. The data are mean $\pm$ s.d. ($n = 4$ biological replicates).

One-way ANOVAs were performed with data from the same effector (one-way ANOVA with Tukey's multiple comparisons tests for 3xHA-LFY + UFO-3xFLAG data and Welch's ANOVA with Games-Howell post hoc tests for 3xHA-LFY-VP16 + UFO-3xFLAG data). The asterisks represent a statistical difference compared with the WT promoter (**** $P < 0.0001$). **d**, In vivo analysis of *pAP3::GUS* fusions. The percentage of transgenic lines with an AP3 pattern, a faint AP3 pattern or absence of staining is shown (top). The pattern distributions differ between the two constructs (χ^2 test; **** $P < 0.0001$). n , number of independent lines. Representative pictures of plants with an AP3 pattern (bottom left) and a faint AP3 pattern (bottom right) are also shown. Scale bars, 50 μ m. Note the staining in the ring corresponding to the second and third whorl primordia in the left picture. Source data are available in Supplementary Data 4.

lfy-12 null mutant²⁷. *lfy-12* plants expressing LFY^{K249R} or LFY^{K249S} under the control of the *LFY* promoter developed flowers with normal sepals and carpels but with defective third-whorl and, more importantly, second-whorl organs, resulting in flowers similar to those observed in weak *ufo* mutants (Fig. 4d). When expressed under the constitutive 35S promoter, LFY^{K249R} triggered ectopic flower formation and early flowering like WT LFY (Extended Data Fig. 8j), consistent with these

LFY functions being independent of UFO and thus not affected by the K249R substitution²⁸.

Structural characterization of the ASK1-UFO-LFY-DNA complex

To understand how the LFY-UFO complex recognizes its cognate DNA binding site and how the Lys249 alteration impedes this interaction,

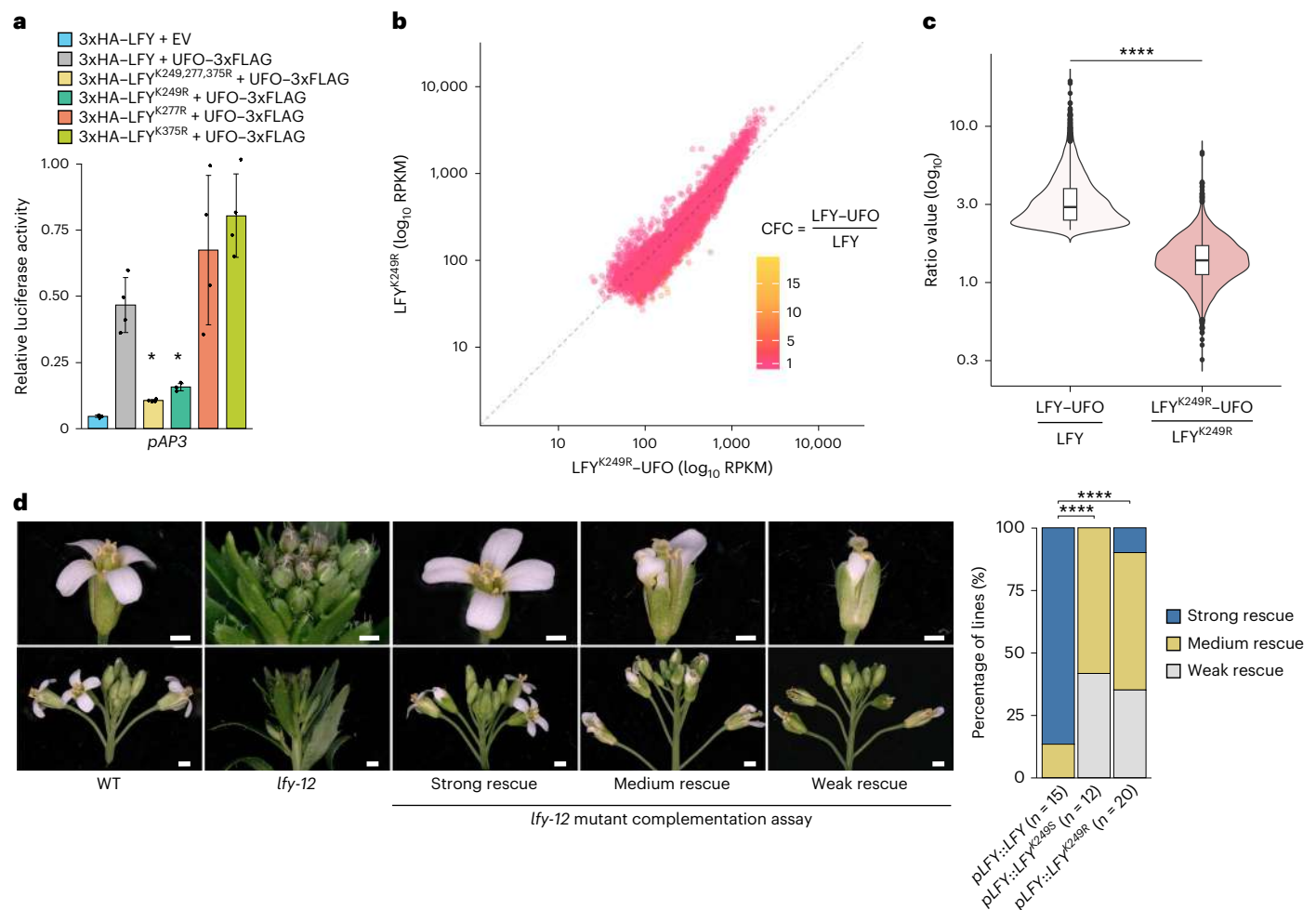


Fig. 4 | The LFY K249R substitution disrupts the LFY-UFO synergy. **a**, pAP3 activation in *Arabidopsis* protoplasts. The data are mean $\pm$ s.d. ($n = 4$ biological replicates). We used Welch's ANOVA with Games-Howell post hoc tests. The asterisks indicate a statistical difference compared with 3xHA-LFY + UFO-3xFLAG ($*P < 0.05$). **b**, Comparison of peak coverage in LFY^{K249R}-UFO (x-axis) and LFY^{K249R} (y-axis) ampDAP-seq experiments, coloured by peak CFC as in Fig. 2d. Note that, in contrast to Fig. 2d, the LFY-UFO-specific regions are mostly absent. **c**, Distribution of coverage ratios for LFY and LFY^{K249R} for LFY-UFO-specific regions (20% highest CFC, $n = 3,843$ genomic regions). We used a Wilcoxon rank sum test ($****P < 0.0001$). The median (solid line), interquartile range

(box edges), $\pm 1.5 \times$ the interquartile range (whiskers) and outliers (dots) are shown. **d**, *lfy-12* mutant complementation assay. The WT, the *lfy-12* mutant and representative plants of the different phenotypic complementation classes are shown on the left. Scale bars, 1 mm for the top row and 1 cm for the bottom row. The distribution of the different lines within the phenotypic complementation classes is shown on the right. Plants complemented with LFY^{K249R} and LFY^{K249S} show different complementation patterns than plants complemented with LFY (χ^2 tests; $****P < 0.0001$, n , number of independent lines. Source data are available in Supplementary Data 4.

we purified the ASK1-UFO-LFY-DBD-LUBS1 complex and structurally characterized it using cryo-electron microscopy (cryo-EM) (Fig. 5a and Extended Data Fig. 9a–d). A structure at a 4.27 Å resolution was obtained (Extended Data Fig. 9g–i) into which were fit the AlphaFold2 predicted structures for UFO and ASK1, and the LFY-DBD dimer/DNA crystallographic structure²⁹ (PDB, 2VY1; Fig. 5b and Extended Data Fig. 9e,f). Due to the modest resolution, specific interacting amino acids could not be unambiguously identified. However, the major protein-protein and protein-DNA interaction surfaces were clearly identifiable.

The structure revealed that UFO directly contacts the DNA in the major groove around the URM (Fig. 5c). This binding probably involves basic residues present on loops projecting from the UFO Kelch-type β -propeller and results in a bend of roughly 30 degrees in the DNA double helix (Extended Data Fig. 9f). The structure also shows an interface between UFO and one LFY-DBD monomer (Fig. 5c). The LFY-DBD loop containing the Lys249 residue lies in this interface and probably interacts with one of the DNA-binding loops of UFO, consistent with the key role of LFY Lys249 in the ternary complex formation. As expected,

ASK1 interacts with the UFO F-box domain¹⁵ (Fig. 5d). These data show how a β -propeller protein is able to modify the specificity of a TF, and they offer a structural explanation of how LFY and UFO synergistically recognize a specific DNA element via direct interactions by both proteins with the DNA.

The LFY-UFO complex might have a deep evolutionary origin

As genetic and physical LFY-UFO interactions have been described in diverse angiosperms, we wondered whether the mechanism unravelled for *Arabidopsis* proteins could also apply to LFY from other species, including non-angiosperm ones. We selected LFY orthologous proteins from several species and with different DNA binding specificities (Fig. 6a). LFY specificity has evolved with three major DNA binding specificities³⁰. Type I specificity is the one described in *Arabidopsis* and is valid for other angiosperms, gymnosperms, ferns and the moss *Marchantia polymorpha*, with two half-sites separated by a 3-bp spacer (Fig. 2e). LFY from the moss *Physcomitrium patens* has a type II

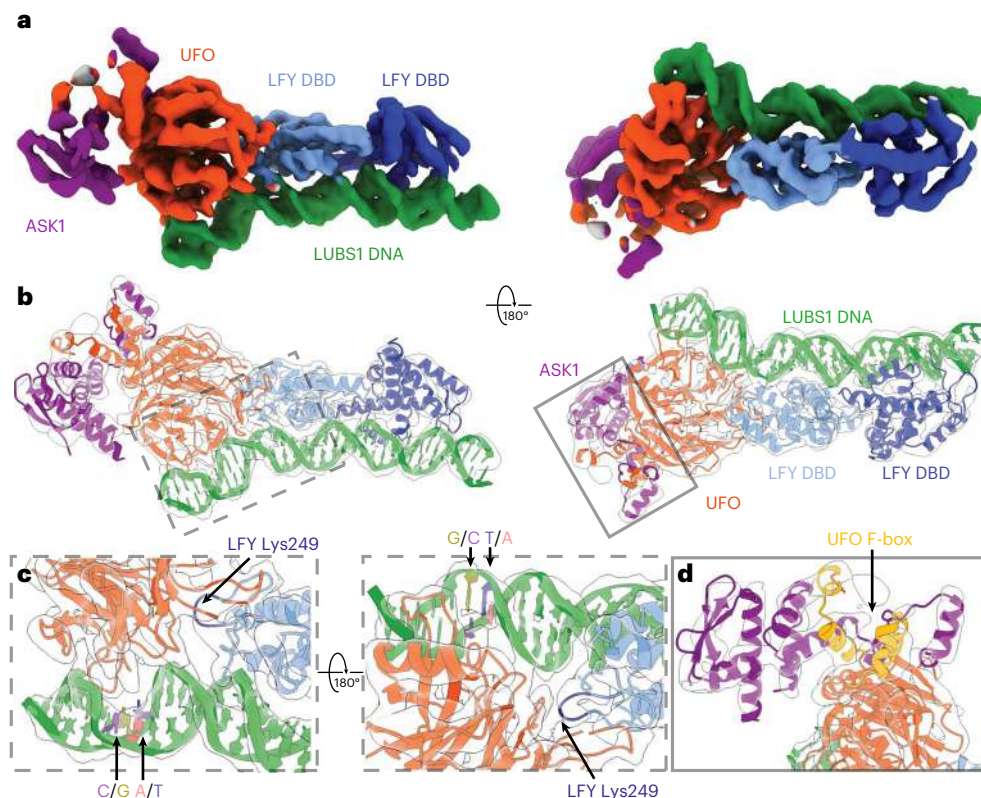


Fig. 5 | Structural characterization of the ASK1-UFO-LFY-DNA complex.

a, Cryo-EM density map of the ASK1-UFO-LFY-DBD-LUBS1 complex under two angles, coloured with regard to the underlying macromolecule (green for LUBS1 DNA, pale and dark blue for LFY-DBD, red for UFO, and purple for ASK1). **b**, The same views of the cryo-EM density map in transparent grey with fitted structures of the LFY-DBD dimer, UFO, ASK1 and LUBS1 DNA. The colours are the same as in **a**. The frames roughly indicate the regions shown in **c** and **d**. **c**, Zoom on the

UFO-DNA contact region (left) and on the LFY-UFO interface (right). Only the high-information CA of the URM and its complement are highlighted by filled colouring the rings for each base (red for A, blue for T, pale green for G and purple for C). The LFY-DBD loop containing the Lys249 residue is highlighted in dark blue. **d**, Zoom on the ASK1-UFO interface, with the UFO F-box highlighted in gold. Source data are available in Supplementary Data 4.

specificity with specific half-sites (different from type I half-sites) also separated by a 3-bp spacer. Finally, type III specificity is found for LFY from algae and corresponds to a type II motif without the spacer. Because functional UFO homologues have not been identified outside angiosperms, we used *Arabidopsis* UFO (AtUFO) in all the following experiments.

We tested the interactions of various LFY orthologues with AtUFO in yeast two-hybrid (Y2H) (Fig. 6b), in DLRAs in protoplasts with *Arabidopsis* pAP3 (Fig. 6c) and in EMSAs (Fig. 6d). In Y2H, all LFYs except LFY from *P. patens* (type II) interacted with AtUFO (Fig. 6b). However, only type I LFYs from angiosperms, gymnosperms and ferns formed a complex on pAP3 LUBS and activated pAP3 in the protoplast assay (Fig. 6c,d). These results suggest that the ability of LFY and UFO to act together by forming a complex is ancient, largely predating the origin of angiosperms. We obtained no evidence that type II and III LFYs (from moss and algae) could form a complex with AtUFO on LUBS1 and LUBS2. A detailed and more trustworthy history of the LFY-UFO interaction will await further analyses, notably with the identification of UFO orthologues from non-angiosperm genomes.

Discussion

LFY has long been known to interact with UFO to control flower and inflorescence development in numerous angiosperm species. However, the molecular nature of their synergistic action remained unknown. As UFO encodes an F-box protein taking part in an SCF complex^{17,31,32}, it was thought to target proteins for SCF^{UFO}-dependent ubiquitination

and possible degradation. LFY was an obvious target candidate, but clear evidence of LFY ubiquitination was missing^{12,18}. The results we present here suggest that the F-box domain, required for ubiquitination, is dispensable for most UFO-dependent LFY activity. Nevertheless, the high conservation level of the UFO F-box sequence in angiosperms, together with slight differences in UFO activity when the F-box is deleted, suggests that this domain might still be needed for some elusive facets of UFO function. UFO may work redundantly with other F-box proteins in ubiquitination pathways, such as the F-box protein HAWAIIAN SKIRT identified in a genetic screen as an enhancer of the *ufo* mutant phenotype³³. It is thus possible that UFO acts as a moonlighting protein³⁴ with functions in both transcription and ubiquitination, and these two activities could be related or independent.

The molecular mechanism we discovered here is consistent with most published data on AP3 and PI regulation^{18,23,35,36}. However, a detailed understanding of the expression patterns of AP3 and RBE will require further work on other *cis*- and *trans*-elements. Why AP3 is not transcribed in floral stage 0–1 despite the expression of LFY and UFO is unclear²⁰. It could be because SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1), AGAMOUS-LIKE 24 (AGL24) and SHORT VEGETATIVE PHASE (SVP) act as early AP3 repressors, as AP3 mRNA is detected in the floral anlage in a *soc1 svp agl24* mutant^{37,38}. Another explanation could be that AP3 expression requires the SEPALLATA3 activator³⁹. Why pAP3 is not activated by LFY (or LFY-VP16) alone through the canonical LFYBS is also an open question.

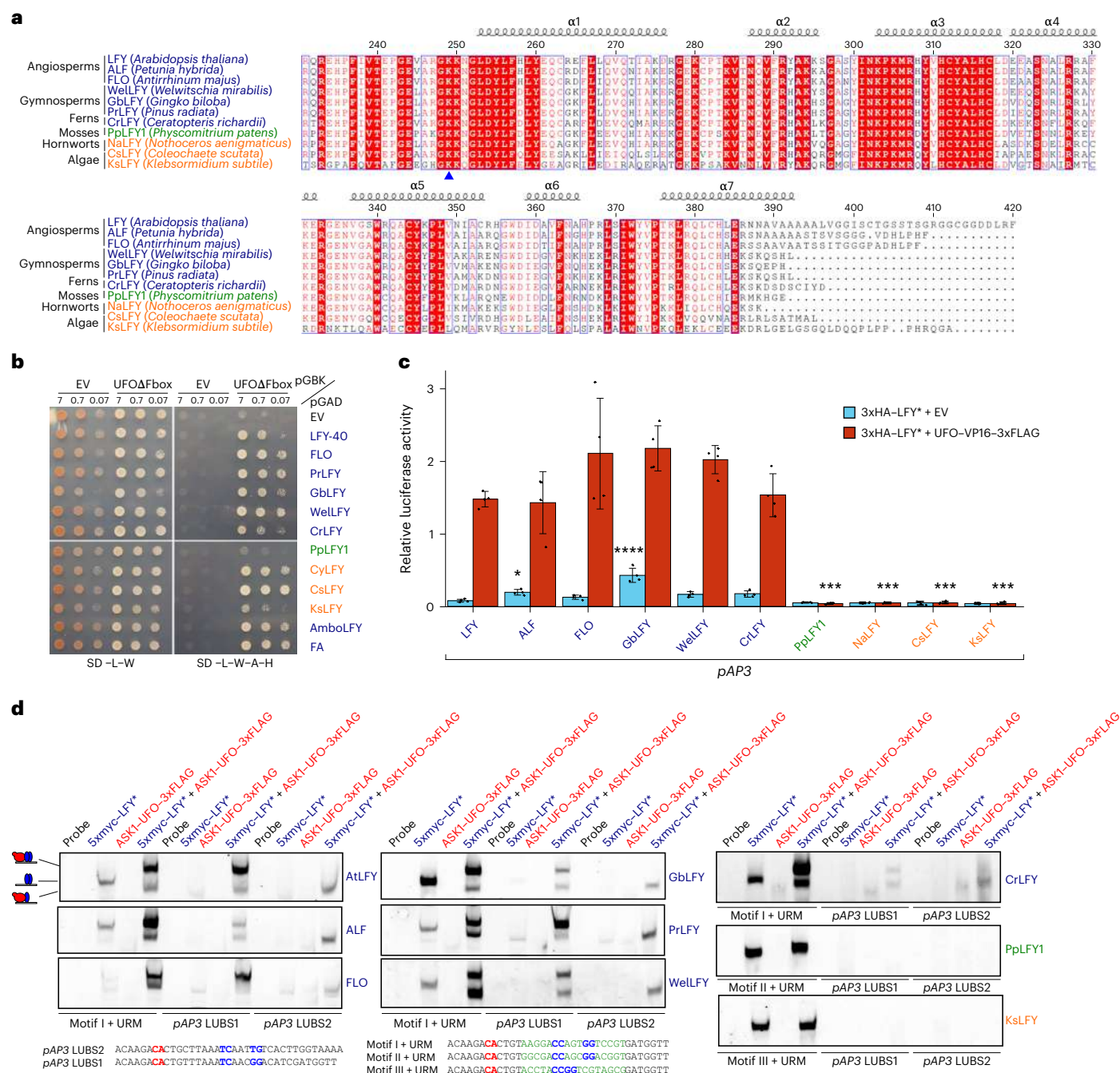


Fig. 6 | LFY-UFO interaction is conserved beyond angiosperm species.

a, Alignment of LFY DBDs. The amino acid numbering and secondary structure annotation are based on LFY from *A. thaliana*. The LFY Lys249 residue is indicated with a blue triangle. DNA binding specificities are colour-coded: types I (blue), II (green) and III (orange). FLO, FLORICAULA; ALF, ABERRANT LEAF AND FLOWER. **b**, Interaction between LFY orthologues and AtUFOΔFbox in Y2H. The LFY orthologues are described in **a** except CyLFY (*Cylindrocystis* sp.), AmbLFY (*Amborella trichopoda*) and FA (FALSIFLORA; *Solanum lycopersicum*). SD, synthetic defined. See Extended Data Fig. 4d for the legends. **c**, pAP3 activation measured by DLARs in *Arabidopsis* protoplasts. 3xHA-LFY* refers to the different LFY orthologues indicated under the x axis. The data represent averages of independent biological replicates and are presented as mean ± s.d., with each dot

representing one biological replicate ($n = 4$). One-way ANOVAs were performed with data from the same effector (one-way ANOVA with Tukey's multiple comparisons tests for 3xHA-LFY* + EV data and Welch's ANOVA with Games-Howell post hoc tests for 3xHA-LFY* + UFO-VP16-3xFLAG data). The asterisks represent a statistical difference compared with AtLFY (* $P < 0.05$; *** $P < 0.001$; **** $P < 0.0001$). **d**, EMSAs with the indicated DNA probes (bottom). URM and LFYBS bases are depicted in red and blue, respectively. The pAP3LUBS1 sequence was modified to insert the perfect sequence of motif I, II or III³⁰ (depicted in green); these DNA probes were used as positive controls for the binding of LFYs alone and for LFY-UFO complex formation. 5xmyc-LFY* refers to the different LFY orthologues indicated next to each EMSA and described in **a**. Source data are available in Supplementary Data 4.

Our work unravelled an unsuspected function unrelated to ubiquitination for UFO: it forms a transcriptional complex with LFY at regulatory sites that are different from the canonical sites bound by a LFY homodimer. UFO was previously proposed to act in transcription,

but in the absence of direct evidence that a LFY-UFO complex forms on new binding sites, it was difficult to understand how UFO controls only a subset of LFY targets. These new regulatory sites (mLUBS and dLUBS) are made of a low-affinity or half LFYBS (poorly or not bound

by LFY alone) and a motif located at a fixed distance from it and responsible for UFO recruitment. The formation of such a sequence-specific complex is explained at the structural level by the capacity of UFO to interact with both LFY and DNA. The poor ability of UFO to bind DNA alone explains its complete dependence on LFY to perform its transcriptional functions in planta^{6,20}. Thus, depending on the *cis*-elements present in regulatory regions, LFY either binds DNA as a homodimer or requires UFO to form a ternary complex. The alteration of the LFY Lys249 residue allows these two types of binding to be uncoupled by specifically disrupting the formation of the LFY–UFO–DNA complex. The position of this residue in the three-dimensional structure at the interface among LFY, UFO and DNA is consistent with the key role of this residue in the complex formation. It is possible that replacing Lys249 with a bulkier Arg residue displaces the UFO loops involved in DNA binding without affecting the LFY–UFO interaction. Obtaining a higher-resolution structure will help us precisely understand the interactions occurring in this complex.

Although it might be a common regulatory mechanism, only a few cases where non-TF proteins modify TF DNA binding specificity have been described so far (for example, Met4 and Met28 modifying the binding of TF Cbfl in yeast⁴⁰, or the herpes simplex virus transcriptional activator VP16 changing the specificity of the Oct-1/HCF-1 complex⁴³). None of these examples involves an F-box protein or a Kelch-type β -propeller protein, and none has been characterized at the structural level. The modification of TF DNA binding specificity by non-TF proteins offers additional possibilities for the combinatorial control of gene expression and explains how a master regulator such as LFY accesses specific *cis*-elements to perform different functions in distinct territories.

Since LFY and UFO play key roles together in numerous plant species (including ornamental, crop and model plants), our findings expand the molecular understanding of flower and inflorescence development in a large variety of angiosperms. Because the LFY–UFO synergy is observed with LFY orthologues from gymnosperms and ferns as well, we speculate that this complex largely predated the origin of flowers and could have been co-opted for flower development from a yet-unknown ancestral role.

Methods

Arabidopsis growth

All mutants and transgenic lines are in the *A. thaliana* Columbia-0 (Col-0) accession. Seeds were sown on soil, stratified for three days at 4 °C and then grown at 22 °C under long-day conditions (16 h light). Transgenic plants were obtained with *Agrobacterium tumefaciens* C58C1 pMP90 using the floral dip method. Transformants were identified using GFP or Basta selection.

Arabidopsis cell suspension culture

Arabidopsis thaliana (ecotype Col-0) cells in suspension cultures were grown under continuous light (90 μ mol of photons per m² per s) at 21 °C with shaking in Murashige and Skoog medium supplemented with 30 g l⁻¹ sucrose and 2 mg l⁻¹ 2,4-dichlorophenoxyacetic acid, pH 5.5. The suspension cells were subcultured every week with a fivefold dilution. Suspension cells at four or five days following subculture were used for protoplast preparation.

Cloning

DNA fragments were amplified by PCR with Phusion high-fidelity polymerase (NEB). Plasmids were all obtained by Gibson Assembly with either PCR-amplified or restriction-enzyme-digested backbone vectors. We used the 420-amino-acid LFY version. For site-directed mutagenesis, primers containing the desired mutations were used for Gibson Assembly mutagenesis. Plasmids were obtained using DH5 α bacteria and were all verified by Sanger sequencing. A list of plasmids and cloning procedures is provided in Supplementary Data 1. The oligonucleotide sequences are listed in Supplementary Data 2.

Y2H

The coding sequences were cloned in pGADT7-AD or pGBKT7 vectors (Clontech) by Gibson Assembly. Y187 and AH109 yeast strains (Clontech) were transformed with pGADT7-AD or pGBKT7 vectors and selected on plates lacking leucine (SD –L) or tryptophan (SD –W), respectively (MP Biomedicals). After mating, the yeasts were restreaked on plates lacking leucine and tryptophan (SD –L–W) for two days. The yeasts were then resuspended in sterile water, and OD_{600nm} was adjusted to the indicated values for all constructions; two tenfold dilutions were performed, and 6 μ l drops were done on SD –L–W or SD –L–W–A–H (lacking leucine, tryptophan, histidine and adenine) plates. The yeasts were grown at 28 °C, and pictures were taken at the indicated times.

DLRAs in *Arabidopsis* protoplasts

Effector plasmids with a 3xHA tag were obtained by cloning the indicated genes in the modified pRT104 vector containing a 3xHA N-terminal tag (pRT104–3xHA)⁴². The pRT104 empty plasmid was reengineered to insert a 3xFLAG C-terminal tag. For reporter plasmids, the indicated promoter fragments were cloned upstream of a firefly luciferase gene in pBB174 (ref. ⁴³). We used a 975-bp *pAP3* fragment and a 2-kilobase (kb) *pRBE* promoter fragment upstream of the ATG, known to induce a WT pattern in planta^{23,44}. *pAG* corresponds to the *AG* second intron fused to a minimal 35S promoter, known to induce a WT pattern in planta²². For *pAPI*, we used a 600-bp fragment upstream of the ATG. This version is sufficient to give a WT pattern in planta⁴⁵, and the use of longer promoter versions induced a very high background noise in protoplasts. The pRLC reference plasmid contains a *Renilla* luciferase sequence under the control of the 35S promoter. Plasmids were obtained in large amounts using a NucleoBond Xtra Maxi Plus kit (Macherey-Nagel). Protoplasts were prepared from *Arabidopsis* Col-0 cell suspension and transformed following the procedure described by Iwata et al.⁴⁶. The cell walls were digested using Onuzuka R-10 cellulase and macerozyme R-10 (Yakult Pharmaceutical). The digested cells were passed through two layers of Miracloth to remove debris, and the protoplast concentration was adjusted to 2–5 $\times$ 10⁵ cells per ml. The protoplasts were then PEG-mediated transformed using 10 μ g of the indicated effector and reporter plasmids and 2 μ g of the reference plasmid. After 17 h of incubation at room temperature, the protoplasts were lysed. Firefly (F-LUC) and *Renilla* luciferase (R-LUC) activities were measured using a Dual Luciferase Reporter Assay System (Promega) and a TECAN Spark 10 M 96-well plate reader. F-LUC/R-LUC luminescence ratios were calculated with background-corrected values. Four biological replicates were done for each plasmid combination.

EMSAs

The DNA probes used in EMSAs are listed in Supplementary Data 2. Complementary oligonucleotides were annealed overnight in annealing buffer (10 mM Tris (pH 7.5), 150 mM NaCl and 1 mM EDTA). Then, 4 pmol of double-stranded DNA was fluorescently labelled with 1 unit of Klenow fragment polymerase (NEB) and 8 pmol of Cy5-dCTP (Cytiva) in Klenow buffer for 1 h at 37 °C. The enzymatic reaction was stopped with a 10 min incubation at 65 °C.

The proteins used in EMSAs were obtained by different methods (bacteria, insect cells or quick coupled transcription/translation (TnT)). The concentrations of recombinant proteins (6xHis–LFY–DBD and UFO Δ Fbox–3xFLAG) and recombinant complexes (ASK1–UFO and ASK1–UFO–3xFLAG) were adjusted to 500 nM for all reactions. All the 5xmyc-tagged proteins were obtained *in vitro* by TnT. We did 50 μ l TnT reactions by mixing 5 μ g of pTNT–5xmyc plasmid containing the gene of interest with TnT SP6 High-Yield Wheat Germ Protein Expression System (Promega) for 2 h at 25 °C. For EMSAs with TnT-produced proteins, 5 μ l of TnT reaction was used. Recombinant protein buffer or TnT mix was used as a control when comparing reactions with multiple proteins.

All binding reactions were performed in 20 μ l of binding buffer (20 mM Tris (pH 7.5), 150 mM NaCl, 1% glycerol, 0.25 mM EDTA, 2 mM

MgCl₂, 0.01% Tween-20 and 3 mM TCEP) with 10 nM labelled probe. The reactions were supplemented with 140 ng μl^{-1} fish sperm DNA (Sigma-Aldrich) for EMSAs performed with in-vitro-produced LFY and 200 ng μl^{-1} for EMSAs performed with recombinant 6xHis-LFY-DBD. The binding reactions were incubated for 20 min on ice and then loaded on a 6% native polyacrylamide gel. The gels were electrophoresed at 90 V for 75 min at 4 °C and revealed with an Amersham ImageQuant 800 imager (Cytiva). The uncropped gels are shown in the Source data.

Recombinant protein production and purification from bacteria

We produced 6xHis-LFY-DBD in *E. coli* Rosetta2 (DE3) cells (Novagen) and purified it as previously described²⁹. ASK1 was cloned into the pETM-11 expression vector⁴⁷, and the resulting plasmid was transformed into *E. coli* BL21 cells (Novagen). The bacteria were grown in LB medium supplemented with kanamycin and chloramphenicol at 37 °C up to an OD_{600nm} of 0.6. The cells were then shifted to 18 °C, and 0.4 mM isopropyl β -D-1-thiogalactopyranoside was added. After an overnight incubation, the cells were sonicated in UFO buffer (25 mM Tris (pH 8), 150 mM NaCl and 1 mM TCEP) supplemented with one EDTA-free Pierce Protease Inhibitor Tablet (Thermo Fisher). The lysed cells were then centrifuged for 30 min at 27,000 g. The supernatant was mixed with Ni Sepharose High Performance resin (Cytiva) previously equilibrated with UFO buffer (25 mM Tris (pH 8), 150 mM NaCl and 1 mM TCEP). The resin was then washed with UFO buffer containing 20 and 40 mM imidazole. Bound proteins were eluted with UFO buffer containing 300 mM imidazole and dialysed overnight at 4 °C against UFO buffer without imidazole.

Recombinant protein production and purification from insect cells

The different tagged versions of ASK1, LFY and UFO were cloned in acceptor and donor plasmids (pACEBac1, pIDK and pIDS, respectively; Geneva Biotech). The final acceptor plasmids containing the desired combination of coding sequences were obtained with Cre recombinase (NEB). DH10EmBacY-competent cells containing the baculovirus genomic DNA (bacmid) were transformed with the final acceptor plasmids. Blue-white selection was used to identify colonies with a recombinant bacmid with the acceptor plasmid inserted. Bacmid was then isolated from bacteria and mixed with X-tremeGENE HP DNA Transfection Reagent (Roche) to transfect Sf21 insect cells. At 96 h after transfection, supernatant containing the recombinant baculovirus (V0) was collected and used to infect fresh Sf21 cells. When the infected cells reached the day post arrest, V1 virus was collected. For large expression, Sf21 cells were infected with either V1 virus or frozen baculovirus-infected cells. The pellet of a 0.75 l culture was sonicated in 50 ml of UFO buffer supplemented with one EDTA-free Pierce Protease Inhibitor Tablet (Thermo Fisher). The sonicated cells were centrifuged for 1.5 h at 120,000 g at 4 °C. The supernatant was then incubated for 1 h at 4 °C with Ni Sepharose High Performance resin (Cytiva) previously equilibrated with UFO buffer. The beads were transferred into a column and washed with 20 column volumes of UFO buffer, then UFO buffer + 50 mM imidazole. Proteins were eluted with UFO buffer containing 300 mM imidazole. The elution was dialysed overnight at 4 °C against UFO buffer. TEV protease was added to cleave tags (0.01% w/w). When ASK1 was limiting compared with UFO, recombinant 6xHis-ASK1 from bacteria was added. The following day, the elution was repassed on Dextrin Sepharose High Performance (Cytiva) and Ni Sepharose High Performance resins (Cytiva) to remove tags and contaminants. For ASK1-UFO, ASK1-UFO-3xFLAG or UFO Δ Fbox-3xFLAG, the proteins were concentrated with a 30 kDa Amicon Ultra Centrifugal filter (Millipore) and further purified by SEC. For ASK1-UFO-LFY-DBD complex purification, contaminant DNA was removed by passing proteins on Q Sepharose High Performance resin (Cytiva) pre-equilibrated with UFO buffer. Increasing salt concentrations allowed us to obtain

DNA-free proteins. The indicated annealed HPLC-purified oligonucleotides (Supplementary Data 2) were then added and incubated with proteins on ice for 20 min. The proteins were concentrated with a 30 kDa Amicon Ultra Centrifugal filter (Millipore) and further purified by SEC.

SEC and SEC-MALLS

SEC was performed with a Superdex 200 Increase 10/300 GL column (Cytiva) equilibrated with UFO buffer. Unaggregated proteins of interest were frozen in liquid nitrogen and stored at -80 °C. SEC-MALLS was performed with a Superdex 200 Increase 10/300 GL column (Cytiva) equilibrated with UFO buffer. For each run, 50 μl containing 1 mg ml^{-1} of complex was injected. Separations were performed at room temperature with a flow rate of 0.5 ml min^{-1} . The elutions were monitored by using a Dawn Heleos II for MALLS measurement (Wyatt Technology) and an Optilab T-rEX refractometer for refractive index measurements (Wyatt Technology). Molecular mass calculations were performed using ASTRA software with a refractive index increment (dn/dc) of 0.185 ml g^{-1} .

ampDAP-seq

We used pTnT-5xmyc-LFY⁴⁸ to produce 5xmyc-LFY in vitro using a TnT SP6 High-Yield Wheat Germ Protein Expression System (Promega). We used the ampDAP-seq libraries described in Lai et al.⁴⁸. The ampDAP-seq experiments were performed in triplicates (LFY-UFO) or in duplicates (LFY^{K249R} and LFY^{K249R}-UFO).

A 50 μl TnT reaction producing 5xmyc-LFY was mixed with an excess of recombinant ASK1-UFO-3xFLAG (2 μg) and 20 μl of Pierce Anti-c-Myc Magnetic Beads (ThermoScientific). DAP buffer (20 mM Tris (pH 8), 150 mM NaCl, 1 mM TCEP and 0.005% NP40) was added to reach 200 μl . The mix was incubated for 1 h at 4 °C on a rotating wheel. The beads were then immobilized and washed three times with 100 μl of DAP buffer, moved to a new tube and washed once again. The ampDAP-seq input libraries (50 ng) were then added, and protein-DNA mixes were incubated for 1.5 h at 4 °C on a rotating wheel. The beads were immobilized and washed five times with 100 μl of DAP buffer, moved to a new tube and washed two more times. Finally, the beads were mixed with 30 μl of elution buffer (10 mM Tris (pH 8.5)) and heated for 10 min at 90 °C.

Immunoprecipitated DNA fragments contained in the elution were amplified by PCR according to the published protocol⁴⁹ with Illumina TruSeq primers. The remaining beads were mixed with 20 μl of 1 $\times$ SDS-PAGE Protein Sample Buffer, and western blots were performed to check the presence of tagged proteins. The PCR products were purified using AMPure XP magnetic beads (Beckman Coulter) following the manufacturer's instructions. Library molar concentrations were determined by quantitative PCR using a NEBNext Library Quant Kit for Illumina (NEB). The libraries were then pooled with equal molarity. Sequencing was done on Illumina HiSeq (Genewiz) with the specification of paired-end sequencing of 150 cycles.

GUS staining

The different promoter versions were cloned upstream of the *GUS* gene in the pRB14 backbone vector⁴⁵. Transformants were selected with GFP seed fluorescence. The number of independent lines analysed for each construct is indicated in each figure. GUS staining was performed on the apex of primary inflorescences of T₂ plants. Tissues were placed in ice-cold 90% acetone for 20 min at room temperature and then rinsed in GUS buffer without X-Gluc (0.2% Triton X-100, 50 mM NaPO₄ (pH 7.2), 2 mM potassium ferrocyanide and 2 mM potassium ferricyanide). The tissues were transferred to GUS buffer containing 2 mM X-Gluc substrate (X-Gluc DIRECT) and placed under vacuum for 5 min. The samples were then incubated overnight at 37 °C unless otherwise specified in the legend. Finally, the tissues were washed with different ethanol solutions (35%, 50% and 70%), and pictures were taken with a Keyence VHX-5000 microscope with a VH-Z100R objective.

In planta overexpression and mutant complementation assays

Tagged versions of UFO and UFOΔFbox were cloned under the control of the 35S promoter in pEGAD⁵⁰. Transformants were selected with Basta treatment. Overexpressing lines with a strong gain-of-function phenotype were crossed with the strong *ufo-1* mutant. Basta-resistant F₂ plants were individually genotyped to select *ufo-1* ^{-/-} homozygous plants. For this, a fragment was amplified by PCR with the oligonucleotides oGT1085 and oPR578 (Supplementary Data 2) and digested with DpnII enzyme (NEB). On the basis of the digestion profile, *ufo-1* ^{-/-} plants were kept and analysed once they reached flowering.

Altered versions of LFY were cloned in pETH29 (ref. ²⁹) or pCA26 (ref. ⁵¹) to express LFY complementary DNA under the control of its endogenous promoter or the 35S promoter, respectively. For the *lfy-12* complementation assay, heterozygous *lfy-12*/+ plants were transformed. The transformants were selected with GFP fluorescence and genotyped with a previously described protocol⁴⁵ to select *lfy-12* ^{-/-} plants. The complementation assay was performed with T₂ plants and was based on the analysis of the first ten flowers from the primary inflorescence. Pictures were taken with a Keyence VHX-5000 microscope with a VH-Z20R objective.

Western blots

For western blots on plant total protein extracts, the indicated tissues were crushed in 2× SDS–PAGE Protein Sample Buffer (100 mM Tris (pH 6.8), 20% glycerol, 2% SDS, 0.005% Bromophenol blue and 0.8% w/v dithiothreitol) at a 1:2 w/v ratio and boiled for 5 min. The samples were then loaded on a 12% acrylamide SDS–PAGE gel. For all western blots, transfer was performed with an iBlot2 Dry Blotting System (Invitrogen) using the default parameters. Membranes were blocked for 1 h at room temperature with 5% milk TBST and then incubated overnight at 4 °C with 5% milk TBST solution containing HRP-conjugated antibody (1:1,000 for anti-FLAG (Sigma-Aldrich; Cat. No. A8592) and 1:5,000 for anti-myc (Invitrogen; Cat. No. R951-25)). Revelation was performed with Clarity Western ECL substrate (Bio-Rad). Pictures were taken with a ChemiDoc MP Imaging System (BioRad). The uncropped gels are shown in the Source data.

Cryo-EM sample preparation, data collection and data processing

An aliquot of the SEC-purified ASK1–UFO–LFY–LUBS1 complex was thawed on ice (see Supplementary Data 2 for the LUBS1 DNA sequence). Subsequently, 3.5 µl of the complex at 1 mg ml⁻¹ was deposited onto glow-discharged (25 mA, 30 s) C-flat Au grid R 1.2/1.3 300 mesh (Electron Microscopy Sciences), blotted for 5.5 s with force 0 at 20 °C and 100% humidity using a Mark IV Vitrobot (FEI, Thermo Fisher Scientific), and plunge-frozen in liquid ethane for specimen vitrification. A dataset of about 1,000 videos of 40 frames was acquired on a 200 kV Glacios (Thermo Fisher Scientific) electron microscope (Supplementary Data 3) at a nominal magnification of 36,000 with a physical pixel size of 1.145 Å.

The raw videos, acquired with SerialEM on a Gatan K2 Summit camera (Supplementary Data 3), were imported to Cryosparc live⁵² for motion correction and CTF estimation. The dose-weighted micrographs were used for particle picking with crYOLO v1.7.6 and the general model for low-pass filtered images⁵³. Particle coordinates were imported to Cryosparc, where all subsequent steps were performed. After manual inspection, a subset of 761 micrographs was selected on the basis of CTF fit resolution, total and per-frame motion, average defocus and relative ice thickness. A raw particle stack of 282,567 images was extracted at a box size of 256 × 256 pixels², binned twice and subjected to two-dimensional classification to remove false positive picks. A total of 207,392 particles from the selected class averages were re-extracted, re-centred at full size and submitted for a second round of two-dimensional classification. All class averages showing clear protein features were selected, and the resulting 147,849 particles were used

for ab initio reconstruction with three classes and subsequent heterogeneous refinement of the resulting volumes. Of those three classes, two looked like a protein–DNA complex, with the most apparent difference being the presence or absence of an extra electron density at one edge of the DNA helix. The last class had no recognizable features and was used as a decoy to remove ‘junk’ particles. Each subset and volume of the two first classes was refined separately with non-uniform refinement⁵⁴, resulting in two distinct reconstructions of about 4.2 Å resolution, where the DNA model, the crystal structure of LFY–DBD and the AlphaFold2 models of UFO and ASK1 could be unambiguously fitted into the electron density. The second of these classes could fit a LFY–DBD dimer, while in the first class there was density only for the LFY–DBD molecule that directly interacts with UFO (Extended Data Fig. 9d). The unsharpened maps of each reconstruction were used for post-processing with DeepEMhancer⁵⁵. The figures were prepared with Chimera⁵⁶ or ChimeraX⁵⁷.

Cryo-EM model building

Ideal B-form DNA was generated in Coot⁵⁸ and then manually built into the electron density. The resulting model was further refined using phenix.real_space_refine⁵⁹. A single monomer of LFY–DBD was manually placed in the electron density, followed by fitting in ChimeraX⁵⁷. The biological LFY–DBD dimer was then downloaded from the RCSB PDB (2VY1)²⁹ and used as a guide to place the second LFY monomer, followed by fitting to density in ChimeraX. AlphaFold models⁶⁰ of ASK1 (uniprot ID: Q39255) and UFO (uniprot ID: Q39090) were both downloaded from the EBI, preprocessed to remove low-confidence regions in phenix.process_predicted_model⁶¹, and then placed manually and fit to density in ChimeraX.

Bioinformatic analyses

Read mapping and peak calling. Read processing and peak calling of LFY, LFY–UFO, LFY^{K249R} and LFY^{K249R}–UFO ampDAP-seq data were performed as previously published⁶². Briefly, the quality of sequencing data was analysed with fastQC v.0.11.7, and adapters were removed with NGmerge v.0.2_dev⁶³. Bowtie2 v.2.3.4.1 was used for mapping to the TAIR10 *A. thaliana* reference genome⁶⁴. Reads mapped to a single location and with a maximum of two mismatches were retained. Duplicates were removed with the samtools dedup program v.1.8. Bound regions (that is, peaks) were identified with MACS2 v.2.2.7.1, using input DNA from Lai et al. as a control⁴⁸. Consensus peaks were selected with MSPC v.4.0.0 (ref. ⁶⁵) by retaining peaks called in all replicates and resizing them by ±200 bp around the peak maximum for further analysis.

Analyses of ampDAP-seq experiments. To compare binding in different experiments, peaks were merged according to a previously published procedure⁶². Bound peaks were considered as common if they overlapped by at least 80%, while the remaining non-overlapping portion of either peak was <50%. Peaks that did not overlap by at least 50% were considered as new peaks. The same procedure was used to assess experimental reproducibility (comparisons between replicates of the same experiment), where peaks were normalized by the number of reads mapped in the library (RPKM).

As the fraction of reads mapped in peaks is much lower for LFY than LFY–UFO ampDAP-seq (~25% versus ~40%, respectively), normalizing the read count by all reads mapped along the genome would introduce a bias and estimate the LFY relative coverage (RPKM) towards lower values compared with LFY–UFO. In addition to this consideration, experimental proof from EMSAs suggests that UFO does not strongly affect the binding intensity of the complex at canonical LFYBS (which represent most peaks). Hence, the read count at each peak was normalized by the total number of reads mapped within all LFY and LFY–UFO merged peaks. Then, the mean normalized coverage from each experiment, divided by the peak size, was computed for each peak. The same strategy was applied when comparing LFY^{K249R}

and LFY^{K249R}-UFO (Fig. 4b), LFY^{K249R} and LFY (Extended Data Fig. 8h), and LFY, LFY-UFO, LFY^{K249R} and LFY^{K249R}-UFO (Fig. 4c). The CFC was computed on merged peaks as the ratio between the mean normalized peak coverage in LFY-UFO and LFY (Fig. 2d) or the mean normalized coverage in LFY^{K249R}-UFO and LFY^{K249R} (Fig. 4b).

Motif search in bound regions. Merged peaks of the LFY and LFY-UFO datasets were sorted on the basis of decreasing CFC value. The top 600 peaks (that is, the highest CFC values) were used for a motif search using MEME-ChIP v.4.12.0 using the options nmeme, 600; meme-maxsize, 600*1000; meme-nmotifs, 1; dreme-m, 0; and noecho and the JASPAR 2018 core plants non-redundant database⁶⁶. For dLUBS, we used the options meme-minw, 20; meme-maxw, 30; while for mLUBS, we used meme-minw, 16; meme-maxw, 19. To retrieve the LFY motif in Fig. 2e, the 600 LFY ampDAP-seq peaks with the strongest coverage were fed to MEME-ChIP with the options nmeme, 600; meme-nmotifs, 1; meme-minw, 19; meme-maxw, 19; pal.

Receiver operating characteristics analysis. From the dataset of merged peak sets (peaks found in LFY or in LFY-UFO experiments or in both), the peaks were sorted on the basis of decreased CFC value, the top 20% peaks were selected, and among these, the first 600 used for motif determination were excluded to avoid overfitting, for a total of 3,243 final peaks. A negative set of the same size was created using a previously published method, which allows searching for sequences from the *A. thaliana* genome (TAIR10 reference) with the same GC content and genomic origin as the positive set⁶⁷. Both sets were scanned with dLUBS and mLUBS PWMs as well as with the LFY PWM with dependencies as published previously⁶⁸ using an in-house script available on our GitHub page. The receiver operating characteristics plot was then created with the R package plotROC v.2.2.1 (ref. 69).

LFY in dLUBS within LFY-UFO-specific regions versus LFY in LFY-specific regions. To assess whether the scores of LFYBS within dLUBS were comparable to the scores of canonical LFYBS, we used the peaks from the comparison of LFY versus LFY-UFO ampDAP-seq and resized them (± 50 bp around the peak maximum). We used the dLUBS matrix to scan the resized sequences and retained the best site per sequence. We then retrieved sequences corresponding to the dLUBS site and computed the score of the LFYBS present in dLUBS using the LFY PWM⁶⁸. The values obtained in the 20% most LFY-UFO-specific sequences (20% highest CFC) are shown in the box plot. The 20% lowest CFC peaks were scanned with the LFY PWM to generate the box plot in Extended Data Fig. 4f.

Microarray data analysis. Microarray data were retrieved from AtGen-Express⁷⁰ for inflorescence tissue in the *ufo* background (ATGE_52A-C) versus the Col-0 background (ATGE_29A-C). The R package gcrma⁷¹ was used to adjust probe intensities and convert them to expression measures, and then the limma package⁷² was used to fit the model and smooth standard errors. A Benjamini-Hochberg correction was applied to the *P* values, and fold change (FC) was computed as the ratio between expression in the WT and that in the *ufo* mutant. Only genes with $|\log_2(\text{FC})| > 0.5$ and adjusted *P* < 0.05 were considered as significantly differentially expressed.

ChIP-seq datasets and analysis of ChIP-seq versus ampDAP-seq. We collected the raw data of all available LFY ChIP-seq datasets: GSE141704 (ref. 73), GSE96806 (ref. 25), GSE64245 (ref. 26) and GSE24568 (ref. 68). Mapping and peak-calling analysis were performed with the same procedure as for ampDAP-seq, except that the peaks were resized to 600 bp around the peak maximum, and the q option of MACS2 was set to 0.1. Coverage of the resulting peaks was calculated as the average of the normalized read coverage for each replicate. Peaks from the four datasets were merged through a four-way comparison following the

same procedure used for ampDAP-seq. Bedtools intersect (v.2.30.0)⁷⁴ was used with the options wa; f, 0.8; F, 0.8; and e to find the peaks common to the merged ChIP-seq peaks and the 20% most LFY-UFO-specific genomic regions (the highest CFC value from ampDAP-seq). The peaks were assigned to genes by extending gene regions 3 kb upstream of the transcription start site and 1 kb downstream of the transcription termination site and using bedtools intersect (options f, 0.8; F, 0.8; e). The bound genes obtained were crossed with the list of differentially expressed genes in *ufo* inflorescences.

Identification of the URM from published LFY ChIP-seq data. To test whether the URM could be identified de novo (Extended Data Fig. 4g), we collected the 298 regions bound by LFY ChIP-seq data from inflorescence tissue²⁵ for which the binding intensity was twice greater in vivo relative to in vitro (LFY ampDAP-seq). We resized these regions to ± 55 bp around the ChIP-seq peak maximum. The corresponding sequences were searched with the LFY PWM⁶⁸ to identify all LFYBS with a PWM score greater than -23. Assuming that a recruiting motif should be at a fixed distance from the LFYBS, we created 140 batches, corresponding to sequences with sizes ranging from 4 to 10 bp, distant from 1 to 20 bp at both sides of the canonical LFYBS. Each of the 140 batches of sequences was used as input with MEME-ChIP for motif discovery with the motif size constrained to the length of the sequences in a given batch.

Statistics and reproducibility

All DLRA data were analysed using RStudio software⁷⁵ and are presented as mean $\pm$ s.d. All statistical methods are indicated in the figure legends. One-way ANOVA was used to analyse experimental data with more than two experimental groups (with two-sided Tukey's multiple comparisons tests). Welch's ANOVA was performed when the homogeneity-of-variance assumption was not met (with two-sided Games-Howell post hoc tests). For Fig. 1a,b, the strong promoter activation by 3xHA-LFY-VP16 + UFO-3xFLAG skewed the model and did not allow us to analyse other differences; a log-transformation was applied to the data before performing ANOVA. Two-tailed unpaired Student's *t*-tests were used for the other data analyses. For the GUS experiments and plant complementation assays, two-sided χ^2 tests were used to test for independency between constructs and measured phenotypes. The raw data and exact *P* values are provided in the Source data files as well as the number of independent repetitions for each experiment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The ampDAP-seq data have been deposited at GEO and are publicly available as of the date of publication (GSE204793). The cryo-EM structure determined in this study is deposited in the EM data bank under the reference number EMD-15145. The .pdb file of the model is available in the Supplementary Information. Any additional information required to reanalyse the data reported in this paper is available from the corresponding author upon request. The biological materials generated in this study are available from the corresponding author without restriction. Source data are provided with this paper.

Code availability

All original code has been deposited at GitHub (https://github.com/Bioinfo-LPCV-RDF/LFYUFO_project) and is publicly available as of the date of publication.

References

1. Moyroud, E., Kusters, E., Monniaux, M., Koes, R. & Parcy, F. LEAFY blossoms. *Trends Plant Sci.* **15**, 346–352 (2010).

2. Irish, V. F. The flowering of *Arabidopsis* flower development. *Plant J.* **61**, 1014–1028 (2010).
3. Parcy, F., Nilsson, O., Busch, M. A., Lee, I. & Weigel, D. A genetic framework for floral patterning. *Nature* **395**, 561–566 (1998).
4. Wagner, D., Sablowski, R. W. M. & Meyerowitz, E. M. Transcriptional activation of APETALA1 by LEAFY. *Science* **285**, 582–584 (1999).
5. Lohmann, J. U. et al. A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. *Cell* **105**, 793–803 (2001).
6. Lee, I., Wolfe, D. S., Nilsson, O. & Weigel, D. A LEAFY co-regulator encoded by UNUSUAL FLORAL ORGANS. *Curr. Biol.* **7**, 95–104 (1997).
7. Levin, J. Z. & Meyerowitz, E. M. UFO: an *Arabidopsis* gene involved in both floral meristem and floral organ development. *Plant Cell* **7**, 529–548 (1995).
8. Wilkinson & Haughn UNUSUAL FLORAL ORGANS controls meristem identity and organ primordia fate in *Arabidopsis*. *Plant Cell* **7**, 1485–1499 (1995).
9. Krizek, B. A. & Meyerowitz, E. M. The *Arabidopsis* homeotic genes APETALA3 and PISTILLATA are sufficient to provide the B class organ identity function. *Development* **122**, 11–22 (1996).
10. Ikeda-Kawakatsu, K., Maekawa, M., Izawa, T., Itoh, J.-I. & Nagato, Y. ABERRANT PANICLE ORGANIZATION 2/RFL, the rice ortholog of *Arabidopsis* LEAFY, suppresses the transition from inflorescence meristem to floral meristem through interaction with APO1. *Plant J.* **69**, 168–180 (2012).
11. Lippman, Z. B. et al. The making of a compound inflorescence in tomato and related nightshades. *PLoS Biol.* **6**, e288 (2008).
12. Souer, E. et al. Patterning of inflorescences and flowers by the F-box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of petunia. *Plant Cell Online* **20**, 2033–2048 (2008).
13. Kuzay, S. et al. WAPO-A1 is the causal gene of the 7AL QTL for spikelet number per spike in wheat. *PLoS Genet.* **18**, e1009747 (2022).
14. Ingram, G. C. et al. Dual role for fimbriata in regulating floral homeotic genes and cell division in *Antirrhinum*. *EMBO J.* **16**, 6521–6534 (1997).
15. Samach, A. et al. The UNUSUAL FLORAL ORGANS gene of *Arabidopsis thaliana* is an F-box protein required for normal patterning and growth in the floral meristem. *Plant J.* **20**, 433–445 (1999).
16. Simon, R., Carpenter, R., Doyle, S. & Coen, E. Fimbriata controls flower development by mediating between meristem and organ identity genes. *Cell* **78**, 99–107 (1994).
17. Wang, X. et al. The COP9 signalosome interacts with SCF UFO and participates in *Arabidopsis* flower development. *Plant Cell* **15**, 1071–1082 (2003).
18. Chae, E., Tan, Q. K.-G., Hill, T. A. & Irish, V. F. An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. *Development* **135**, 1235–1245 (2008).
19. Geng, F., Wenzel, S. & Tansey, W. P. Ubiquitin and proteasomes in transcription. *Annu. Rev. Biochem.* **81**, 177–201 (2012).
20. Risseuw, E. et al. An activated form of UFO alters leaf development and produces ectopic floral and inflorescence meristems. *PLoS ONE* **8**, e83807 (2013).
21. Krizek, B. A., Lewis, M. W. & Fletcher, J. C. RABBIT EARS is a second-whorl repressor of AGAMOUS that maintains spatial boundaries in *Arabidopsis* flowers. *Plant J.* **45**, 369–383 (2006).
22. Busch, M. A., Bomblies, K. & Weigel, D. Activation of a floral homeotic gene in *Arabidopsis*. *Science* **285**, 585–587 (1999).
23. Hill, T. A., Day, C. D., Zondlo, S. C., Thackeray, A. G. & Irish, V. F. Discrete spatial and temporal cis-acting elements regulate transcription of the *Arabidopsis* floral homeotic gene APETALA3. *Development* **125**, 1711–1721 (1998).
24. Lamb, R. S., Hill, T. A., Tan, Q. K.-G. & Irish, V. F. Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. *Development* **129**, 2079–2086 (2002).
25. Goslin, K. et al. Transcription factor interplay between LEAFY and APETALA1/CAULIFLOWER during floral initiation. *Plant Physiol.* **174**, 1097–1109 (2017).
26. Sayou, C. et al. A SAM oligomerization domain shapes the genomic binding landscape of the LEAFY transcription factor. *Nat. Commun.* **7**, 11222 (2016).
27. Weigel, D., Alvarez, J., Smyth, D. R., Yanofsky, M. F. & Meyerowitz, E. M. LEAFY controls floral meristem identity in *Arabidopsis*. *Cell* **69**, 843–859 (1992).
28. Weigel, D. & Nilsson, O. A developmental switch sufficient for flower initiation in diverse plants. *Nature* **377**, 495–500 (1995).
29. Hamès, C. et al. Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *EMBO J.* **27**, 2628–2637 (2008).
30. Sayou, C. et al. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* **343**, 645–648 (2014).
31. Zhao, D., Yu, Q., Chen, M. & Ma, H. The ASK1 gene regulates B function gene expression in cooperation with UFO and LEAFY in *Arabidopsis*. *Development* **128**, 2735–2746 (2001).
32. Ni, W. et al. Regulation of flower development in *Arabidopsis* by SCF complexes. *Plant Physiol.* **134**, 1574–1585 (2004).
33. Levin, J. Z., Fletcher, J. C., Chen, X. & Meyerowitz, E. M. A genetic screen for modifiers of UFO meristem activity identifies three novel FUSED FLORAL ORGANS genes required for early flower development in *Arabidopsis*. *Genetics* **149**, 579–595 (1998).
34. Singh, N. & Bhalla, N. Moonlighting Proteins. *Annu. Rev. Genet.* **54**, 265–285 (2020).
35. Honma, T. & Goto, K. The *Arabidopsis* floral homeotic gene PISTILLATA is regulated by discrete cis-elements responsive to induction and maintenance signals. *Development* **127**, 2021–2030 (2000).
36. Tilly, J. J., Allen, D. W. & Jack, T. The CARG boxes in the promoter of the *Arabidopsis* floral organ identity gene APETALA3 mediate diverse regulatory effects. *Development* **125**, 1647–1657 (1998).
37. Liu, C., Xi, W., Shen, L., Tan, C. & Yu, H. Regulation of floral patterning by flowering time genes. *Dev. Cell* **16**, 711–722 (2009).
38. Gregis, V., Sessa, A., Colombo, L. & Kater, M. M. AGL24, SHORT VEGETATIVE PHASE, and APETALA1 redundantly control AGAMOUS during early stages of flower development in *Arabidopsis*. *Plant Cell* **18**, 1373–1382 (2006).
39. Castillejo, C., Romera-Branchat, M. & Pelaz, S. A new role of the *Arabidopsis* SEPALLATA3 gene revealed by its constitutive expression. *Plant J.* **43**, 586–596 (2005).
40. Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. & Bulky, M. L. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 555 (2011).
41. Babb, R., Huang, C., Aufiero, D. J. & Herr, W. DNA recognition by the herpes simplex virus transactivator VP16: a novel DNA-binding structure. *Mol. Cell. Biol.* **21**, 4700–4712 (2001).
42. Chahtane, H. et al. LEAFY activity is post-transcriptionally regulated by BLADE ON PETIOLE2 and CULLIN3 in *Arabidopsis*. *N. Phytol.* **220**, 579–592 (2018).
43. Blanvillain, R. et al. The *Arabidopsis* peptide kiss of death is an inducer of programmed cell death. *EMBO J.* **30**, 1173–1183 (2011).
44. Takeda, S., Matsumoto, N. & Okada, K. RABBIT EARS, encoding a SUPERMAN-like zinc finger protein, regulates petal development in *Arabidopsis thaliana*. *Development* **131**, 425–434 (2004).

45. Benlloch, R. et al. Integrating long-day flowering signals: a LEAFY binding site is essential for proper photoperiodic activation of APETALA1. *Plant J.* **67**, 1094–1102 (2011).
46. Iwata, Y., Lee, M. H. & Koizumi, N. Analysis of a transcription factor using transient assay in *Arabidopsis* protoplasts. *Methods Mol. Biol.* **754**, 107–117 (2011).
47. Dümmler, A., Lawrence, A. M. & de Marco, A. Simplified screening for the detection of soluble fusion constructs expressed in *E. coli* using a modular set of vectors. *Microb. Cell Fact.* **4**, 34 (2005).
48. Lai, X. et al. The LEAFY floral regulator displays pioneer transcription factor properties. *Mol. Plant* **14**, 829–837 (2021).
49. Bartlett, A. et al. Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* **12**, 1659–1672 (2017).
50. Cutler, S. R., Ehrhardt, D. W., Griffiths, J. S. & Somerville, C. R. Random GFP::cDNA fusions enable visualization of subcellular structures in cells of *Arabidopsis* at a high frequency. *Proc. Natl Acad. Sci. USA* **97**, 3718–3723 (2000).
51. Chahtane, H. et al. A variant of LEAFY reveals its capacity to stimulate meristem development by inducing RAX1. *Plant J.* **74**, 678–689 (2013).
52. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
53. Wagner, T. et al. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Commun. Biol.* **2**, 218 (2019).
54. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020).
55. Sanchez-Garcia, R. et al. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* **4**, 874 (2021).
56. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
57. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
58. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
59. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D Struct. Biol.* **74**, 531–544 (2018).
60. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
61. Terwilliger, T. C. et al. Improved AlphaFold modeling with implicit experimental information. *Nat. Methods* **19**, 1376–1382 (2022).
62. Lai, X. et al. Genome-wide binding of SEPALLATA3 and AGAMOUS complexes determined by sequential DNA-affinity purification sequencing. *Nucleic Acids Res.* **48**, 9637–9648 (2020).
63. Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinform.* **19**, 536 (2018).
64. Berardini, T. Z. et al. The Arabidopsis Information Resource: making and mining the ‘gold standard’ annotated reference plant genome. *Genesis* **53**, 474–485 (2015).
65. Jalili, V., Matteucci, M., Masseroli, M. & Morelli, M. J. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics* **31**, 2761–2769 (2015).
66. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
67. Stigliani, A. et al. Capturing auxin response factors syntax using DNA binding models. *Mol. Plant* **12**, 822–832 (2019).
68. Moyroud, E. et al. Prediction of regulatory interactions from genome sequences using a biophysical model for the *Arabidopsis* LEAFY transcription factor. *Plant Cell* **23**, 1293–1306 (2011).
69. Sachs, M. C. plotROC: A Tool for Plotting ROC Curves. *J. Stat. Softw.* **79**, 1–19 (2017).
70. Schmid, M. et al. A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506 (2005).
71. Wu J. et al. gcrma: Background Adjustment Using Sequence Information. R package version 2.70.0. (2022). <https://bioconductor.org/packages/release/bioc/manuals/gcrma/man/gcrma.pdf>
72. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
73. Jin, R. et al. LEAFY is a pioneer transcription factor and licenses cell reprogramming to floral fate. *Nat. Commun.* **12**, 626 (2021).
74. Quinlan, A.R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
75. RStudio Team RStudio: Integrated Development for R. (RStudio, PBC, Boston, MA, 2020). <http://www.rstudio.com/>
76. Gagne, J. M., Downes, B. P., Shiu, S. H., Durski, A. M. & Vierstra, R. D. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **99**, 11519–11524 (2002).
77. Zhang, S. et al. 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 (2003).
78. Zhao, Y. et al. Evolutionary co-option of floral meristem identity genes for patterning of the flower-like Asteraceae inflorescence. *Plant Physiol.* **172**, 284–296 (2016).
79. Chen, Y. et al. CsUFO is involved in the formation of flowers and tendrils in cucumber. *Theor. Appl. Genet.* **3**, 2141–2150 (2021).
80. Ikeda, K., Ito, M., Nagasawa, N., Kyoizuka, J. & Nagato, Y. Rice ABERRANT PANICLE ORGANIZATION 1, encoding an F-box protein, regulates meristem fate. *Plant J.* **51**, 1030–1040 (2007).
81. Li, F. et al. Reduced expression of CbUFO is associated with the phenotype of a flower-defective *Cosmos bipinnatus*. *Int. J. Mol. Sci.* **20**, 2503 (2019).
82. Sasaki, K. et al. Mutation in *Torenia fournieri* Lind. UFO homolog confers loss of TtLFY interaction and results in a petal to sepal transformation. *Plant J.* **71**, 1002–1014 (2012).
83. Sharma, B. et al. Homologs of LEAFY and UNUSUAL FLORAL ORGANS promote the transition from inflorescence to floral meristem identity in the cymose *Aquilegia coerulea*. *Front. Plant Sci.* **10**, 1218 (2019).
84. Taylor, S., Hofer, J. & Murfet, I. Stamina pistilloida, the pea ortholog of Fim and UFO, is required for normal development of flowers, inflorescences, and leaves. *Plant Cell* **13**, 31–46 (2001).
85. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).

Acknowledgements

We thank A. M. Boisson for preparing the suspension cells, X. Lai for the ampDAP-seq libraries and technical assistance and R. Koes for sharing data and materials. We acknowledge C. Marondedze, G. Vachon, M. Le Masson, C. Berthollet, B. Orlando Marchesano and J. Bourenane-Vieira for help with the experiments. We thank G. Vert, U. Dolde and R. Dumas for discussion. The electron microscopy facility is supported by the Rhône-Alpes Region, the FRM, the FEDER and the GIS-IBISA. This work used the platforms of the Grenoble Instruct-ERIC

centre (ISBG; UAR 3518 CNRS-CEA-UGA-EMBL) within the Grenoble Partnership for Structural Biology, supported by FRISBI (ANR-10-INBS-0005-02). We thank C. Mas for assistance and access to the biophysical platform. This work was supported by the GRAL Labex financed within the University Grenoble Alpes graduate school (Ecoles Universitaires de Recherche) CBH-EUR-GS (ANR-17-EURE-0003), the CEA (PhD fellowship to P.R.) and the ANR-17-CE20-0014-01 Ubiflor project to F.P.

Author contributions

F.P. and P.R. designed the project. P.R. performed the plant experiments with assistance from G.T. P.R. and E.T. performed the biochemical experiments with assistance from H.C. on the evolutionary analyses. L.T. performed the bioinformatics analyses with assistance from J.L. and R.B.-M. E.Z. and G.S. performed the cryo-EM experiments, and M.N., E.Z., C.Z. and G.S. analysed the data. P.R. and L.T. assembled the figures. P.R. and F.P. wrote the paper with contributions from L.T. and C.Z.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41477-022-01336-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41477-022-01336-2>.

Correspondence and requests for materials should be addressed to François Parcy.

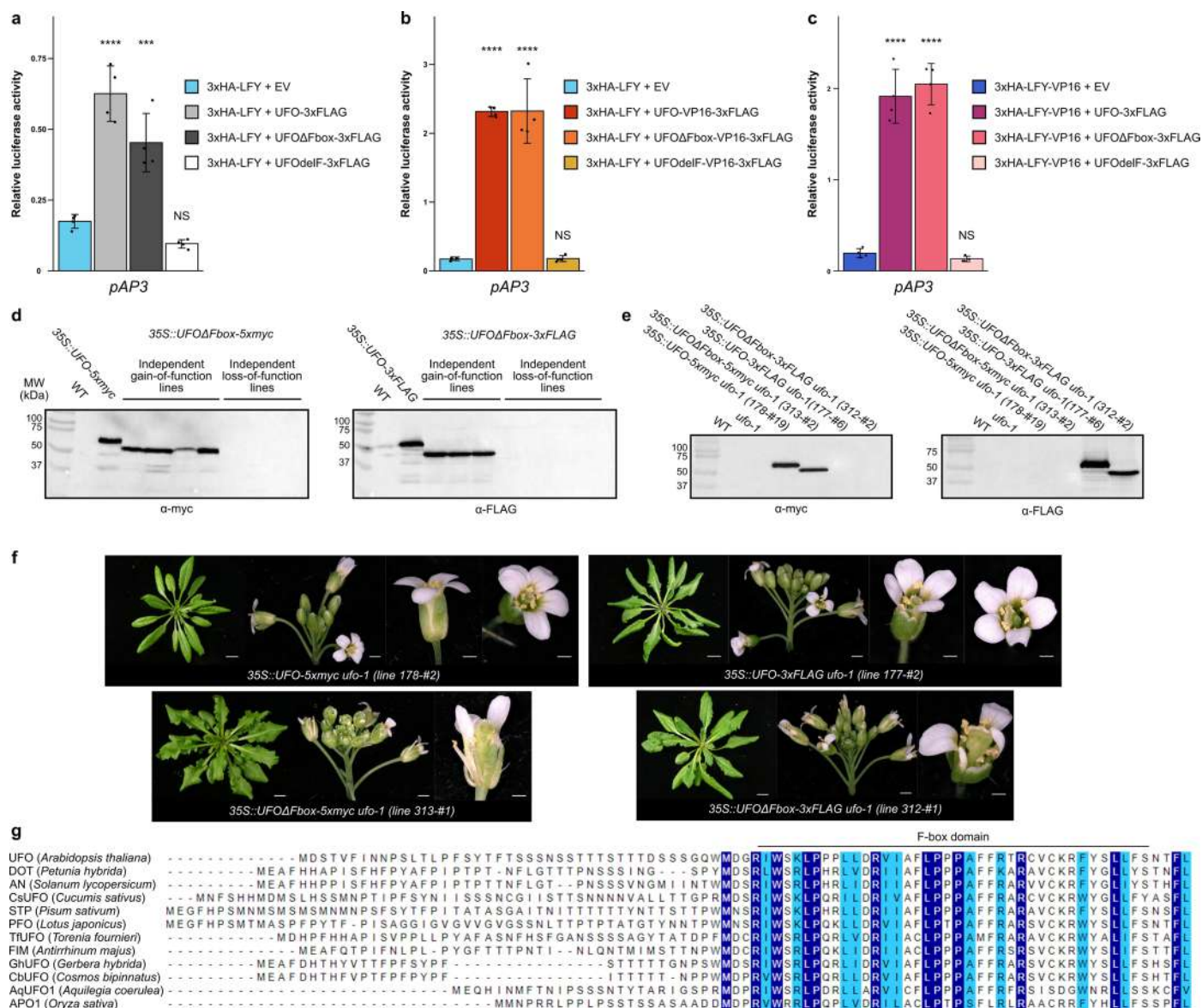
Peer review information *Nature Plants* thanks Nobutoshi Yamaguchi, Aiwu Dong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

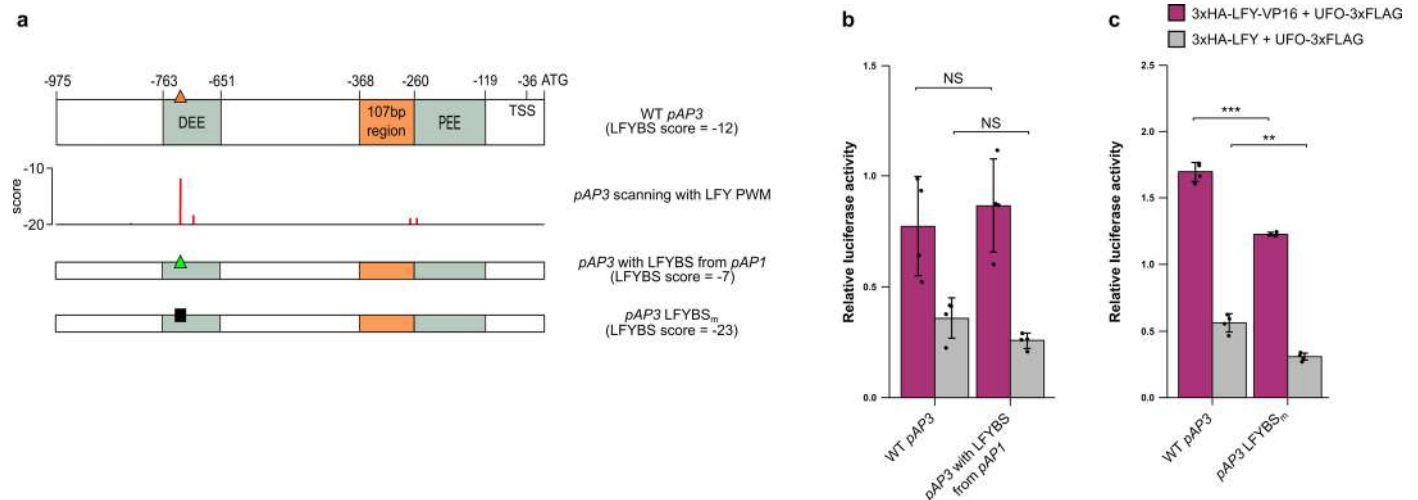
© The Author(s), under exclusive licence to Springer Nature Limited 2023



Extended Data Fig. 1 | UFO has SCF-dependent and independent functions.

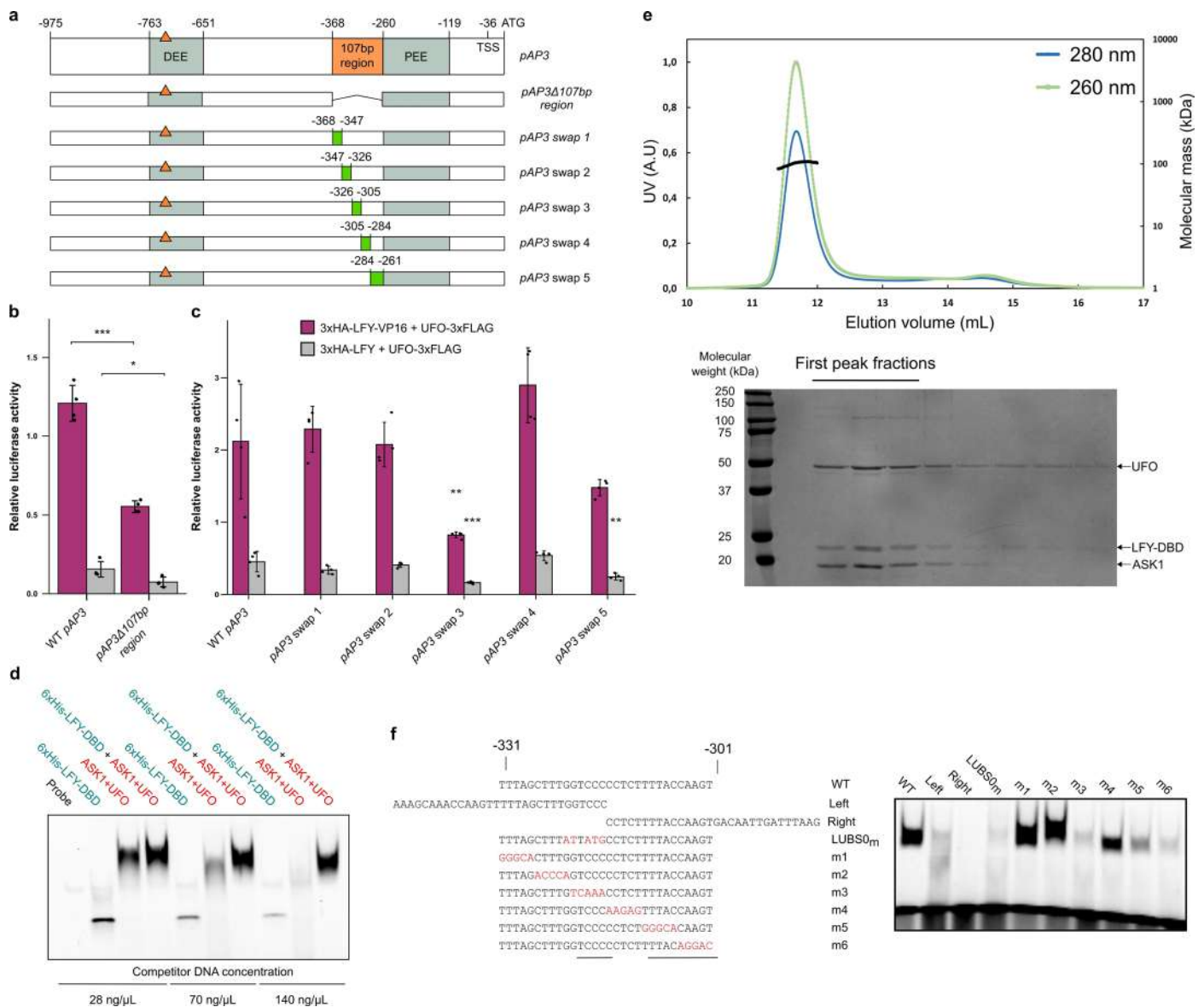
a-c, pAP3 activation measured by DLRA in Arabidopsis protoplasts. EV = Empty Vector (pRT104-3xHA). UFOΔFbox corresponds to a deletion of the whole N-terminal part comprising the F-box domain (aa. 1-90), while UFOΔFbox corresponds to a previously-described internal deletion in the F-box domain (aa. 50-62)²⁰. Data represent averages of independent biological replicates and are presented as mean ± SD, each dot representing one biological replicate (n = 4). One-way ANOVA with Tukey's multiple comparisons tests. Stars above bars represent a significant statistical difference compared to 3xHA-LFY + EV or 3xHA-LFY-VP16 + EV negative controls (NS: p > 0.05, *: p < 0.05, **: p < 0.01, ***: p < 0.001 and ****: p < 0.0001). **d**, Western Blot on protein extracts from independent T1 plants from different phenotypic classes described in Fig. 1g (one independent line per lane). 35S::UFO-5xmyc (line 178-#19) and 35S::UFO-3xFLAG (line 177-#6) plants were used as positive controls. Total proteins were extracted from rosette leaves. Note the difference of molecular weight between UFO and

UFOΔFbox. Loss-of-function defects are likely due to silencing of both transgene-encoded UFOΔFbox and endogenous UFO. **e**, Western Blot on protein extracts from F2 plants described in Fig. 1h. Total proteins were extracted from rosette leaves. **f**, ufo-1 complementation assay with other 35S::UFO and 35S::UFOΔFbox lines. Rosette leaves (right, scale bar, 1 cm), inflorescence (middle, scale bar 1 mm) and flower (left, scale bar, 0.5 mm) phenotypes are shown. Primary inflorescences were removed to observe rosette phenotype. For each construct, at least 5 plants were analyzed per line. As in Risseuw et al, our 35S::UFO lines displayed relatively milder phenotypes than the 35S::UFO phenotypes reported by Lee et al.^{6,20}. Note that the 35S::UFO-5xmyc 178-#2 line did not display the serrated leaves phenotype. **g**, Sequence alignment of UFO N-terminal region. The F-box domain is represented⁷⁶. In selected species, presented proteins were identified as UFO homologs and their role was confirmed genetically^{7,11,12,16,77-84}. Source data are available in Supplementary Data 4.



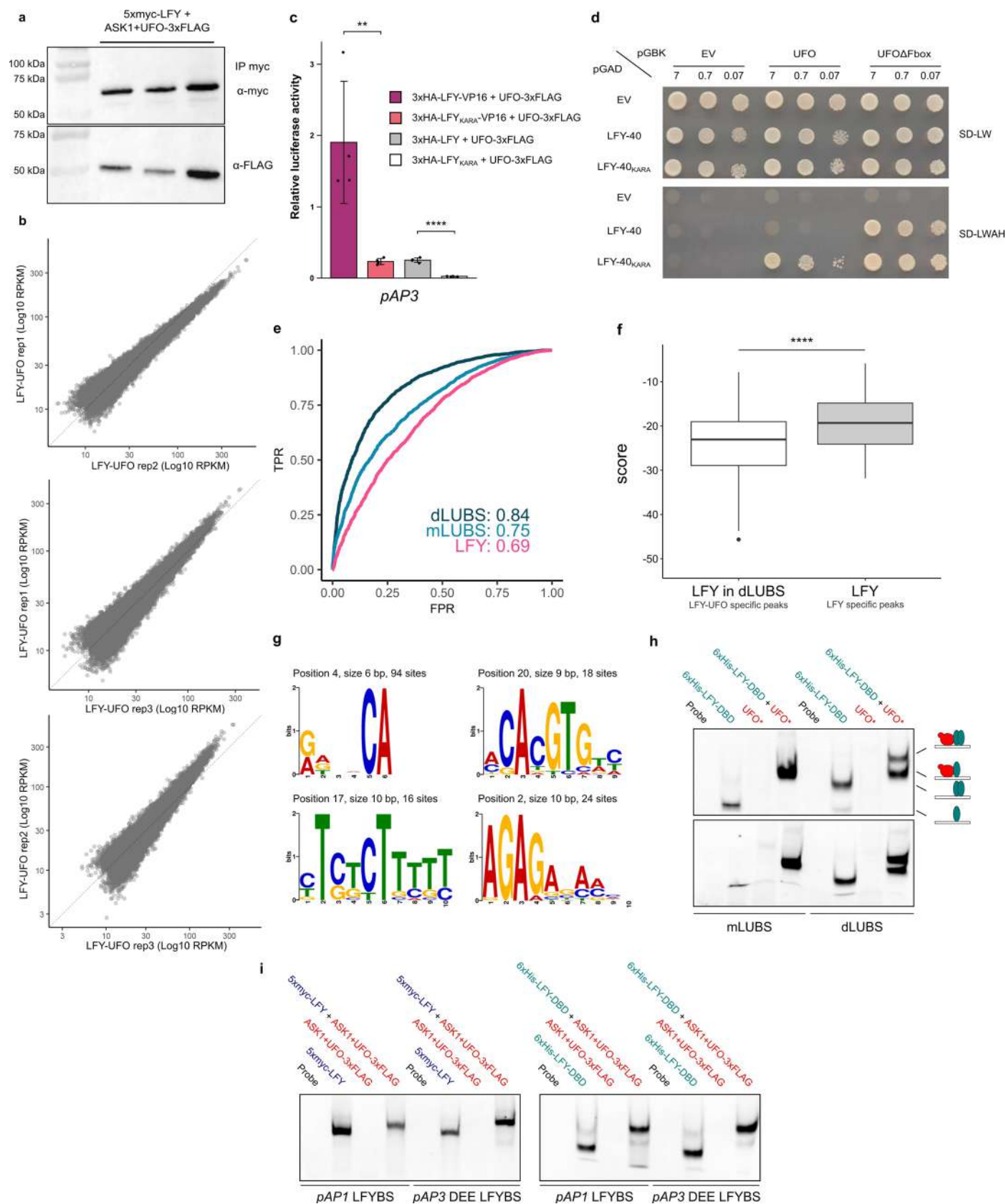
Extended Data Fig. 2 | *pAP3* DEE LFYBS is not required for LFY-UFO-dependent *pAP3* activation. **a**, Schematic representation of *pAP3*. Top row represents WT *pAP3* with regulatory regions and *cis*-elements. Orange triangle represents LFYBS. The second row represents the scores for the best LFYBS obtained by scanning WT *pAP3* sequence with LFY PWM⁶⁸ (the best binding sites correspond to the less negative score values). Other rows represent the different *pAP3* versions used in **(b)** and **(c)**. LFYBS mutation corresponds to the previously

described *site1m-site2m* mutation²⁴. **b,c**, *pAP3* activation with promoter versions described in **(a)** and indicated effectors. For bar charts, data represent averages of independent biological replicates and are presented as mean $\pm$ SD, each dot representing one biological replicate ($n = 4$). Unpaired t-tests **(b,c)**. (NS: $p > 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$). Source data are available in Supplementary Data 4.



Extended Data Fig. 3 | Analysis of *pAP3* activation by LFY-UFO. a, Description of *pAP3*. Top line represents WT *pAP3* with regulatory regions and *cis*-elements. Coordinates are relative to *AP3* start codon. TSS: Transcription Start Site. Orange triangle represents LFYBS. Other rows show the promoter versions used in (b) and (c). Green rectangles in swapped versions correspond to the same random sequence. **b, c**, *pAP3* LFY-UFO response element mapping with *pAP3* versions described in (a) by DLRA in Arabidopsis protoplasts. Data represent averages of independent biological replicates and are presented as mean ± SD, each dot representing one biological replicate (n = 4). One-way ANOVA with Tukey's multiple comparisons test (c). One-way ANOVA was performed with data from the same effector, and stars represent a statistical difference compared to WT *pAP3*. Unpaired t-tests (b). (NS: p > 0.05; *, p < 0.05; **, p < 0.01; ***, p < 0.001). **d**, EMSA with ASK1-UFO, LFY-DBD and LUBSO DNA probe. Different competitor DNA concentrations were tested as indicated. **e**, Molecular mass determination for ASK1-UFO-LFY-DBD in complex with LUBSO DNA by SEC-MALLS (top). Elution

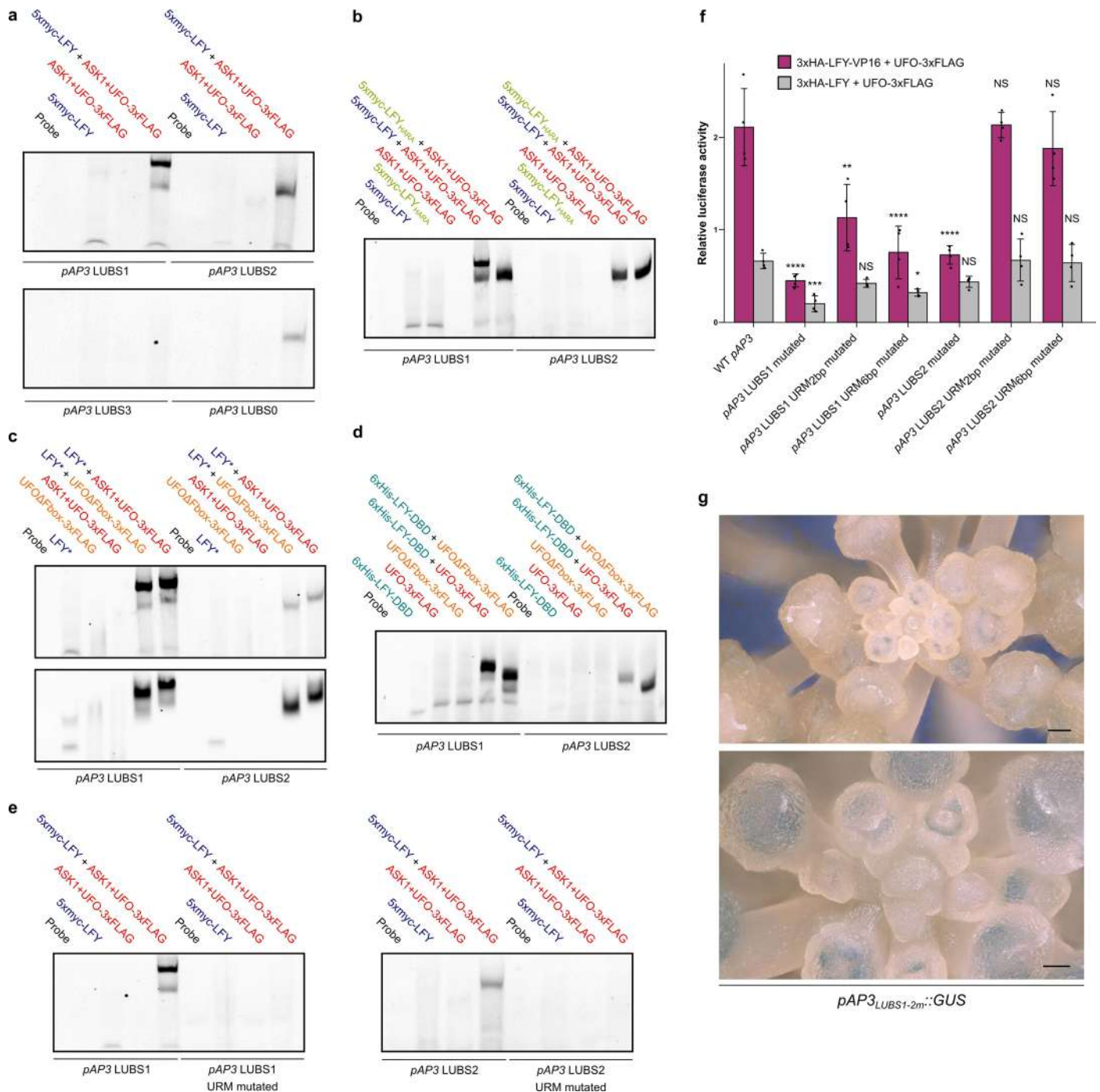
profiles correspond to absorbance at 280 nm and 260 nm (left ordinate axis, A.U: Arbitrary Unit). The black line shows the molecular mass distribution (right ordinate axis). A mass of 102 ± 3.3 kDa was found for this ASK1-UFO-LFY-DBD-LUBSO complex, consistent with one copy of each protein per DNA molecule (theoretical mass of 108 kDa). Coomassie-stained SDS-PAGE gel of the different SEC-MALLS fractions (bottom). Each lane corresponds to a 0.5 mL fraction. Molecular weights of the protein standards are indicated (BioRad Precision Plus). Faint bands above UFO likely correspond to contaminants. **f**, EMSA with ASK1-UFO, LFY-DBD and indicated DNA probes. Sequences with coordinates relative to *AP3* start codon (left). Red letters indicate mutated bases. Bars under sequences represent the regions required for ASK1-UFO-LFY-DBD binding. EMSA with described DNA probes (right). Each DNA probe was mixed with the same ASK1-UFO-LFY-DBD protein mix. Note that the LUBSO mutation also reduced *pAP3* activation in protoplasts (Fig. 2b). Source data are available in Supplementary Data 4.



Extended Data Fig. 4 | See next page for caption.

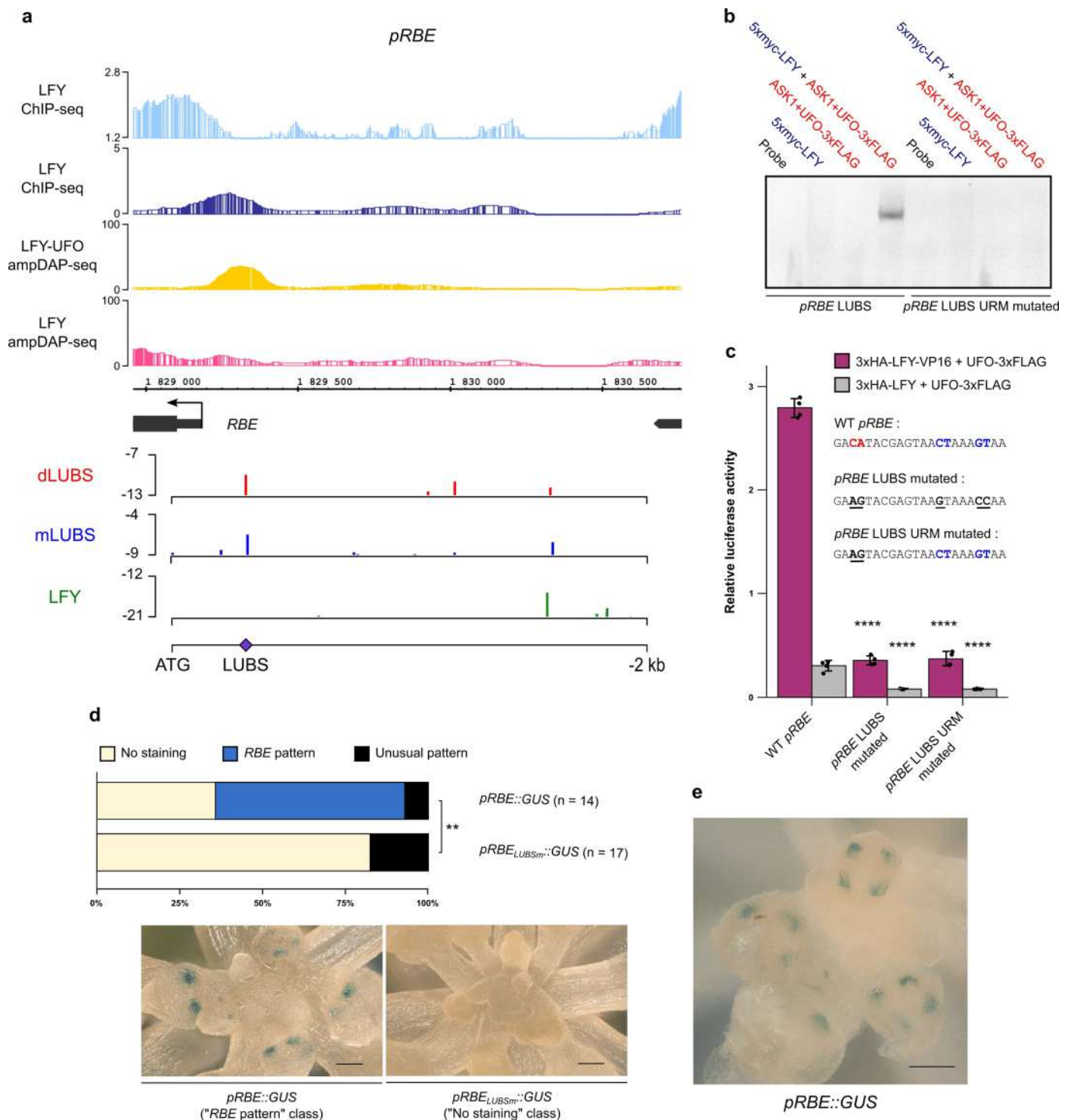
Extended Data Fig. 4 | Genome-wide analysis of LFY-UFO binding. **a**, Western Blot after DNA elution during ampDAP-seq experiment. After DNA elution, 20 μ L of 1X SDS-PAGE Protein Sample Buffer was added to the remaining beads to run WB. Each lane represents one replicate. **b**, Assessment of experimental reproducibility of ampDAP-seq experiment through the comparison of replicates datasets 2 by 2. **c**, Effect of the LFY KARA mutation (K303A-R233A)⁵¹ on *pAP3* activation in Arabidopsis protoplasts. Data represent averages of independent biological replicates and are presented as mean $\pm$ SD, each dot representing one biological replicate (n = 4). Unpaired t-tests (**: p < 0.01; ****: p < 0.0001). **d**, The LFY KARA mutation (K303A-R233A) does not disrupt LFY-UFO interaction in Yeast-Two-Hybrid (Y2H). EV = Empty Vector. LFY-40 is a LFY version lacking the first 40 aa and better tolerated by yeast cells. Values correspond to the different dilutions (OD = 7, 0.7 and 0.07). Top picture corresponds to the non-selective plate lacking Leucine and Tryptophan (SD -L-W), and bottom picture to the selective plate lacking Leucine, Tryptophan, Histidine and Adenine (SD -L-W-A-H). Pictures were taken at day + 4. **e**, Receiver operating characteristics (ROC) curves for mLUBS, dLUBS and LFY using the top 20% high-CFC LFY-UFO-specific peaks. Area under the curve (AUC) values are shown. TPR: True Positive Rate, FPR: False Positive Rate. **f**, Score distribution of LFY PWM with dependencies⁶⁸ within dLUBS (best site on 20% most LFY-UFO-specific genomic regions, high

CFC, n = 3843 genomic regions) and in canonical LFYBS (best site on 20% most LFY-specific genomic regions, low CFC, n = 3843 genomic regions). Best sites were selected within ± 25 bp around the peak maximum. Wilcoxon rank sum test (****: p < 0.0001). Median (solid line), interquartile range (box edges), $\pm 1.5 \times$ interquartile range (whiskers) and outliers (black dot) are shown. **g**, *De novo* identification of URM from LFY ChIP-seq data²⁵. Motifs identified at a fixed distance from LFY canonical binding sites in 298 regions harboring high LFY ChIP-seq to LFY ampDAP-seq coverage ratio. The text above each motif gives the motif's start position relative to the canonical LFYBS, its length and the number of sites used to build the motif. **h**, EMSA with mLUBS and dLUBS highest score sequences. 6xHis-LFY-DBD is recombinant. UFO* refers to either recombinant ASK1-UFO-3xFLAG complex (top gel) or *in vitro* produced UFO-3xFLAG (bottom gel). Drawings represent the different types of complexes involving LFY-DBD (pale blue) and ASK1-UFO (red) on DNA. LFY-DBD binds as a monomer as previously reported²⁹. The fact that *in vitro* produced UFO-3xFLAG shifts DNA in the presence of LFY indicates that ASK1 is not required for the UFO-LFY-DNA complex formation *in vitro*. **i**, EMSA with DNA probes corresponding to *pAP1* and *pAP3* DEE LFYBS and indicated proteins. Note that probes used here have the same length as those used to study LUBS. Source data are available in Supplementary Data 4.



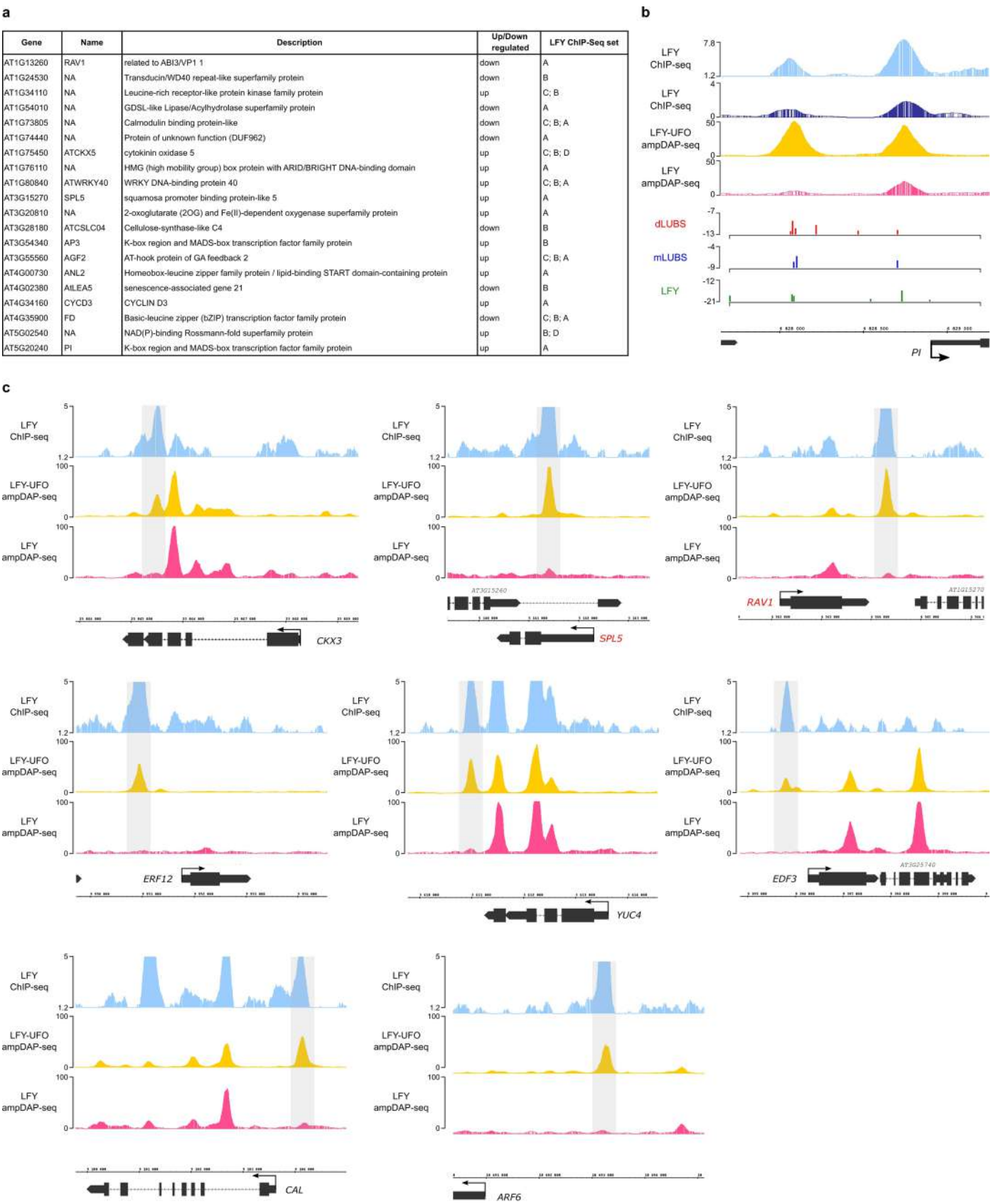
Extended Data Fig. 5 | *pAP3* LUBS are required for LFY-UFO-dependent activation. **a**, EMSA with indicated probes and proteins. LUBS3 is the third highest-score *pAP3* LUBS. Because LUBS0 is bound with a lower affinity by LFY-UFO compared to LUBS1 and LUBS2, we then focused on LUBS1 and LUBS2. **b**, EMSA with *pAP3* LUBS1 and LUBS2 DNA probes and indicated proteins. LFY^{H383A-R386A} (LFY^{HARA}) is a LFY mutated version affected in its ability to dimerize^{29,51}. Note the absence of the complex with a slower mobility on LUBS1 with LFY^{HARA}. **c**, EMSA with *pAP3* LUBS1 and LUBS2 DNA probes and indicated proteins. LFY* refers to either *in vitro*-produced 5xmyc-LFY (top) or recombinant 6xHis-LFY-DBD (bottom). Note the difference of complex size between UFO and UFOΔFbox. **d**, Same as in (c) except that UFO-3xFLAG and UFOΔFbox-3xFLAG were produced *in vitro*. Note that *in vitro* produced UFO-3xFLAG and UFOΔFbox-3xFLAG behave similarly as recombinant UFO versions. **e**, EMSA with indicated proteins and DNA probes corresponding to *pAP3* LUBS1 (left) and LUBS2 (right), WT or with URM mutated. **f**, Promoter activation measured by

DLRA in Arabidopsis protoplasts with indicated effectors. Different promoter versions were tested as indicated under x-axis. Either 2 bp (high-informative CA) or 6 bp (whole URM) of *pAP3* LUBS1 and LUBS2 URM were mutated. Data represent averages of independent biological replicates and are presented as mean ± SD, each dot representing one biological replicate (n = 4). One-way ANOVA with Tukey's multiple comparisons tests. One-way ANOVA were performed with data from the same effector and stars represent a statistical difference compared to WT *pAP3* promoter. (NS: p > 0.05, *: p < 0.05, **: p < 0.01, ***: p < 0.001 and ****: p < 0.0001). **g**, *In vivo* analysis of *pAP3*_{LUBS1-2m}::*GUS* fusions. Same as in Fig. 3d, except that staining incubation time was increased to 17 h (4 h incubation in Fig. 3d). Representative pictures are shown (top scale bar, 100 μm, bottom scale bar, 50 μm). The faint *AP3* pattern suggests that other LUBS (such as LUBS0) may take over but less efficiently. Note that with this staining incubation time, all plants expressing *pAP3*::*GUS* showed a highly saturated staining. Source data are available in Supplementary Data 4.



Extended Data Fig. 6 | *pRBE* LUBS is required for LFY-UFO-dependent activation. **a**, IGB view of *pRBE* showing LFY ChIP-seq in inflorescences (light blue)²⁵ or seedlings (dark blue)²⁶, LFY-UFO ampDAP-seq (yellow), LFY ampDAP-seq (pink)⁴⁸, numbers indicate read number range (top). Identification of LUBS in *pRBE* (bottom). Predicted binding sites using dLUBS and mLUBS models from Fig. 2e and LFY PWM with dependencies⁶⁸, y-axis represents score values (bottom). The best binding sites correspond to the less negative score values. Studied LUBS is indicated with a purple square. **b**, EMSA with probes corresponding to *pRBE* LUBS, WT or with URM mutated. **c**, *pRBE* activation in Arabidopsis protoplasts. Effect of mutations (underlined) in URM (red) and in LFYBS (blue) bases of *pRBE* LUBS were assayed. Data represent averages of independent biological replicates and are presented as mean $\pm$ SD,

each dot representing one biological replicate (n = 4). One-way ANOVA with Tukey's multiple comparisons test. One-way ANOVA were performed with data from the same effector, and stars represent a statistical difference compared to WT promoters (****; p < 0.0001). **d**, *In vivo* analysis of *pRBE::GUS* fusions. The percentage of transgenic lines with *RBE* pattern, unusual pattern or absence of staining was scored (top; χ^2 test, **; p < 0.01). n = number of independent lines. Unusual pattern refers to staining in unexpected tissues, each pattern seen in a single line. Representative pictures of plants with no staining (bottom left) and a *RBE* pattern (bottom right) are shown (scale bar, 50 μ m). **e**, *In vivo* analysis of *pRBE::GUS* fusions. Same as in (d), with another view showing staining in the four petal primordia (scale bar, 50 μ m). Source data are available in Supplementary Data 4.

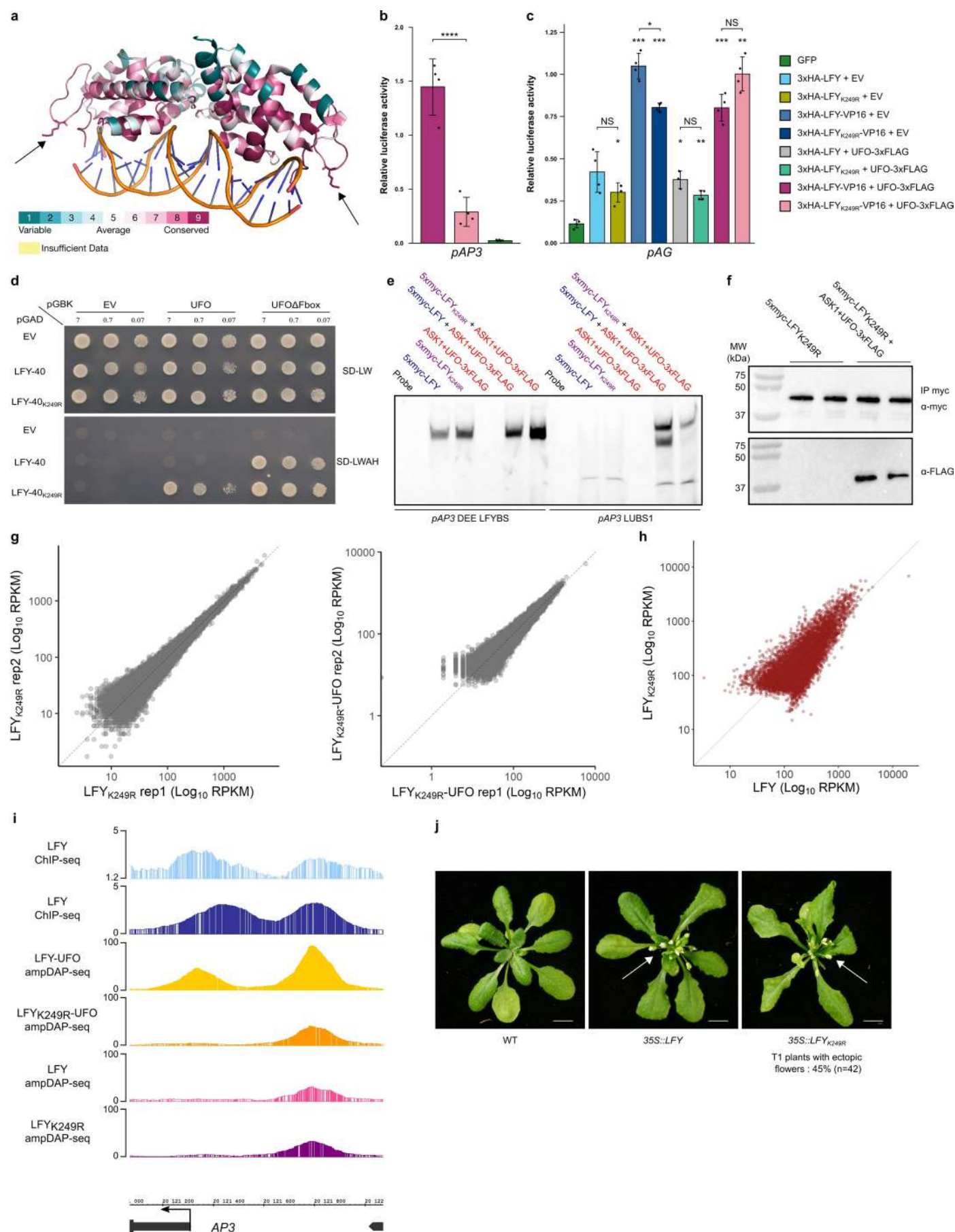


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | LFY and UFO likely regulate other genes in Arabidopsis.

a. List of candidate LFY-UFO target genes selected as i) present in regions specifically bound by LFY-UFO in ampDAP-seq (high CFC) ii) bound *in vivo* in LFY ChIP-seq experiments (A²⁵; B²⁶; C⁶⁸; D⁷³) and iii) deregulated in *ufo* inflorescences⁷⁰. **b.** IGB view of *PISTILLATA* promoter region showing LFY ChIP-seq in inflorescences (light blue)²⁵ or seedlings (dark blue)²⁶, LFY-UFO ampDAP-seq (yellow), LFY ampDAP-seq (pink)⁴⁸, numbers indicate read number

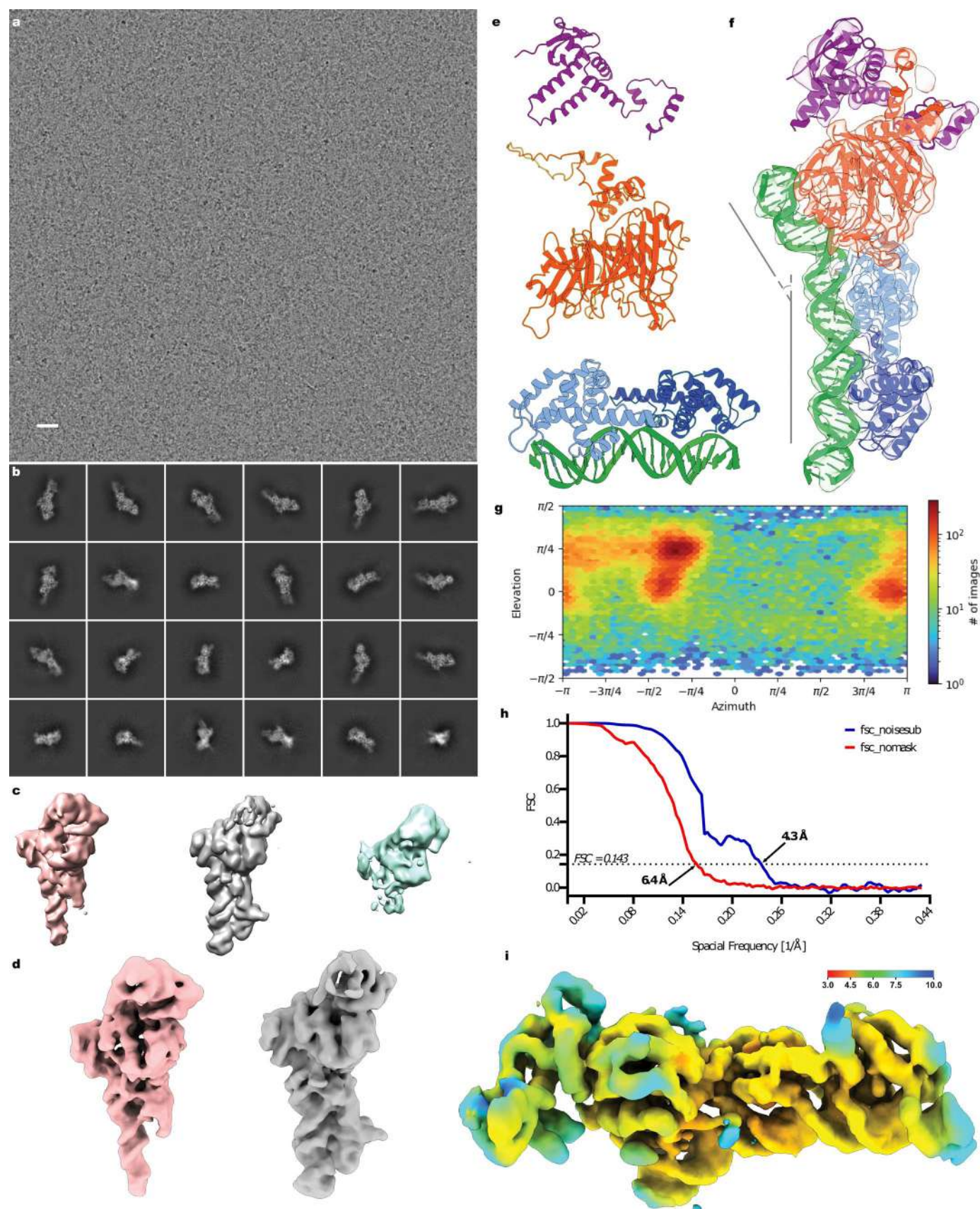
range (top). Predicted binding sites using the dLUBS, mLUBS models from Fig. 2e and LFY PWM with dependencies⁶⁸, y-axis represents score values (bottom). **c.** IGB view of selected genes showing LFY ChIP-seq in inflorescences (light blue)²⁵, LFY-UFO ampDAP-seq (yellow), LFY ampDAP-seq (pink)⁴⁸, numbers indicate read number range. Genes in red are deregulated in *ufo* inflorescences⁷⁰. ChIP-seq peaks better explained by LFY-UFO than by LFY alone are shaded in grey. Source data are available in Supplementary Data 4.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | The LFY K249 is essential for LFY-UFO-LUBS complex formation. **a**, Structure of LFY-DBD²⁹. Residues were colored by conservation using ConSurf with default parameters⁸⁵. K249 residues on each LFY monomer are represented as sticks and indicated with arrows. Note that the K249-containing loop is highly conserved. **b, c**, Promoter activation measured by DLRA in Arabidopsis protoplasts with indicated effectors (right). EV = Empty Vector (pRT104-3xHA). Tested promoters are indicated below each graph. Note that for 3xHA-LFY + UFO-3xFLAG on *pAG* only $n = 3$ biological replicates are shown. Data represent averages of independent biological replicates and are presented as mean $\pm$ SD, each dot representing one biological replicate ($n = 4$ unless specified). One-way ANOVA with Tukey's multiple comparisons tests (**b**) or Welch's ANOVA with Games-Howell post-hoc test (**c**). In (**c**), stars above bars represent a statistical difference compared to GFP. Other comparisons are indicated with brackets. (NS: $p > 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ and ****: $p < 0.0001$). **d**, Effect of the LFY_{K249R} mutation on LFY-UFO interaction in Y2H. EV = Empty Vector. LFY-40 is a LFY version lacking the first 40 aa and better tolerated by yeast cells. Values correspond to the different dilutions (OD = 7, 0.7 and 0.07). Top picture corresponds to the non-selective plate lacking Leucine and Tryptophan (SD -L-W), and bottom picture corresponds to the selective plate

lacking Leucine, Tryptophan, Histidine and Adenine (SD -L-W-A-H). Pictures were taken at day + 4. **e**, EMSA with DNA probes corresponding to *pAP3* DEE LFYBS and *pAP3* LUBS1 and indicated proteins. *pAP3* DEE LFYBS DNA probe was used as a control for binding on canonical LFYBS. **f**, WB after DNA elution during ampDAP-seq experiment. After DNA elution, 20 μ L of 1X SDS-PAGE Protein Sample Buffer was added to the remaining beads to run WB. Each lane represents one replicate. **g**, Reproducibility of ampDAP-seq experiments with LFY_{K249R} (left) and LFY_{K249R}-UFO (right) through the comparison of replicates datasets 2 by 2. **h**, Comparison of peak coverage in LFY_{K249R} (y-axis, this study) and LFY (x-axis)⁴⁸ ampDAP-seq experiments. **i**, Integrated Genome Browser (IGB) view of *pAP3* showing LFY ChIP-seq in inflorescences (light blue)²⁵ or seedlings (dark blue)²⁶, LFY-UFO ampDAP-seq (yellow; this study), LFY ampDAP-seq (pink)⁴⁸ and LFY_{K249R} ampDAP-seq (purple; this study). Numbers indicate read number range. **j**, Pictures of WT and representative transgenic plants expressing 35S::LFY or 35S::LFY_{K249R} (scale bar, 1 cm). The white arrows indicate ectopic rosette flowers. 35S::LFY was obtained previously²⁶. 42 T1 plants expressing 35S::LFY_{K249R} were analyzed; the percentage of plants with a LFY overexpressing phenotype is comparable to the one obtained with 35S::LFY²⁶. Source data are available in Supplementary Data 4.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | UFO binds DNA and LFY DBD. **a**, A representative micrograph of the ASK1-UFO-LFY-DNA complex in vitreous ice (scale bar, 20 nm). **b**, Selected 2D class averages of the particles submitted to *ab initio* reconstruction and heterogeneous refinement for 3D classification. **c**, Intermediate reconstructions of the 3D classes after heterogeneous refinement. **d**, Final reconstructions of ASK1-UFO-LFY-DNA complexes (involving either a LFY-DBD monomer (pink) or a LFY-DBD dimer (gray)) after Non-Uniform refinement. **e**, Unprocessed AlphaFold2 model for ASK1 (top, purple; uniprot ID, [Q39255](#)), UFO (middle, red; uniprot ID, [Q39090](#)) and the LFY-DBD dimer/DNA crystallographic structure (bottom, pale and dark blue for

the LFY-DBD dimer and green for the DNA; PDB, [2VY1](#)). **f**, Cryo-EM density map color-coded by fitted molecule. Note the kink on DNA induced by the presence of UFO. **g**, Heat map of the angular distribution of particle projections contributing for the final reconstruction of the complete ASK1-UFO-LFY-DNA complex (with a LFY-DBD dimer). **h**, Gold-standard Fourier shell correlation (FSC) curves. The dotted line represents the 0.143 FSC threshold, which indicates a nominal resolution of 6.4 Å for the unmasked (red) and 4.3 Å for the masked (blue) reconstruction. **i**, View of the post-processed map of the complete ASK1-UFO-LFY-DNA complex, colored according to the local resolution.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

For DLRA: TECAN Spark 10M 96-well plate reader
For EMSAs: Amersham ImageQuant 800 imager
For Western Blots: ChemiDoc MP Imaging System
For plant pictures: Keyence VHX-5000 microscope
For SEC-MALLS: Dawn Heleos II for MALLS measurement (Wyatt Technology) and Optilab T-REX refractometer for refractive index measurements (Wyatt Technology)
For cryoEM: serialEM v3.8.17

Data analysis

Data analysis procedures are described in the Methods section.
For DLRA: R studio v1.3.959
For SEC-MALLS: ASTRA v6.1.7.17
For cryoEM data analysis: cryoSPARC v3.0, crYOLO v1.7.6, DeepEMhancer, UCSF Chimera v1.15, UCSF ChimeraX v1.3
For cryoEM model building: Coot v0.9.7, UCSF Chimera v1.15, UCSF ChimeraX v1.3, PHENIX v1.20.1
Tools and versions used for bioinformatic analyses:
fastqc v0.11.7
bowtie2 v2.3.4.1
NGmerge v0.2_dev
MACS2 v2.2.7.1
bedtools v2.30.0
samtools v1.8
R 'plotROC' package v2.2.1

mSPC v4.0.0
R v3.5.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Data used in this study:

Microarray data: AtGenExpress (ATGE_52A-C and ATGE_29A-C)

LFY ampDAP-seq data: GSE160013

LFY ChIP-seq data: GSE141704, GSE96806, GSE64245, GSE24568

Structures: LFY-DBD dimer structure (RCSB PDB, 2VY1), AlphaFold models of ASK1 (uniprot ID: Q39255) and UFO (uniprot ID: Q39090).

ampDAP-seq data have been deposited and are publicly available (GSE204796).

The cryo-EM structure determined in this study is deposited in the EM databank under the reference number EMD-15145.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

For DLRA, we followed standard practice for data replication and data analysis. Sample size was sufficient for statistical significance and reproducibility. For plant studies, we analyzed several independent lines as indicated in figure legends to minimize variability.

Data exclusions

No data excluded.

Replication

The number of replication for each experiment is indicated in Source data.
GUS staining experiments were performed twice for all lines with similar results.
ampDAP-seq was performed in triplicate or in duplicate as indicated in the Method section and was reproduced 3 times with similar results.

Randomization

For plant experiments, all T1 resistant plants were further analyzed without any selection criteria.

Blinding

Investigators performing the biochemical experiments also analyzed the data so blinding was not applied. For ampDAP-seq, libraries were analyzed without prior knowledge of the tested protein(s).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	HRP-conjugated anti-myc antibody (Invitrogen; Cat# R951-25; RRID:AB_2314045) and HRP-conjugated anti-FLAG antibody (Sigma-Aldrich; Cat# A8592; RRID:AB_439702).
Validation	Provided by the supplier.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	Sf21 insect cells.
Authentication	None of the cell lines used were authenticated.
Mycoplasma contamination	N/A
Commonly misidentified lines (See ICLAC register)	N/A

4.3 Additional results and discussion

4.3.1 The role of LFY-UFO extends beyond petals and stamens development

In addition to its role in determining floral meristem identity and patterning, LFY has also been shown to control floral meristem emergence in the early stages of flower development (Moyroud et al., 2009, 2010). In *Arabidopsis*, LFY contributes to meristem emergence through the induction of *REGULATOR OF AXILLARY MERISTEMS1* (*RAX1*), a Myb family TF (Figure 4.3-1G) (Chahtane et al., 2013; Denay et al., 2018). The LFY pathway to trigger meristem emergence works in parallel with another one involving *REVOLUTA* (*REV*), a homeodomain leucine zipper (HD-ZIP) family TF that is also involved in lateral meristem formation and in the establishment of the dorsoventral axis in leaves (Figure 4.3-1G) (Denay et al., 2018; Otsuga et al., 2001; Prigge et al., 2005). While *rev* flowers often lack inner structures compared to WT (Figure 4.3-1C and A, respectively), their flower phenotype is not as affected as the one observed in *lfy-12* plants (Figure 4.3-1B) (Denay et al., 2018; Otsuga et al., 2001; Weigel et al., 1992). However, *rev-c4 lfy-12* double mutant plants display a strong spike phenotype where lateral structures are reduced to filaments (Figure 4.3-1D), suggesting that the *rev* background provides a sensitized context to study the early meristematic role of LFY in *Arabidopsis* (Denay et al., 2018).

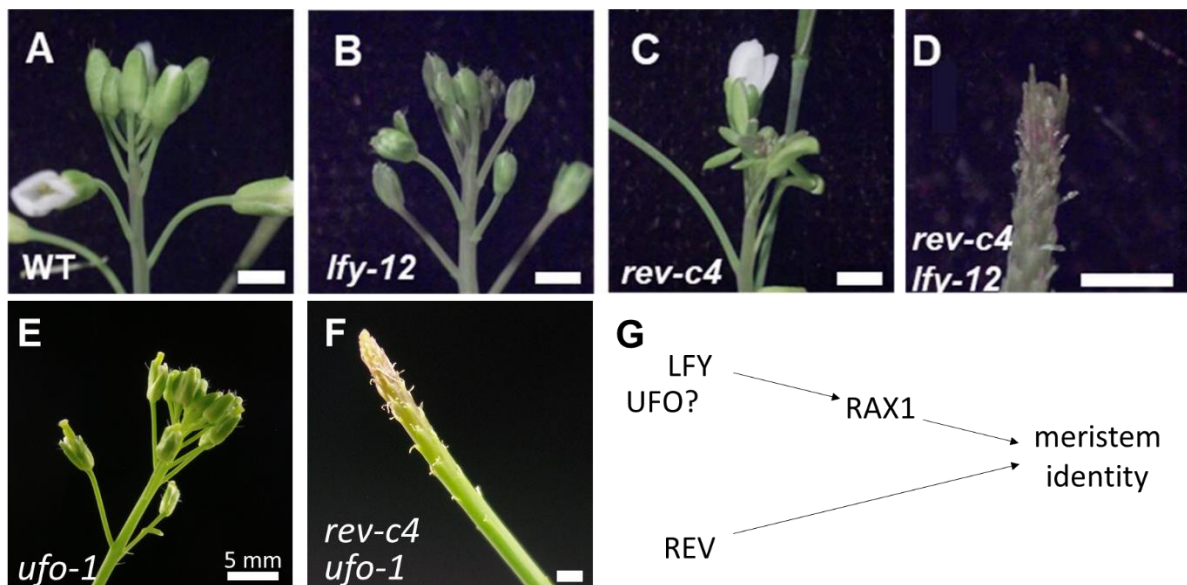


Figure 4.3-1 Investigating the role of LFY and UFO in floral meristem identity establishment. Panels A-D are from (Denay et al., 2018). Scale bars are 1 mm unless indicated differently.

Interestingly, while *ufo-1* flowers only lack petals and stamens (Figure 4.3-1E) (Durfee et al., 2003; Hepworth et al., 2006), double *rev-c4 ufo-1* mutants have a spike phenotype indistinguishable from the one observed in *rev-c4 lfy-12* plants (Figure 4.3-1F and D, respectively). This could suggest that the meristematic role of LFY also involves UFO (Figure 4.3-1G). I performed an RNA-seq experiment to test whether the gene expression profiles of *rev-c4 lfy-12* and *rev-c4 ufo-1* apices were mirroring their phenotypic similarity, and whether it could reveal new targets of the LFY-UFO complex beyond petal and stamen development. On top of that, the sensitized *rev-c4* background could reveal the presence of LFY- or UFO-specific targets.

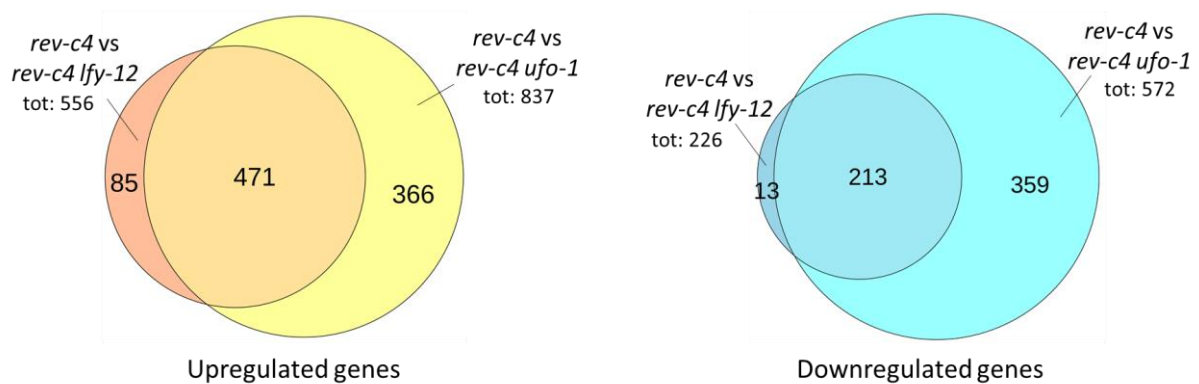


Figure 4.3-2 Differentially expressed genes in RNA-seq experiments with *rev-c4*, *rev-c4 lfy-12* and *rev-c4 ufo-1* inflorescences. LFY- and UFO-upregulated genes (left) are those with $\log_2(\text{fold change}) > 1$ compared to the *rev-c4* background and false discovery rate (FDR) < 0.01. Downregulated genes (right) have $\log_2(\text{fold change}) < -1$ and FDR < 0.01.

When comparing the expression profile of *rev-c4* and *rev-c4 lfy-12* inflorescences, more than 500 genes are significantly upregulated, and more than 80% of them (471) are also upregulated in *rev-c4* vs *rev-c4 ufo-1* inflorescences (Figure 4.3-2, left). A similar trend is observed when comparing downregulated genes in the two experiments: in this case, over 90% of genes displaying significantly lower expression levels in *rev-c4* plants compared to *rev-c4 lfy-12* plants are also downregulated when comparing *rev-c4* to *rev-c4 ufo-1* (Figure 4.3-2, right).

These results suggest that the sensitized *rev* background reveals many targets of LFY + UFO in early flower meristem development. However, different transcriptional levels can also simply indicate indirect targets, especially given that, while we have inflorescence tissue in *rev-c4* plants, *rev-c4 lfy-12* and *rev-c4 ufo-1* apices are different enough to represent a

separate tissue (see Figure 4.3-1C, D and F, p. 102). Such tissue differences affect the comparison of the transcriptional profiles of the three genotypes.

The availability of LFY and LFY-UFO ampDAP-seq data allowed me to make a list of genes that could be direct targets of the LFY-UFO complex and also be involved in an early meristematic role. Of the genes upregulated in both *rev-c4/rev-c4 lfy-12* and *rev-c4/rev-c4 ufo-1*, 87 (18%) are also bound by the LFY-UFO complex in ampDAP-seq. These genes include known targets of LFY-UFO such as *AP3*, but also new potential targets such as *CAULIFLOWER (CAL)*, a paralog of *AP1* also involved in flower meristem identity that was previously suggested as a LFY target (William et al., 2004), and *SQUAMOSA PROMOTER-LIKE4 (SPL4)*, which is involved in the transition to flowering (Jung et al., 2016). For downregulated genes, 44 out of 213 (20%) genes differentially expressed in RNA-seq are also bound by LFY-UFO, but none of them has an apparent flowering or meristem-related function.

While comparing differentially expressed genes (DEGs) to a list of genes bound in ampDAP-seq is missing the crucial information of *in vivo* binding of the complex, these results suggest that the role of LFY and UFO in floral development may also include floral meristem establishment.

4.3.2 UFO could have LFY-independent targets

In addition to LFY and UFO having common targets in floral meristem establishment, as shown by the large overlap covering almost the entirety of LFY-related DEGs, the *rev-c4/rev-c4 ufo-1* comparison reveals a large share of DEGs that are not LFY-dependent (Figure 4.3-2). This is surprising given that the phenotypes of *rev-c4 lfy-12* and *rev-c4 ufo-1* plants are very similar, and even more so considering LFY's central role in several flowering-related processes, as compared with UFO being mostly known as LFY's cofactor to regulate B-class genes. I did not investigate UFO-specific DEGs in greater detail but I suggest here some possible ways to broaden our understanding of UFO's LFY-independent roles.

The observation of UFO-specific, LFY-independent DEGs suggests a broader role of UFO in plant development, supported by the fact that UFO is also expressed in the peripheral zone of the SAM in Arabidopsis plants (Durfee et al., 2003; Laufs et al., 2003; Reddy, 2008). These

additional roles could act at the transcriptional level, like the LFY-UFO complex, with UFO collaborating with other TFs. However, finding new UFO partners requires more experimental data and computational power. For starters, UFO genome-wide binding profiling *in vivo*, whether through a ChIP-seq or techniques such as CUT&Tag or CUT&RUN, would be valuable to better characterize UFO's transcriptional role, with or without LFY. Prior to or in combination with this approach, Y2H screens and/or co-immunoprecipitation with potential TF partners could help identify new interactors, whose DNA-binding profile could be probed like LFY-UFO's. In the era of protein folding predictions with AlphaFold2 and protein-protein interaction predictions with AlphaFold Multimer, a prior computational exploration of potential UFO interactors could significantly increase the success of experimental interaction assays (Evans et al., 2021; Jumper et al., 2021). More exploratory computational approaches could leverage RNA-seq results for UFO to find common TFs with TFBSs on their regulatory regions, and possibly checking if the flanks of their TFBSs show an enrichment in an element of fixed length such as the UFO-recruiting motif found in the LFY-UFO binding site.

Although we showed that the F-box domain of UFO is dispensable for its transcriptional role with LFY, it is possible that the LFY-independent roles of UFO in Arabidopsis are related to its involvement in protein ubiquitination and degradation. In the absence of UFO-mediated degradation, its targets could go unrepressed and activate their own target genes. Potential degradation targets could be found by scanning the promoter regions of LFY-independent, UFO-dependent DEGs, to see if there are any known TFs enriched with TFBSs in those promoter regions. An additional ubiquitination-related role of UFO in development would be supported by recent findings that will be discussed in the next section.

4.3.3 Double action of UFO as a transcriptional cofactor in the nucleus and an F-box in the cytoplasm

Our ampDAP-seq experiment revealed that UFO works as a LFY cofactor for the expression of flower development genes, and that this function is independent of its protein-ubiquitination role as *AP3* expression is not lost in the absence of UFO's F-box. Moreover, transcriptional data suggests UFO's involvement in the early establishment of the flower

meristem, at least partially alongside LFY. However, this does not mean that UFO exclusively works as a cofactor, but rather that it could have a dual role in transcriptional regulation (where its F-box is dispensable) and in protein homeostasis control (F-box required).

A recent report found that LFY can form liquid droplets in the cytoplasm, and that those droplets can be targeted by UFO to trigger LFY's ubiquitination and subsequent degradation (Dolde et al., 2023). This process could be a way to control the turnover of LFY protein, and it is compatible with previous reports of LFY exiting the nucleus and working cell non-autonomously (X. Wu et al., 2003), as well as with LFY's UFO-dependent ubiquitination and degradation (Chae et al., 2008).

However, many aspects remain unclear. First, while the authors specifically focus on UFO-dependent LFY degradation, the same mechanism of cytoplasmic-specific ubiquitination and degradation by UFO could also apply to the regulation of other proteins, and possibly other TFs. More importantly, it remains unclear how the same complex that works in the nucleus to regulate developmental genes, can also lead to the ubiquitination and degradation of LFY in the cytoplasm. Further investigation will be required to dig into the mechanistic details of how LFY's turnover is regulated, what are the signals and protein modifications that lead to LFY's exit from the nucleus and UFO-mediated degradation. Finally, it should be noted that both our results and those of Dolde et al. relied on UFO overexpression, which highlights the need to see what happens when UFO is expressed under its endogenous promoter to check whether this function truly has a role *in planta*.

5 Conclusions and perspectives

5.1 Genomic context can be used to characterize transcriptional regulation of LFY sites and shed new light on LFY's regulatory properties

Gene regulation is a highly complex process involving several players, and TFs play a key role in this process as they bind to regulatory regions to control gene expression. In particular, master TFs are involved in the regulation of major developmental switches, such as the transition from vegetative to reproductive development.

The main aim of my PhD was to develop a model to predict transcriptional regulation of TFBSs genome-wide based on genetically-encoded information in their surroundings, namely genomic context and evolutionary conservation. I focused on LFY, a plant-specific master TF with a central role in flower development and on the model plant *Arabidopsis*.

My approach leveraged differences in genomic context to distinguish transcriptionally active ('functional') LFY TFBSs from inactive ('nonfunctional') ones. Such context included crucial determinants of the cis-regulatory code, from state-of-the-art LFY TFBS models (a PWM with nucleotide dependencies and POcc) to the presence of TFBSs belonging to other TFs at given distances from LFY, their density and diversity. The distance of LFY sites from each other was also included.

This approach allowed the classification of functional and nonfunctional LFY sites (Figure 3.2-1, p. 50). Moreover, it revealed that LFY has specific regulatory preferences in *Arabidopsis*, and that they are genetically encoded: the presence and quality of other LFY TFBSs in the surroundings of a LFY site is particularly important to distinguish functional and nonfunctional sites (Figure 3.2-3, p. 55).

Our model also reveals that, while to a lesser extent, the presence of non-LFY TFBSs around LFY sites and their diversity is also an important determinant of functional LFY sites (Figure 3.2-3, p. 55). Some TFs and TF groups in particular seem to differ between functional and nonfunctional sites, although their exact role and their presence on regulatory regions with LFY remain to be determined.

We used our trained model to look for functional LFY sites among those that we could not confidently label as functional nor as nonfunctional. While I do not have experimental validation of these predictions yet, some of these sites predicted as functional are likely to be transcriptionally active based on prior biological knowledge. Experimental validation will be crucial to better assess the biological relevance of our findings and the importance of co-occurring TFs.

Our approach could be applied to other TFs different from LFY, and this would give insights about whether the features important for LFY binding (e.g. presence of other LFY sites close by) are also important for other TFs, or if this is a particular characteristic of LFY's binding mode. Like other SAM-containing proteins, LFY can dimerize and form higher-order complexes that are lost upon SAM mutation (Hope et al., 2018; Qiao et al., 2020; Sayou et al., 2016). While most TFs can di- and tetramerize, examples of higher-order complexes are more rare (Amoutzias et al., 2008; Blanc-Mathieu et al., 2023; Puranik et al., 2014). Using our approach on other TFs could reveal whether features related to cooperative binding and oligomerization are important for other factors, or if it is specific to (e.g. SAM-containing) proteins that can form higher-order complexes.

Additionally, our model could be trained on Arabidopsis data and tested on other plant species, to see whether it can predict transcriptional regulation of LFY sites at increasing evolutionary distances. To this end, the pipeline I developed for LFY on Arabidopsis could be applied to virtually any other plant species to recover LFY sites and their surrounding non-LFY TFBS context. As an example, the availability of ChIP-seq and RNA-seq data for two GLK TFs in five plant species (Tu et al., 2022) could be a good TF to start testing our approach with other TFs and to probe its potential for transfer learning.

More generally, the tradeoff between model performance and interpretability is central to the use of machine learning algorithms to describe biological phenomena, and particularly in genomics, where the availability of vast data offers endless opportunities to study gene regulation (Meyer & Saez-Rodriguez, 2021; Watson, 2022). Our approach, based on the inclusion of relevant genomic features for supervised learning, offers a major advantage in model interpretability, which has led us to important biological insights into how LFY regulates the genome.

5.2 Could conservation still be useful to model transcriptional regulation?

In our quest to find biologically relevant features to include in our model, we hypothesized that the conservation of LFY TFBSs in flowering plants could be informative about their transcriptional activity, with conserved sites more likely to have a functional role. Moreover, the documented conservation of LFY's sequence and specificity in the plant lineage (Guo et al., 2019; Sayou et al., 2014) made it the ideal TF, once again, to test our hypothesis.

Nevertheless, our attempts at including conservation in our Random Forest model highlighted that the conservation of LFY sites in flowering plants, expressed as the average conservation score (PhyloP and PhastCons) over the LFY TFBS, was not informative to distinguish functional sites from nonfunctional ones (Figure 3.2-2A, p. 53). The same was true when we combined conservation information with all genomic context features (Figure 3.2-2A, p. 53).

Rather than being due to the evolutionary distance used to compute conservation scores, this result could be linked to the fact that, in our model, we studied LFY sites found both in coding and noncoding regions (Figure 3.2-2C, p. 53, compared to Figure S3.4-1D, p. 65). Our results show that, when looking at LFY sites exclusively in noncoding regions, there is a stark difference between the enrichment in CNSs in functional and nonfunctional sites, with a stronger proportion of functional sites overlapping with CNSs. This difference remains marked even at increasing evolutionary scales, including the one we included in our model, although it slightly decreased. When we looked at all LFY sites in our model, found in both coding and noncoding regions, the levels of conserved sequence enrichment were very similar in functional and nonfunctional sites. Therefore, in our model, high conservation of LFY sites regardless of their transcriptional activation status could hinder the potential of this important feature for predictions.

Previous work on the first intron of *AG*, which provides a restricted regulatory sequence across many plant species, highlighted that (i) the location of LFY sites could change in distantly related species, and that (ii) POcc proved instrumental to reconstruct the evolution of *AG* regulation by LFY without relying on alignment-based sequence conservation (Moyroud et al., 2011). While this can be a powerful strategy when looking at specific,

restricted regions, as POcc is sensitive to sequence length, it is also sensitive to changes in promoter size and TFBS location across multiple species.

More generally, given the functional importance of conserved regulatory modules rather than sequence conservation of the TFBS itself (Maher et al., 2018; Nitta et al., 2015), the insights provided by our model on LFY binding preferences and co-occurring factors could be employed to further develop comparative genomics approaches.

I only tested a portion of the endless possible strategies to include conservation in our model, and to try to understand why our method was not useful to classify LFY sites as functional or nonfunctional. Nevertheless, I think that this information will be crucial in future models aiming at studying gene regulation, and in particular the differences between TFBSs with comparable quality but different transcriptional status, in plants and beyond.

5.3 The action of LFY-UFO in plant development could extend beyond petal and stamen development

The model presented in Chapter 1 focused on the study of LFY binding when LFY acts alone to regulate gene expression. However, LFY can also work with other protein partners, including the F-box protein UFO. Our recent paper reveals that LFY and UFO form a transcriptional complex to orchestrate, in particular, petal and stamen development, and that UFO modifies LFY's canonical binding specificity to bring it to new regulatory regions (Rieu et al., 2023).

LFY and UFO also seem to share an early function in floral meristem establishment, further extending the regulatory role of the LFY-UFO complex in plant development, and the implication of LFY in the early stages of floral meristem establishment (Figure 4.3-1, p. 102). Moreover, our results show that UFO could play a larger role in these early stages, in opposition to LFY's generally wider function in flower development in Arabidopsis. This change in power balance between LFY and UFO in floral meristem establishment is in line with what has been observed in other plant species such as pea and petunia (Moyroud et al., 2009), and opens up new possibilities in their joint role as transcriptional regulators.

It would be interesting to test our approach for LFY functional sites prediction on the genomic regions specifically bound by the LFY-UFO complex (Rieu et al., 2023), and on the DEGs revealed by the *rev* background (Figure 4.3-2, p. 103), to determine whether the *cis*-regulatory context of LFY sites changes when it works alone or with UFO. In the coming years, new techniques such as CUT&Tag or CUT&RUN will be crucial to determine UFO's *in vivo* genome-wide binding profile, which could be used in approaches like ours to investigate if and how the *cis*-regulatory context of transcriptionally active LFY sites changes when LFY works alone or with other binding partners.

In addition to its role as a LFY partner for transcriptional regulation in flower development, UFO has recently been shown to regulate LFY's turnover by degrading the TF in liquid droplets in the cytoplasm (Dolde et al., 2023). While it remains unclear how UFO exerts both roles, this evidence opens new questions concerning the convergence of protein turnover and transcriptional regulation on the same TF-F-box complex in Arabidopsis.

Overall, we have probably barely scratched the surface of LFY's role in transcriptional regulation with partners. Indeed, UFO does not explain the entirety of genomic regions bound by LFY *in vivo* and not *in vitro* (Rieu et al., 2023), and there is previous evidence of LFY working with WUS, another TF, to regulate *AG* expression (Lohmann et al., 2001). In the future, comparing binding data *in vivo* and *in vitro* will make it possible to determine at a higher scale which TFs require partners to regulate their target genes, how widespread this behavior is and whether TF families differ in their cofactor preferences.

I started this manuscript talking about the complexity of gene regulation, and then I focused on TFs and their role in the control of gene expression. While I could not cover the endless exciting research avenues stemming from the study of TF-dependent gene regulation, some recent examples that I find particularly exciting include the role of intrinsically disordered regions in TF-target recognition (Brodsky et al., 2020; Jonas et al., 2023; Staller, 2022), and the relevance of TF-RNA interactions (Oksuz et al., 2023) as well as 3D genome organization and dynamics (Kim & Shendure, 2019; Wagh et al., 2021) in gene regulation. These findings show that the TF paradigm that I focused on, i.e. that TFs are proteins that recognize specific motifs on gene regulatory regions through their DBD, is a non-exhaustive definition that is still being questioned (Samee, 2023), and there is still much to be done in the coming years to increase our understanding of how TFs regulate gene expression in all organisms.

6 Materials and Methods

6.1 ChIP-seq and ampDAP-seq analysis

LFY ChIP-seq data from (Goslin et al., 2017; Jin et al., 2021; Moyroud et al., 2011; Sayou et al., 2016) were taken from GEO (GSE96806, GSE141706, GSE24568, GSE64245, respectively). LFY ampDAP-seq data was from GEO GSE160013 (Lai, Blanc-Mathieu, et al., 2021).

Fastq files were processed as in (Rieu et al., 2023). Sequencing data quality was evaluated with fastQC v.0.11.7, and adapters were removed with NGmerge v.0.2_dev (Gaspar, 2018). Bowtie2 v.2.3.4.1 was used to map reads to the TAIR10 *A. thaliana* reference genome (Lamesch et al., 2012). We only retained reads mapped to a single location and with a maximum of two mismatches. We used samtools dedup v.1.8 to remove duplicates. We identified bound regions (“peaks”) with MACS2 v.2.2.7.1 (Y. Zhang et al., 2008), with input DNA from Lai et al. as a control (Lai, Blanc-Mathieu, et al., 2021) and with $-q$ 0.05 for ChIP-seq and $-q$ 0.0001 for ampDAP-seq. Consensus peaks called in all replicates were identified with MSPC v.4.0.0 (Jalili et al., 2015) and, finally, peaks were resized around the peak maximum (± 200 bp) for further analysis.

6.2 Microarray analysis

Microarray data were retrieved from GEO (GSE28062) for 35S::LFY-GR seedlings after the addition of dexamethasone (Winter et al., 2011), from (Chahtane et al., 2013) for 35S::LFY experiments and from AtGen-Express (Schmid et al., 2004) for inflorescence tissue in the *lfy* background. Each mutant or overexpressing genotype was compared to WT (Col-0). The R package gcrma (Z. Wu & Irizarry, 2022) was used to adjust probe intensities and convert them to expression measures, and then the limma package (Ritchie et al., 2015) was used for differential expression analysis. A Benjamini–Hochberg correction was applied to the P values, and fold change (FC) was computed as the ratio between expression in the overexpression line/WT (for 35S::LFY-GR and 35S::LFY) or WT/*lfy* mutant. Only genes with $|\log_2(\text{FC})| > 1$ and adjusted $P < 0.05$ were considered as significantly differentially expressed.

6.3 RNA-seq experiments and analyses

6.3.1 WT vs *lfy* RNA-seq experiment

WT (Col-0) and *lfy-12* seeds were grown for 5 weeks in short day (SD) conditions (8h light, 16h dark), then moved to long days (LD) conditions (16h light, 8h dark) for another two weeks before inflorescence dissection. All flower-looking structures were removed, and only the inner part of the inflorescence was sampled (2-3 mm). We only sampled primary inflorescences for a total of four samples per genotype, each sample containing multiple inflorescences. RNA was extracted with the Qiagen Rneasy Kit, and sent for paired-end mRNA sequencing. After sequencing, fastq files were trimmed with BBduk to remove adapter sequences. Trimmed reads were fed to Salmon (Patro et al., 2017) using the `--gcBias` and `--validateMappings` flags for mapping on the Arabidopsis TAIR10 transcriptome (Lamesch et al., 2012) and read counting. The tximport R package (Soneson et al., 2015) was used to import read counts computed by Salmon of all samples to RStudio v1.3.959 (RStudio Team, 2020), and the DESeq2 package v1.28.1 (Love et al., 2014) was used to normalize raw counts by transcript length. Transcript length-normalized counts were analyzed with the NOISeq package (Tarazona et al., 2015). The ARSyNseq function was applied for noise removal, followed by the noiseqbio function for differential analysis with the options `lc = 0`, `norm = "tmm"`, `cpm = 1`, `filter = 1` and all other options with default values. Genes were considered as differentially expressed if their $|\log_2(FC)|$ was > 1 and the associated adjusted p-value (corresponding to 1 - Probability calculated by the noiseqbio function) was less than 0.01.

6.3.2 *lfy rev* vs *rev*, *ufo rev* vs *rev* RNA-seq experiment

rev-c4, *lfy-12 rev-c4* and *ufo-1 rev-c4* plants were grown in long days conditions (16h light, 8h dark) for 6 to 7 weeks prior to the collection of inflorescences. All flower-looking structures were removed, and only the inner part of the inflorescence was sampled (2-3 mm). We only sampled primary inflorescences for a total of four samples per genotype, each sample containing multiple inflorescences. RNA was extracted with the Qiagen Rneasy Kit and sent for QuantSeq 3' mRNA-Seq sequencing with Unique Molecular Identifiers (UMIs). After sequencing, UMIs were extracted with `umi_tools extract` command by UMI-tools v1.1.2 (Smith et al., 2017) and reads were trimmed with Bbduk v 38.18 to remove adapter

sequences. STAR v2.7.9a (Dobin et al., 2013) was used to align reads on the Arabidopsis TAIR10 genome and then the `umi_tools dedup` command by UMI-tools v1.1.2 was used to remove PCR duplicates. Reads were counted with HTSeq v0.13.5 (Putri et al., 2022) and differential analysis was performed with DESeq2 v1.28.1 (Love et al., 2014) on RStudio v1.3.959 (RStudio Team, 2020). Genes were considered as differentially expressed if their $|\log_2(\text{FC})|$ was > 1 and the associated adjusted p-value was less than 0.01.

6.4 Data matrix to classify LFY sites based on genomic context and conservation

6.4.1 Definition of LFY sites genome-wide

The LFY PWM with dependencies (Moyroud et al., 2011) was used to scan the Arabidopsis genome (TAIR10) (Lamesch et al., 2012) to predict LFY TFBS, and a PWM score was assigned to every genomic position. All genomic scores were used to determine an overall distribution (Figure S3.4-1A, p. 65), which was used to select LFY TFBSs based on percentile thresholds. I tested multiple score-thresholds ranging from the 99th to the 99.99th percentile (Figure S3.4-1C, p. 65), but we ended up choosing the 99.9th percentile threshold for further analyses, i.e. we only retained the top 0.1% best-scoring sites. LFY site names (unique identifiers) were defined with the following nomenclature: chromosome number, start and end position of the site's genomic coordinates (e.g. "chr1:10000017-10000036").

6.4.2 Integration of binding and expression data and definition of nonfunctional/functional/'unknown' LFY sites

To build a model that is able to predict for which LFY TFBS binding elicits a transcriptional response (i.e. 'functional' LFY TFBS) *in vivo*, we integrated LFY binding and expression data. We created a table where each row contained a LFY site on the Arabidopsis genome over a given PWM score threshold, and the three columns had information about experimental evidence of binding (*in vivo* with ChIP-seq data, *in vitro* thanks to ampDAP-seq data) or differential expression (microarrays and RNA-seq experiment). A site was considered to be associated with a significant change in gene expression if it was found in the genomic interval from 3 kb upstream of the transcription start site (TSS) of an Arabidopsis gene to 1 kb downstream of its transcription termination site (TTS). All columns are binary, with '1' for

overlap between a given LFY site and the corresponding source of data and '0' for no overlap. LFY sites overlapping with genomic regions bound by LFY in ChIP-seq and ampDAP-seq experiments, and associated with differentially expressed genes, were labeled as 'functional' ('1'). The rationale behind this choice is that we wanted to select LFY TFBS bound *in vivo* and by LFY alone, not in a complex. LFY sites with no evidence of binding or of differential expression close by were labeled as 'nonfunctional' ('0'). Functional and nonfunctional sites were then used to train random forest models. All sites satisfying one of the previous criteria but not all of them at once, e.g. found in a region bound in ChIP-seq but not in ampDAP-seq, or bound in ChIP-seq and DAP-seq but close to a gene that is not differentially expressed, were labeled as 'unknown' ('U'), as they could not be confidently labeled as either functional or nonfunctional.

6.4.3 Computing POcc around LFY TFBS

POcc was calculated using the method published in (Moyroud et al., 2011) and explained in Equations 1 and 2 (see Computing POcc around LFY TFBS, p. 116) with LFY's PWM with dependencies. We used DNA sequences spanning ± 250 bp or ± 500 bp around each LFY TFBS (Figure 6.4-1). To define the presence of LFY TFBS around each reference site, we used the same score threshold as to identify LFY sites genome-wide, i.e. the top 0.1% best scoring sites (99.9th percentile). We chose the ± 250 bp genomic window for further analyses based on a greater PR curve AUC compared to the one obtained for the ± 500 bp interval (Figure 6.4-1).

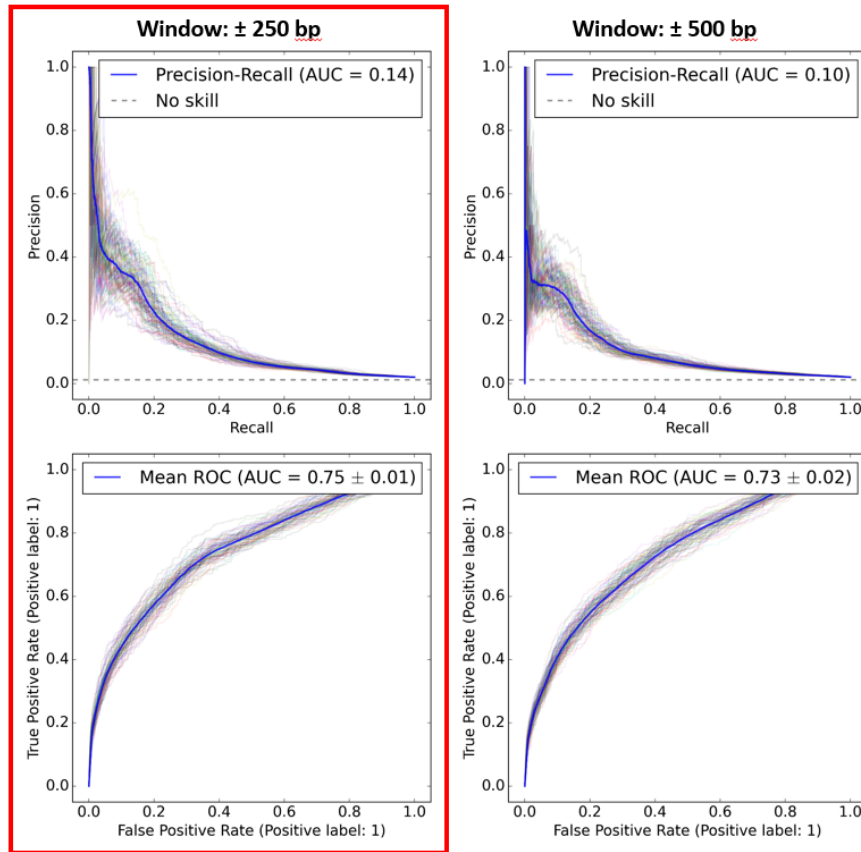


Figure 6.4-1 PR (top row) and ROC (bottom row) curves of random forest models trained with only one feature, POcc, computed over a window of ± 250 bp (graphs on the left) or ± 500 bp (graphs on the right) around each LFY site. Each thin line represents a separate random forest model trained on 75% of LFY sites and tested on the remaining 25%, for a total of 100 models (see Training and testing Random Forest models, p. 124); the thick blue line shows the mean of all models. Mean AUC is shown for both PR and ROC curves, at both genomic windows around LFY sites. The red square indicates the genomic window that is used in Figure 3.2-1, p. 50.

6.4.4 Computing co-occurrence and LFY-LFY distances

To compute co-occurrence distances, I downloaded PWMs of 47 TF clusters representing over 500 plant TFs from JASPAR 2022 (Castro-Mondragon et al., 2022) and I used them to scan the Arabidopsis genome. As for LFY sites (see Definition of LFY sites genome-wide, p. 115), I computed each cluster's genome-wide score distribution and set a PWM score threshold at the 99.9th percentile. Then, I used *bedtools closest* with options `-t all -m db each -d` (Quinlan & Hall, 2010) to find the distance, in bp, of each LFY site from the closest site of each TF cluster. I then transformed such distance to the distance between the central position of each matrix (LFY-cluster) based on matrix length.

One of the clusters, 'cluster_47', represented a single TF, LFY, and was excluded from further analyses. Instead, to compute LFY-LFY TFBS distances, I used the LFY sites detected with the PWM with dependencies (Moyroud et al., 2011) with score above the 99.9th percentile threshold, as explained in Definition of LFY TFBS genome-wide. This time, I used *bedtools closest* with options -t all -mdb each -d -io (Quinlan & Hall, 2010) to avoid overlaps.

6.4.5 Computing LFY-TSS distances

The same way I computed distances between LFY and TF clusters or LFY and other LFY sites, I included the distances, in bp, of each LFY site from the closest TSS. This time, I used *bedtools closest* with option -D a (Quinlan & Hall, 2010) to keep positive or negative distance information based on whether the LFY site was upstream or downstream of the closest TSS, respectively. TSS positions were determined based on TAIR10 annotation (Lamesch et al., 2012), taking one isoform per Arabidopsis gene.

6.4.6 Encoding sequence type

Six binary features were dedicated to the type of sequence where each LFY TFBS was found: CDS, 5' UTR, 3' UTR, intron, promoter and downstream regulatory region. Promoters were defined as 3 kb upstream of the TSS, and downstream regulatory regions as 1 kb downstream of the TTS. Both promoters and downstream regulatory sequences can overlap neighboring genes, thus some genomic regions can be categorized as multiple types at once. For each feature column in the data matrix, '1' indicated that the corresponding TFBS was found in that type of sequence. For LFY TFBS overlapping multiple sequence types, multiple feature columns contained a '1'. The TAIR10 gff annotation file was used as a reference (Lamesch et al., 2012).

6.4.7 Computing TFBS density and diversity around LFY TFBSs

TFBS density was computed as the total amount of non-LFY TFBSs ('totTFBS' in Figure 3.2-1A, see example in Figure 6.4-2).

To compute TFBS diversity around LFY sites, I used two methods:

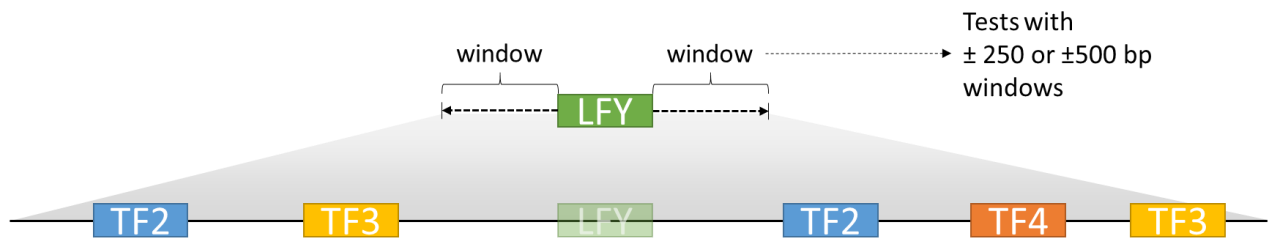
1. The first method ('sindex' in Figure 3.2-1A, p. 50, and Figure 6.4-2, p. 120, example in Figure 6.4-2, p. 120) was an adaptation of the Shannon entropy formula to quantify non-LFY TFBS diversity as in Equation 3:

$$H' = - \sum_{i=1}^S p_i \ln p_i$$

Equation 3 Shannon's entropy formula adapted to quantify non-LFY TFBS diversity around LFY sites. p_i is the proportion of non-LFY TFBS of cluster _{i} over the total amount of non-LFY TFBSs in the genomic region considered (± 250 bp or ± 500 bp around each LFY site), and S is the total amount of clusters considered (46 here).

2. For the second method ('diff_tfs' in Figure 3.2-1A, p. 50, and Figure 6.4-3, p. 121), I counted the number of *different* non-LFY TFs with a TFBS around LFY, as shown in the example in Figure 6.4-2, p. 120.

Non-LFY TFBS density and diversity were computed within a window of ± 250 bp or ± 500 bp around each LFY site (Figure 6.4-2, p. 120). I tried different combinations of these three features, either one at a time or two at a time, to train and test 40 Random Forest models (Figure 6.4-2, p. 120). As the AUC of PR curves was always close to 0, I chose a combination of Shannon's entropy (for TFBS diversity) and total non-LFY TFBSs (for TFBS density) features at a ± 500 bp window for further analyses, based on a greater median ROC AUC compared to the other combinations.



Value of non-LFY **TFBS density** around LFY sites: **5**

Value of non-LFY **TFBS diversity** around LFY sites – **Shannon's entropy**: **1.0549**

$$\begin{array}{lcl}
 p_i \text{ TF2} = 2/5 \rightarrow p_i \ln p_i = -0.3665 & & \\
 p_i \text{ TF3} = 2/5 \rightarrow p_i \ln p_i = -0.3665 & & \\
 p_i \text{ TF4} = 1/5 \rightarrow p_i \ln p_i = -0.3219 & \left. \vphantom{\begin{array}{l} p_i \text{ TF2} \\ p_i \text{ TF3} \\ p_i \text{ TF4} \end{array}} \right\} & 1.0549
 \end{array}$$

Value of non-LFY **TFBS diversity** around LFY sites – **alternative method**: **3**

Figure 6.4-2 Schematic example of methods to compute non-LFY TFBS density and diversity around LFY sites. From top to bottom: scheme showing a LFY TFBS with extensions on both sides, representing the two genomic windows we tested: ± 250 bp or ± 500 bp. Below, scheme showing non-LFY TFBSs belonging to TF2, TF3 and TF4 and found within the selected genomic region. Based on the methods explained above, values corresponding to TFBS density and TFBS diversity, both through the adaptation of Shannon's entropy and as the number of different TFs, are circled in red.

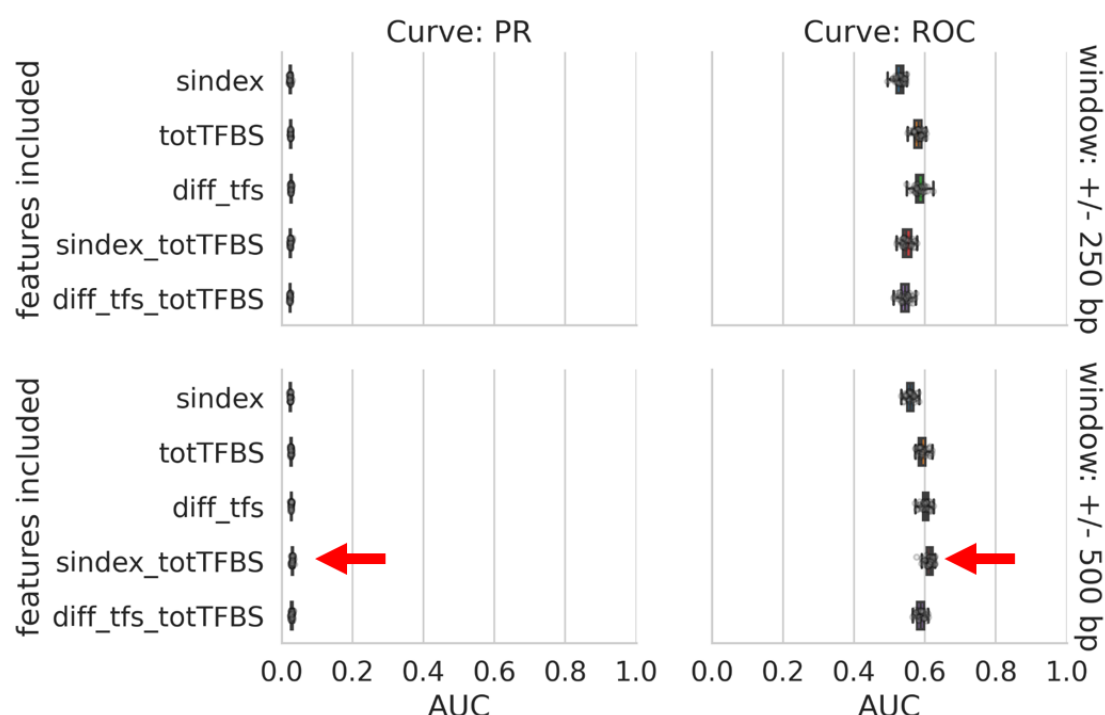


Figure 6.4-3 AUC for PR (graphs on the left) and ROC (graphs on the right) curves of Random Forest models trained with different features representing TFBS density and diversity around LFY sites. Density and diversity values were computed at two genomic windows around LFY sites: ± 250 bp (top row) and ± 500 bp (bottom row). ‘sindex’: models trained with one feature representing TFBS diversity as computed with Equation 3, in the previous page; ‘totTFBS’: models trained with one feature representing the total amount of non-LFY TFBSs around LFY sites; ‘diff_TFs’: models trained with one feature representing the total number of different non-LFY TFBSs around LFY sites; ‘sindex_totTFBS’: models trained with two features, ‘sindex’ and ‘totTFBS’; ‘diff_tfs_totTFBS’: models trained with two features, ‘diff_tfs’ and ‘totTFBS’. Each dot represents the AUC of the (PR or ROC) curve obtained with a separate random forest model trained on 75% of LFY sites and tested on the remaining 25%, for a total of 40 models (this is the only case with 10 repeats instead of 25; see Training and testing Random Forest models, p. 124). Boxplots display lower and upper quartile values (box edges), median (line between the two box edges), and whiskers extend from the box at 1.5 the interquartile range. The red arrow indicates the combination of TFBS diversity and density features used in Figure 3.2-1A and B, p. 50.

6.4.8 Computing average conservation at LFY sites

PhastCons and PhyloP conservation scores on the Arabidopsis genome were downloaded from the PlantRegMap database (Tian et al., 2020). Conservation at each position of each LFY TFBS was retrieved with bwtool extract (Pohl & Beato, 2014) and several strategies were tested to encode conservation:

1. Average conservation score at each site, for both types of scores (PhyloP and PhastCons). See ‘Average’ in Figure 6.4-4, p. 123.

2. Weighted average of the conservation score, where the IC of each nucleotide in a LFY palindromic PWM (obtained from ampDAP-seq peaks) was used as weight, depending on its position within LFY's binding site. This meant that e.g. if the selected LFY site had an A in position 1, the IC of the A nt in the LFY's palindromic PWM was used as weight, while if the selected LFY site had a T in position 1, the IC of the T nt in the PWM was used as weight. See 'Weighted average*' in Figure 6.4-4, p. 123. The same strategy was used for both types of scores (PhyloP and PhastCons).
3. Weighted average of each conservation score, using the total IC per position as weight. This meant that a fixed weight was used per position, independently of the nature of the nt at each particular position of the selected LFY site. See 'Weighted average**' in Figure 6.4-4, p. 123. The same strategy was used for both types of scores (PhyloP and PhastCons).

I used conservation features computed with each strategy to train and test 100 Random Forest models (see Training and testing Random Forest models, p. 124, for more details about my cross-validation strategy). As precision-recall curves for all models had an AUC of 0.01, I decided to keep conservation features as computed with the average method described above based on a slightly higher average ROC AUC (Figure 6.4-4, p. 123, red box).

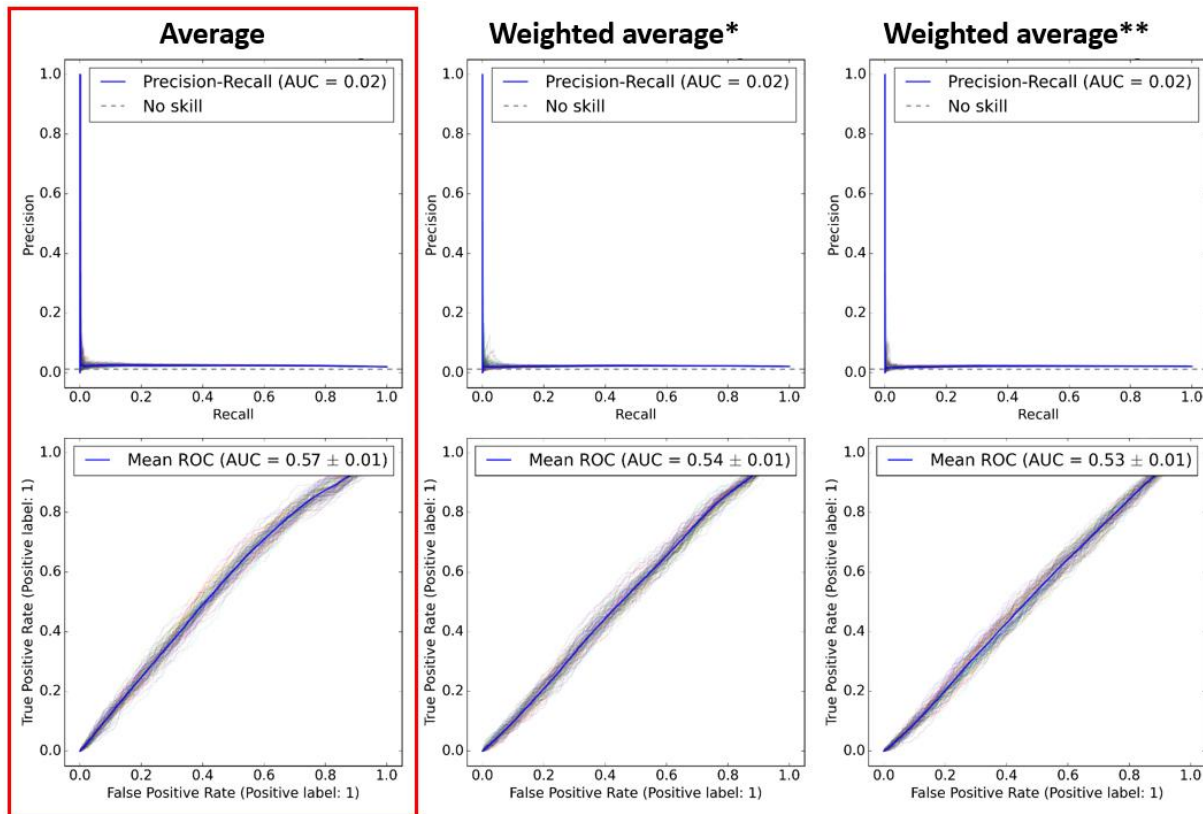


Figure 6.4-4 PR (top row) and ROC (bottom row) curves of random forest models trained with two conservation features (PhastCons and PhyloP scores) computed with three different methods: Average (graphs on the left), Weighted average* (graphs in the middle) and Weighted average** (graphs on the right). Average: mean conservation score over the LFY site; Weighted average*: weighted average of each conservation score, where each nucleotide's IC is used as weight, depending on its position within LFY's binding site; Weighted average**: weighted average of each conservation score, where the total IC per position is used as weight. Each thin line represents a separate random forest model trained on 75% of LFY sites and tested on the remaining 25%, for a total of 100 models (see Training and testing Random Forest models, p. 124); the thick blue line shows the mean between all the separate models. The mean AUC is shown for both PR and ROC curves, at both genomic windows around LFY sites. The red square indicates the conservation computation method used in Figure 3.2-2, p. 53.

6.5 Calculating CNS enrichment at LFY sites

To determine whether LFY functional, nonfunctional and 'unknown' sites were overlapping with CNSs, I retrieved CNS datasets from four publications and I kept their genomic coordinates (Haudry et al., 2013; Tian et al., 2020; Velde et al., 2014, 2016). For each dataset, I used bedtools intersect with option `-wao` to check whether any CNSs were overlapping with LFY sites coordinates by at least 1 nt. Then, for each LFY site class, I calculated and plotted the proportion of sites overlapping with CNSs comparing to those with no overlap. For the analysis in Figure 3.2-2C (p. 53), I removed LFY sites found in coding

regions (as previously determined: Encoding sequence type) before calculating the proportion of sites overlapping with CNSs. For Figure S3.4-1 (p. 65) I used all LFY sites to calculate proportions of overlapping CNSs.

For Figure 3.2-2D (p. 53), for each CNS dataset, I calculated the ratio between the proportion of functional sites overlapping with CNSs in Figure 3.2-2C (p. 53) and the proportion of nonfunctional sites of nonfunctional overlapping with CNSs shown in the same figure.

6.6 Training and testing Random Forest models

I used the scikit-learn python package (v1.2.0) to run Random Forest algorithms on our data with the RandomForestClassifier function and max_depth option set to 20 (Pedregosa et al., 2011). Then, I used cross-validation to evaluate whether the model was capable of distinguishing functional LFY sites from nonfunctional ones.

In cross-validation, a dataset is divided into a given amount of equal subsets (“folds”), and all except one are used to train a model. The remaining fold is used to test the trained model, i.e. to see whether it is capable of predicting the label of an entry in the dataset, which it has never seen, based on the entry’s features. There are many possible strategies for cross-validation, and the best one depends on one’s data. In our case, as the data are highly unbalanced, two popular options are to use Repeated Stratified k-fold cross-validation (Figure 6.6-2) or to undersample the majority class (i.e. nonfunctional sites) to match the amount of the minority class (functional sites) in each fold, and thus obtain a balanced dataset.

To determine the best cross-validation strategy for our data, I set up a procedure to compare the performance of Stratified vs Balanced cross-validation on the dataset containing all features except conservation, for a total of 58 features. My procedure worked as follows: I trained a model with the balanced strategy and another with the stratified one, and I tested each model on the leftover functional and nonfunctional LFY sites, i.e. those not used to train either model. For both strategies I used 4 folds. At each round of my procedure I trained a Balanced and a Stratified model, I tested them on the same unused data and I fed true labels and model predictions to scikitlearn’s function ‘precision_recall_curve’ to plot PR curves and retrieve AUC values for each. Then I repeated by changing the folds used for

training. As the resulting PR AUC was greater for models cross-validated with the stratified strategy, I used it all models presented in this manuscript (Figure 6.6-1).

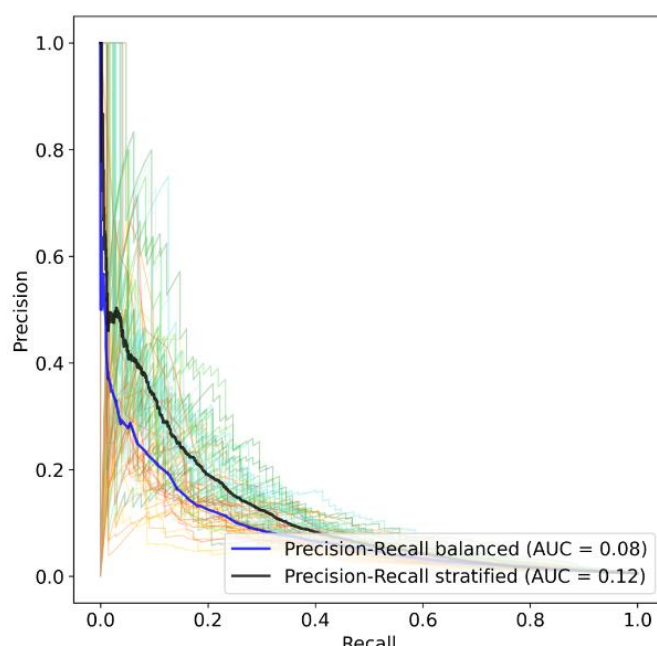


Figure 6.6-1 PR curve to compare balanced vs stratified cross-validation strategies. Green and orange curves obtained when testing stratified and balanced Random Forest models, respectively.

All figures in Chapter 1: A machine-learning model to predict transcriptional regulation of LFY sites genome-wide based on genomic context and evolutionary conservation were generated with a Repeated Stratified k-fold cross-validation strategy with $k=4$ and 25 repeats (Figure 6.6-2, in the next page). As each round of cross-validation (one train-test iteration) produces a model of its own, I generated a total of 100 models with the same parameters but with different training and testing sets (Figure 6.6-2). The only case in which I used less than 25 repeats was to determine which combination of TFBS diversity and density features to include in the model, where I ran a Repeated Stratified k-fold cross-validation strategy with $k=4$ and 10 repeats for a total of 40 models (see Computing TFBS density and diversity around LFY TFBSs, p. 118).

At each cross-validation test step, I fed true labels and model predictions to scikitlearn's functions `RocCurveDisplay.from_predictions` and `precision_recall_curve` to plot ROC and PR curves, respectively, and retrieve AUC values for each.

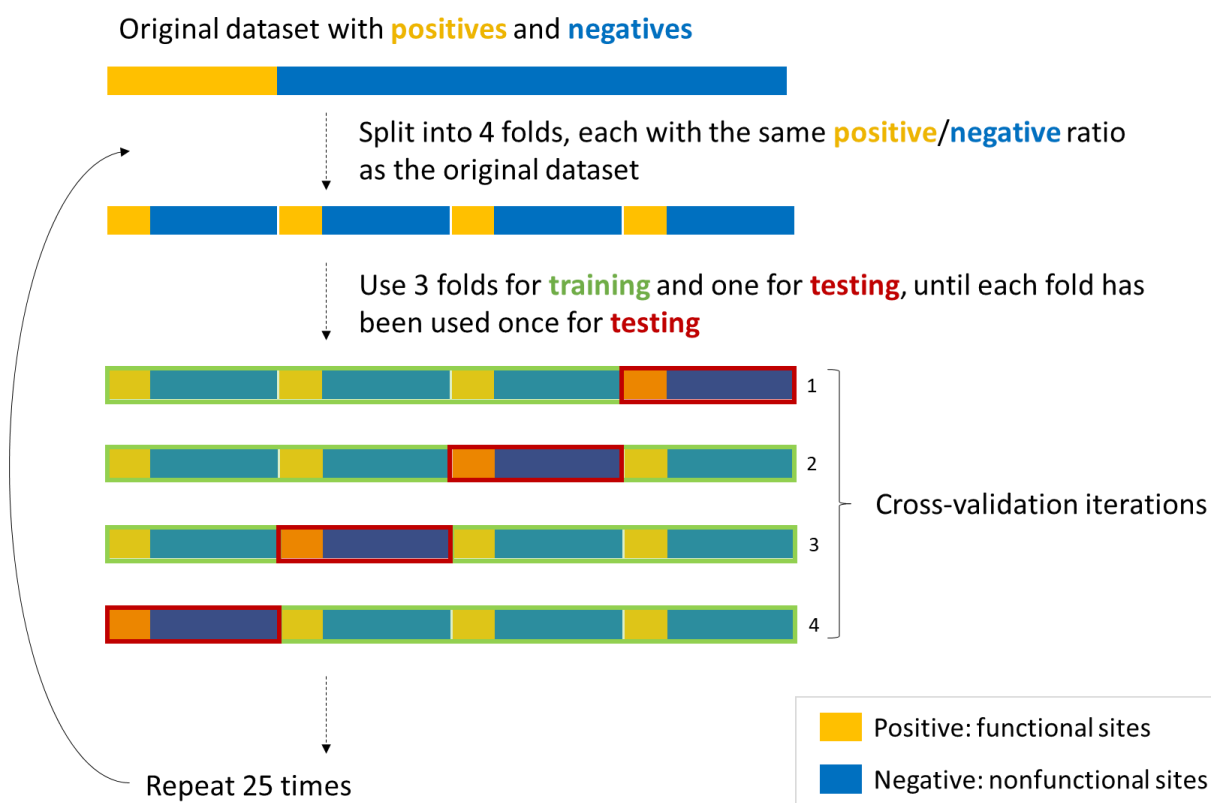


Figure 6.6-2 Schematic representation of the Repeated Stratified k -fold cross-validation strategy used in all Random Forest models shown in this manuscript. The original dataset, which contains a certain amount of positive and negative entries (in our case, an unbalanced dataset with few positives, i.e. functional LFY sites, and many negatives, i.e. nonfunctional LFY sites), is split into 4 folds, each one with the same positive to negative ratio as the original one. Then, three of these four folds are used to train a model (with labels), and the fourth one is used to test the trained model (hidden labels); the same thing is repeated three more times until each fold has been used once for testing. When this process is completed, the original dataset is once again split in four folds, different from the previous ones, and the training-testing process is repeated.

6.7 Extracting feature importance from Random Forest models

Following the cross-validation strategy explained above, after each training round I used the function `feature_importances_` from `scikitlearn` to retrieve Gini importance for each feature included in the model.

6.8 Using trained Random Forest models to make predictions on ‘unknown’ sites

Following the cross-validation strategy explained in Training and testing Random Forest models, after each training round I tested the trained model on LFY sites labeled as ‘unknown’ and I retrieved prediction probabilities for the ‘1’ (LFY functional site) class.

7 References

- Akagi, T., Masuda, K., Kuwada, E., Takeshita, K., Kawakatsu, T., Ariizumi, T., Kubo, Y., Ushijima, K., & Uchida, S. (2022). Genome-wide cis-decoding for expression design in tomato using cistrome data and explainable deep learning. *The Plant Cell*, 34(6), 2174-2187. <https://doi.org/10.1093/plcell/koac079>
- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), 831-838. <https://doi.org/10.1038/nbt.3300>
- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., ... Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2), 481-491. <https://doi.org/10.1016/j.cell.2016.05.063>
- Ambrosini, G., Groux, R., & Bucher, P. (2018). PWMScan : A fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, 34(14), 2483-2484. <https://doi.org/10.1093/bioinformatics/bty127>
- Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist : Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1), 1. <https://doi.org/10.1038/s41598-019-45839-z>
- Amoutzias, G. D., Robertson, D. L., Van de Peer, Y., & Oliver, S. G. (2008). Choose your partners : Dimerization in eukaryotic transcription factors. *Trends in Biochemical Sciences*, 33(5), 220-229. <https://doi.org/10.1016/j.tibs.2008.02.002>
- Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10), 1196-1203. <https://doi.org/10.1038/s41592-021-01252-x>
- Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., & Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 1-13. <https://doi.org/10.1038/s41588-021-00782-6>
- Back, G., & Walther, D. (2021). Identification of cis-regulatory motifs in first introns and the prediction of intron-mediated enhancement of gene expression in *Arabidopsis thaliana*. *BMC Genomics*, 22(1), 390. <https://doi.org/10.1186/s12864-021-07711-1>
- Ballester, B., Medina-Rivera, A., Schmidt, D., González-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A. J., Funnell, A. P., Goncalves, A., Kutter, C., Lukk, M., Menon, S., McLaren, W. M., Stefflova, K., Watt, S., Weirauch, M. T., Crossley, M., Marioni, J. C., ... Wilson, M. D. (2014). Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife*, 3, e02626. <https://doi.org/10.7554/eLife.02626>
- Bargmann, B. O. R., Marshall-Colon, A., Efroni, I., Ruffel, S., Birnbaum, K. D., Coruzzi, G. M., & Krouk, G. (2013). TARGET : A Transient Transformation System for Genome-Wide Transcription Factor Target Discovery. *Molecular Plant*, 6(3), 978-980. <https://doi.org/10.1093/mp/sst010>
- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved Elements in the Human Genome. *Science*, 304(5675), 1321-1325. <https://doi.org/10.1126/science.1098119>
- Bennett, H. M., Stephenson, W., Rose, C. M., & Darmanis, S. (2023). Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods*, 20(3), 3. <https://doi.org/10.1038/s41592-023-01791-5>
- Benos, P. V., Bulyk, M. L., & Stormo, G. D. (2002). Additivity in protein–DNA interactions : How good an approximation is it? *Nucleic Acids Research*, 30(20), 4442-4451. <https://doi.org/10.1093/nar/gkf578>

- Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The arabidopsis information resource : Making and mining the “gold standard” annotated reference plant genome. *Genesis*, 53(8), 474-485. <https://doi.org/10.1002/dvg.22877>
- Bernard, V., Lecharny, A., & Brunaud, V. (2010). Improved detection of motifs with preferential location in promoters. *Genome*, 53(9), 739-752. <https://doi.org/10.1139/G10-042>
- Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T., & Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*, 2(1), 1. <https://doi.org/10.1038/s41559-017-0377-2>
- Blanc-Mathieu, R., Dumas, R., Turchi, L., Lucas, J., & Parcy, F. (2023). Plant-TFClass : A structural classification for plant transcription factors. *Trends in Plant Science*, 0(0). <https://doi.org/10.1016/j.tplants.2023.06.023>
- Blázquez, M. A., Soowal, L. N., Lee, I., & Weigel, D. (1997). LEAFY expression and flower initiation in Arabidopsis. *Development*, 124(19), 3835-3844. <https://doi.org/10.1242/dev.124.19.3835>
- Boeva, V. (2016). Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Frontiers in Genetics*, 7. <https://www.frontiersin.org/articles/10.3389/fgene.2016.00024>
- Bradley, D., Ratcliffe, O., Vincent, C., Carpenter, R., & Coen, E. (1997). Inflorescence Commitment and Architecture in Arabidopsis. *Science*, 275(5296), 80-83. <https://doi.org/10.1126/science.275.5296.80>
- Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D. K., Carmi, M., & Barkai, N. (2020). Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Molecular Cell*, 79(3), 459-471.e4. <https://doi.org/10.1016/j.molcel.2020.05.032>
- Brooks, M. D., Reed, K. M., Krouk, G., Coruzzi, G. M., & Bargmann, B. O. R. (2023). The TARGET System : Rapid Identification of Direct Targets of Transcription Factors by Gene Regulation in Plant Cells. In Q. Song & Z. Tao (Éds.), *Transcription Factor Regulatory Networks* (p. 1-12). Springer US. https://doi.org/10.1007/978-1-0716-2815-7_1
- Brownlie, P., Ceska, T., Lamers, M., Romier, C., Stier, G., Teo, H., & Suck, D. (1997). The crystal structure of an intact human Max–DNA complex : New insights into mechanisms of transcriptional control. *Structure*, 5(4), 509-520. [https://doi.org/10.1016/S0969-2126\(97\)00207-4](https://doi.org/10.1016/S0969-2126(97)00207-4)
- Burgess, D., & Freeling, M. (2014). The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates. *The Plant Cell*, 26(3), 946-961. <https://doi.org/10.1105/tpc.113.121905>
- Busch, C. A., Supriya, K., Cooper, K. M., & Brownell, S. E. (2022). Unveiling Concealable Stigmatized Identities in Class : The Impact of an Instructor Revealing Her LGBTQ+ Identity to Students in a Large-Enrollment Biology Course. *CBE—Life Sciences Education*, 21(2), ar37. <https://doi.org/10.1187/cbe.21-06-0162>
- Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., ... Mathelier, A. (2022). JASPAR 2022 : The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1), D165-D173. <https://doi.org/10.1093/nar/gkab1113>
- Chae, E., Tan, Q. K.-G., Hill, T. A., & Irish, V. F. (2008). An Arabidopsis F-box protein acts as a transcriptional co-factor to regulate floral development. *Development*, 135(7), 1235-1245. <https://doi.org/10.1242/dev.015842>
- Chahtane, H., Vachon, G., Masson, M. L., Thévenon, E., Pérignon, S., Mihajlovic, N., Kalinina, A., Michard, R., Moyroud, E., Monniaux, M., Sayou, C., Grbic, V., Parcy, F., & Tichtinsky, G. (2013). A variant of LEAFY reveals its capacity to stimulate meristem development by inducing RAX1. *The Plant Journal*, 74(4), 678-689. <https://doi.org/10.1111/tpj.12156>
- Chahtane, H., Zhang, B., Norberg, M., LeMasson, M., Thévenon, E., Bakó, L., Benlloch, R., Holmlund, M., Parcy, F., Nilsson, O., & Vachon, G. (2018). LEAFY activity is post-transcriptionally

- regulated by BLADE ON PETIOLE2 and CULLIN3 in Arabidopsis. *New Phytologist*, 220(2), 579-592. <https://doi.org/10.1111/nph.15329>
- Chen, H., & Pugh, B. F. (2021). What do Transcription Factors Interact With? *Journal of Molecular Biology*, 433(14), 166883. <https://doi.org/10.1016/j.jmb.2021.166883>
- Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6), 323-329. <https://doi.org/10.1016/j.ygeno.2012.04.003>
- Cheng, J.-Z., Zhou, Y.-P., Lv, T.-X., Xie, C.-P., & Tian, C.-E. (2017). Research progress on the autonomous flowering time pathway in Arabidopsis. *Physiology and Molecular Biology of Plants*, 23(3), 477-485. <https://doi.org/10.1007/s12298-017-0458-3>
- Cho, L.-H., Yoon, J., & An, G. (2017). The control of flowering time by environmental factors. *The Plant Journal*, 90(4), 708-719. <https://doi.org/10.1111/tpj.13461>
- Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology*, 15(9), e2003243. <https://doi.org/10.1371/journal.pbio.2003243>
- Costa, S., & Dean, C. (2019). Storing memories : The distinct phases of Polycomb-mediated silencing of Arabidopsis FLC. *Biochemical Society Transactions*, 47(4), 1187-1196. <https://doi.org/10.1042/BST20190255>
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50), 21931-21936. <https://doi.org/10.1073/pnas.1016071107>
- Crick, F. H. C. (1958). *On protein synthesis*. <https://doi.org/pmid:13580867>
- Denay, G., Gabrielle, T., Marie, L. M., Hicham, C., Sylvie, H., Irene, L.-V., Christian, W., Manuel, F.-Z. J., Rüdiger, S., Lohmann, J. U., & François, P. (2018). Control of stem-cell niche establishment in Arabidopsis flowers by REVOLUTA and the LEAFY-RAX1 module (p. 488114). <https://doi.org/10.1101/488114>
- Dickel, D. E., Ypsilanti, A. R., Pla, R., Zhu, Y., Barozzi, I., Mannion, B. J., Khin, Y. S., Fukuda-Yuzawa, Y., Plajzer-Frick, I., Pickle, C. S., Lee, E. A., Harrington, A. N., Pham, Q. T., Garvin, T. H., Kato, M., Osterwalder, M., Akiyama, J. A., Afzal, V., Rubenstein, J. L. R., ... Visel, A. (2018). Ultraconserved Enhancers Are Required for Normal Development. *Cell*, 172(3), 491-499.e15. <https://doi.org/10.1016/j.cell.2017.12.017>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR : Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21. <https://doi.org/10.1093/bioinformatics/bts635>
- Dolde, U., Muzzopappa, F., Delesalle, C., Neveu, J., Erdel, F., & Vert, G. (2023). LEAFY homeostasis is regulated via ubiquitin-dependent degradation and sequestration in cytoplasmic condensates. *IScience*, 26(6), 106880. <https://doi.org/10.1016/j.isci.2023.106880>
- Durfee, T., Roe, J. L., Sessions, R. A., Inouye, C., Serikawa, K., Feldmann, K. A., Weigel, D., & Zambryski, P. C. (2003). The F-box-containing protein UFO and AGAMOUS participate in antagonistic pathways governing early petal development in Arabidopsis. *Proceedings of the National Academy of Sciences*, 100(14), 8571-8576. <https://doi.org/10.1073/pnas.1033043100>
- Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., ... Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer [Preprint]. *Bioinformatics*. <https://doi.org/10.1101/2021.10.04.463034>
- Fonfría-Subirós, E., Acosta-Reyes, F., Saperas, N., Pous, J., Subirana, J. A., & Campos, J. L. (2012). Crystal Structure of a Complex of DNA with One AT-Hook of HMGA1. *PLOS ONE*, 7(5), e37120. <https://doi.org/10.1371/journal.pone.0037120>
- Freire-Rios, A., Tanaka, K., Crespo, I., Wijk, E. van der, Sizentsova, Y., Levitsky, V., Lindhoud, S., Fontana, M., Hohlbein, J., Boer, D. R., Mironova, V., & Weijers, D. (2020). Architecture of DNA

- elements mediating ARF transcription factor binding and auxin-responsive gene expression in Arabidopsis. *Proceedings of the National Academy of Sciences*, 117(39), 24557-24566. <https://doi.org/10.1073/pnas.2009554117>
- Freytes, S. N., Canelo, M., & Cerdán, P. D. (2021). Regulation of Flowering Time : When and Where? *Current Opinion in Plant Biology*, 63, 102049. <https://doi.org/10.1016/j.pbi.2021.102049>
- Fu, L.-Y., Zhu, T., Zhou, X., Yu, R., He, Z., Zhang, P., Wu, Z., Chen, M., Kaufmann, K., & Chen, D. (2022). ChIP-Hub provides an integrative platform for exploring plant regulome. *Nature Communications*, 13(1), 1. <https://doi.org/10.1038/s41467-022-30770-1>
- Gao, B., Chen, M., Li, X., & Zhang, J. (2019). Ancient duplications and grass-specific transposition influenced the evolution of LEAFY transcription factor genes. *Communications Biology*, 2(1), 1-10. <https://doi.org/10.1038/s42003-019-0469-4>
- Gao, H., Song, W., Severing, E., Vayssières, A., Huettel, B., Franzen, R., Richter, R., Chai, J., & Coupland, G. (2022). PIF4 enhances DNA binding of CDF2 to co-regulate target gene expression and promote Arabidopsis hypocotyl cell elongation. *Nature Plants*, 1-12. <https://doi.org/10.1038/s41477-022-01213-y>
- Gaspar, J. M. (2018). NGmerge : Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics*, 19(1), 536. <https://doi.org/10.1186/s12859-018-2579-2>
- Georgakopoulos-Soares, I., Deng, C., Agarwal, V., Chan, C. S. Y., Zhao, J., Inoue, F., & Ahituv, N. (2023). Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature Communications*, 14(1), 1. <https://doi.org/10.1038/s41467-023-37960-5>
- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., ... Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414), 7414. <https://doi.org/10.1038/nature11245>
- Goslin, K., Zheng, B., Serrano-Mislata, A., Rae, L., Ryan, P. T., Kwaśniewska, K., Thomson, B., Ó'Maoiléidigh, D. S., Madueño, F., Wellmer, F., & Graciet, E. (2017). Transcription Factor Interplay between LEAFY and APETALA1/CAULIFLOWER during Floral Initiation. *Plant Physiology*, 174(2), 1097-1109. <https://doi.org/10.1104/pp.17.00098>
- Grandi, V., Gregis, V., & Kater, M. M. (2012). Uncovering genetic and molecular interactions among floral meristem identity genes in Arabidopsis thaliana. *The Plant Journal*, 69(5), 881-893. <https://doi.org/10.1111/j.1365-313X.2011.04840.x>
- Greene, C. S., & Troyanskaya, O. G. (2010). Integrative Systems Biology for Data-Driven Knowledge Discovery. *Seminars in Nephrology*, 30(5), 443-454. <https://doi.org/10.1016/j.semnephrol.2010.07.002>
- Guo, T., Wang, N., Xue, Y., Guan, Q., van Nocker, S., Liu, C., & Ma, F. (2019). Overexpression of the RNA binding protein MhYTP1 in transgenic apple enhances drought tolerance and WUE by improving ABA level under drought condition. *Plant Science*, 280, 397-407. <https://doi.org/10.1016/j.plantsci.2018.11.018>
- Hajheidari, M., & Huang, S. C. (2022). Elucidating the biology of transcription factor–DNA interaction for accurate identification of cis-regulatory elements. *Current Opinion in Plant Biology*, 68, 102232. <https://doi.org/10.1016/j.pbi.2022.102232>
- Hamès, C., Ptchelkine, D., Grimm, C., Thevenon, E., Moyroud, E., Gérard, F., Martiel, J.-L., Benlloch, R., Parcy, F., & Müller, C. W. (2008). Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. *The EMBO Journal*, 27(19), 2628-2637. <https://doi.org/10.1038/emboj.2008.184>
- Hanano, S., & Goto, K. (2011). Arabidopsis TERMINAL FLOWER1 Is Involved in the Regulation of Flowering Time and Inflorescence Development through Transcriptional Repression. *The Plant Cell*, 23(9), 3172-3184. <https://doi.org/10.1105/tpc.111.088641>

- Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., Forczek, E., Joly-Lopez, Z., Steffen, J. G., Hazzouri, K. M., Dewar, K., Stinchcombe, J. R., Schoen, D. J., Wang, X., Schmutz, J., Town, C. D., Edger, P. P., Pires, J. C., Schumaker, K. S., ... Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8), 8. <https://doi.org/10.1038/ng.2684>
- He, Q., Johnston, J., & Zeitlinger, J. (2015). ChIP-nexus : A novel ChIP-exo protocol for improved detection of in vivo transcription factor binding footprints. *Nature biotechnology*, 33(4), 395-401. <https://doi.org/10.1038/nbt.3121>
- Héberlé, É., & Bardet, A. F. (2019). Sensitivity of transcription factors to DNA methylation. *Essays in Biochemistry*, 63(6), 727-741. <https://doi.org/10.1042/EBC20190033>
- Hepworth, S. R., Klenz, J. E., & Haughn, G. W. (2006). UFO in the Arabidopsis inflorescence apex is required for floral-meristem identity and bract suppression. *Planta*, 223(4), 769-778. <https://doi.org/10.1007/s00425-005-0138-3>
- Hirayama, T., & Shinozaki, K. (1996). A cdc5+ homolog of a higher plant, Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 93(23), 13371-13376. <https://doi.org/10.1073/pnas.93.23.13371>
- Honma, T., & Goto, K. (2000). The Arabidopsis floral homeotic gene PISTILLATA is regulated by discrete cis-elements responsive to induction and maintenance signals. *Development*, 127(10), 2021-2030. <https://doi.org/10.1242/dev.127.10.2021>
- Hope, C. M., Webber, J. L., Tokamov, S. A., & Rebay, I. (2018). Tuned polymerization of the transcription factor Yan limits off-DNA sequestration to confer context-specific repression. *eLife*, 7, e37545. <https://doi.org/10.7554/eLife.37545>
- Hu, H., Tian, S., Xie, G., Liu, R., Wang, N., Li, S., He, Y., & Du, J. (2021). TEM1 combinatorially binds to FLOWERING LOCUS T and recruits a Polycomb factor to repress the floral transition in Arabidopsis. *Proceedings of the National Academy of Sciences*, 118(35), e2103895118. <https://doi.org/10.1073/pnas.2103895118>
- Hugouvieux, V., & Zubietta, C. (2018). MADS transcription factors cooperate : Complexities of complex formation. *Journal of Experimental Botany*, 69(8), 1821-1823. <https://doi.org/10.1093/jxb/ery099>
- Hyun, Y., Richter, R., & Coupland, G. (2017). Competence to Flower : Age-Controlled Sensitivity to Environmental Cues. *Plant Physiology*, 173(1), 36-46. <https://doi.org/10.1104/pp.16.01523>
- Ingram, G. C., & Coena, E. S. (1995). *Parallels between UNUSUAL FLORAL ORGANS and FIMBRIATA, Genes Controlling Flower Development in Arabidopsis and Antirrhinum*. 10.
- Izawa, T. (2021). What is going on with the hormonal control of flowering in plants? *The Plant Journal*, 105(2), 431-445. <https://doi.org/10.1111/tpj.15036>
- Jalili, V., Matteucci, M., Masseroli, M., & Morelli, M. J. (2015). Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics*, 31(17), 2761-2769. <https://doi.org/10.1093/bioinformatics/btv293>
- Jha, P., Ochatt, S. J., & Kumar, V. (2020). WUSCHEL : A master regulator in plant growth signaling. *Plant Cell Reports*, 39(4), 431-444. <https://doi.org/10.1007/s00299-020-02511-5>
- Jin, R., Klasfeld, S., Zhu, Y., Fernandez Garcia, M., Xiao, J., Han, S.-K., Konkol, A., & Wagner, D. (2021). LEAFY is a pioneer transcription factor and licenses cell reprogramming to floral fate. *Nature Communications*, 12(1), 1. <https://doi.org/10.1038/s41467-020-20883-w>
- Johnson, M. P., Havaux, M., Triantaphylidès, C., Ksas, B., Pascal, A. A., Robert, B., Davison, P. A., Ruban, A. V., & Horton, P. (2007). Elevated Zeaxanthin Bound to Oligomeric LHCII Enhances the Resistance of Arabidopsis to Photooxidative Stress by a Lipid-protective, Antioxidant Mechanism. *Journal of Biological Chemistry*, 282(31), 22605-22618. <https://doi.org/10.1074/jbc.M702831200>
- Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., & Taipale, J. (2013). DNA-Binding Specificities of

- Human Transcription Factors. *Cell*, 152(1), 327-339.
<https://doi.org/10.1016/j.cell.2012.12.009>
- Jonas, F., Carmi, M., Krupkin, B., Steinberger, J., Brodsky, S., Jana, T., & Barkai, N. (2023). The molecular grammar of protein disorder guiding genome-binding locations. *Nucleic Acids Research*, 51(10), 4831-4844. <https://doi.org/10.1093/nar/gkad184>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. <https://doi.org/10.1038/s41586-021-03819-2>
- Jung, J.-H., Lee, H.-J., Ryu, J. Y., & Park, C.-M. (2016). SPL3/4/5 Integrate Developmental Aging and Photoperiodic Signals into the FT-FD Module in Arabidopsis Flowering. *Molecular Plant*, 9(12), 1647-1659. <https://doi.org/10.1016/j.molp.2016.10.014>
- Kagawa, W., & Kurumizaka, H. (2021). Structural basis for DNA sequence recognition by pioneer factors in nucleosomes. *Current Opinion in Structural Biology*, 71, 59-64.
<https://doi.org/10.1016/j.sbi.2021.05.011>
- Kardailsky, I., Shukla, V. K., Ahn, J. H., Dagenais, N., Christensen, S. K., Nguyen, J. T., Chory, J., Harrison, M. J., & Weigel, D. (1999). Activation Tagging of the Floral Inducer FT. *Science*, 286(5446), 1962-1965. <https://doi.org/10.1126/science.286.5446.1962>
- Karim, M. R., Hirota, A., Kwiatkowska, D., Tasaka, M., & Aida, M. (2009). A Role for Arabidopsis PUCHI in Floral Meristem Identity and Bract Suppression. *The Plant Cell*, 21(5), 1360-1372.
<https://doi.org/10.1105/tpc.109.067025>
- Karollus, A., Mauermeier, T., & Julien Gagneur. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1), 56. <https://doi.org/10.1186/s13059-023-02899-9>
- Kaufmann, K., & Airoidi, C. A. (2018). Master Regulatory Transcription Factors in Plant Development : A Blooming Perspective. In N. Yamaguchi (Éd.), *Plant Transcription Factors : Methods and Protocols* (p. 3-22). Springer. https://doi.org/10.1007/978-1-4939-8657-6_1
- Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1), 1. <https://doi.org/10.1038/s41467-019-09982-5>
- Kim, S., & Shendure, J. (2019). Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Molecular Cell*, 76(2), 306-319. <https://doi.org/10.1016/j.molcel.2019.08.010>
- Klasfeld, S., Roulé, T., & Wagner, D. (2022). Greenscreen : A simple method to remove artifactual signals and enrich for true peaks in genomic datasets including ChIP-seq data. *The Plant Cell*, 34(12), 4795-4815. <https://doi.org/10.1093/plcell/koac282>
- Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M., & Araki, T. (1999). A Pair of Related Genes with Antagonistic Roles in Mediating Flowering Signals. *Science*, 286(5446), 1960-1962.
<https://doi.org/10.1126/science.286.5446.1960>
- Krämer, U. (2015). Planting molecular functions in an ecological context with Arabidopsis thaliana. *eLife*, 4, e06100. <https://doi.org/10.7554/eLife.06100>
- Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M., Rosen, B. D., Cheng, C.-Y., Moreira, W., Mock, S. A., Stubbs, J., Sullivan, J. M., Krampis, K., Miller, J. R., Micklem, G., Vaughn, M., & Town, C. D. (2015). Araport : The Arabidopsis Information Portal. *Nucleic Acids Research*, 43(Database issue), D1003-D1009.
<https://doi.org/10.1093/nar/gku1200>
- Krogan, N. T., Hogan, K., & Long, J. A. (2012). APETALA2 negatively regulates multiple floral organ identity genes in Arabidopsis by recruiting the co-repressor TOPLESS and the histone deacetylase HDA19. *Development*, 139(22), 4180-4190. <https://doi.org/10.1242/dev.085407>
- Lai, X., Blanc-Mathieu, R., GrandVuillemin, L., Huang, Y., Stigliani, A., Lucas, J., Thévenon, E., Loue-Manifel, J., Turchi, L., Daher, H., Brun-Hernandez, E., Vachon, G., Latrasse, D., Benhamed, M.,

- Dumas, R., Zubieta, C., & Parcy, F. (2021). The LEAFY floral regulator displays pioneer transcription factor properties. *Molecular Plant*. <https://doi.org/10.1016/j.molp.2021.03.004>
- Lai, X., Daher, H., Galien, A., Hugouvieux, V., & Zubieta, C. (2019). Structural Basis for Plant MADS Transcription Factor Oligomerization. *Computational and Structural Biotechnology Journal*, 17, 946-953. <https://doi.org/10.1016/j.csbj.2019.06.014>
- Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., Kaufmann, K., & Parcy, F. (2019). Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. *Molecular Plant*, 12(6), 743-763. <https://doi.org/10.1016/j.molp.2018.10.010>
- Lai, X., Vega-Leon, R., Hugouvieux, V., Blanc-Mathieu, R., Wal, F. van der, Lucas, J., Silva, C. S., Jourdain, A., Muino, J., Nanao, M. H., Immink, R., Kaufmann, K., Parcy, F., Smaczniak, C., & Zubieta, C. (2021). The Intervening Domain Is Required For DNA-binding and Functional Identity of Plant MADS Transcription Factors. *BioRxiv*, 2021.03.10.434815. <https://doi.org/10.1101/2021.03.10.434815>
- Lamb, R. S., Hill, T. A., Tan, Q. K.-G., & Irish, V. F. (2002). Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. *Development*, 129(9), 2079-2086. <https://doi.org/10.1242/dev.129.9.2079>
- Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4), 650-665. <https://doi.org/10.1016/j.cell.2018.01.029>
- Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The Arabidopsis Information Resource (TAIR) : Improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1), D1202-D1210. <https://doi.org/10.1093/nar/gkr1090>
- Laufs, P., Coen, E., Kronenberger, J., Traas, J., & Doonan, J. (2003). Separable roles of UFO during floral development revealed by conditional restoration of gene function. *Development*, 130(4), 785-796. <https://doi.org/10.1242/dev.00295>
- Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3), 204-220. <https://doi.org/10.1038/nrg2719>
- Lee, I., Wolfe, D. S., Nilsson, O., & Weigel, D. (1997). A LEAFY co-regulator encoded by UNUSUAL FLORAL ORGANS. *Current Biology*, 7(2), 95-104. [https://doi.org/10.1016/S0960-9822\(06\)00053-4](https://doi.org/10.1016/S0960-9822(06)00053-4)
- Lee, J., Oh, M., Park, H., & Lee, I. (2008). SOC1 translocated to the nucleus by interaction with AGL24 directly regulates LEAFY. *The Plant Journal*, 55(5), 832-843. <https://doi.org/10.1111/j.1365-3113.2008.03552.x>
- Li, D., Zhang, H., Mou, M., Chen, Y., Xiang, S., Chen, L., & Yu, D. (2019). Arabidopsis Class II TCP Transcription Factors Integrate with the FT–FD Module to Control Flowering1. *Plant Physiology*, 181(1), 97-111. <https://doi.org/10.1104/pp.19.00252>
- Li, S. (2015). The Arabidopsis thaliana TCP transcription factors : A broadening horizon beyond development. *Plant Signaling & Behavior*, 10(7), e1044192. <https://doi.org/10.1080/15592324.2015.1044192>
- Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 6. <https://doi.org/10.1038/nrg3920>
- Lin, Z., Yin, K., Zhu, D., Chen, Z., Gu, H., & Qu, L.-J. (2007). AtCDC5 regulates the G2 to M transition of the cell cycle and is critical for the function of Arabidopsis shoot apical meristem. *Cell Research*, 17(9), 9. <https://doi.org/10.1038/cr.2007.71>
- Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., ... Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370), 7370. <https://doi.org/10.1038/nature10530>

- Liu, L., Zhang, G., He, S., & Hu, X. (2021). TSPTFBS : A Docker image for trans-species prediction of transcription factor binding sites in plants. *Bioinformatics*, 37(2), 260-262. <https://doi.org/10.1093/bioinformatics/btaa1100>
- Lloyd, J. P. B., & Lister, R. (2021). Epigenome plasticity in plants. *Nature Reviews Genetics*, 1-14. <https://doi.org/10.1038/s41576-021-00407-y>
- Lohmann, J. U., Hong, R. L., Hobe, M., Busch, M. A., Parcy, F., Simon, R., & Weigel, D. (2001). A Molecular Link between Stem Cell Regulation and Floral Patterning in Arabidopsis. *Cell*, 105(6), 793-803. [https://doi.org/10.1016/S0092-8674\(01\)00384-1](https://doi.org/10.1016/S0092-8674(01)00384-1)
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M. W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S. M., & Deal, R. B. (2018). Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. *The Plant Cell*, 30(1), 15-36. <https://doi.org/10.1105/tpc.17.00581>
- Maizel, A. (2005). The Floral Regulator LEAFY Evolves by Substitutions in the DNA Binding Domain. *Science*, 308(5719), 260-263. <https://doi.org/10.1126/science.1108229>
- Mathelier, A., & Wasserman, W. W. (2013). The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Computational Biology*, 9(9), e1003214. <https://doi.org/10.1371/journal.pcbi.1003214>
- Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., & Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*, 3(3), 278-286.e4. <https://doi.org/10.1016/j.cels.2016.07.001>
- Matías-Hernández, L., Aguilar-Jaramillo, A. E., Marín-González, E., Suárez-López, P., & Pelaz, S. (2014). RAV genes : Regulation of floral induction and beyond. *Annals of Botany*, 114(7), 1459-1470. <https://doi.org/10.1093/aob/mcu069>
- Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., & Wingender, E. (2006). TRANSFAC and its module TRANSCompel : Transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue), D108-110. <https://doi.org/10.1093/nar/gkj143>
- McKim, S., & Hay, A. (2010). Patterning and evolution of floral structures—Marking time. *Current Opinion in Genetics & Development*, 20(4), 448-453. <https://doi.org/10.1016/j.gde.2010.04.007>
- McKim, S. M., Stenvik, G.-E., Butenko, M. A., Kristiansen, W., Cho, S. K., Hepworth, S. R., Aalen, R. B., & Haughn, G. W. (2008). The BLADE-ON-PETIOLE genes are essential for abscission zone formation in Arabidopsis. *Development*, 135(8), 1537-1546. <https://doi.org/10.1242/dev.012807>
- Mendes, M. A., Guerra, R. F., Berns, M. C., Manzo, C., Masiero, S., Finzi, L., Kater, M. M., & Colombo, L. (2013). MADS domain transcription factors mediate short-range DNA looping that is essential for target gene expression in Arabidopsis. *The Plant cell*, 25(7), 2560-2572. <https://doi.org/10.1105/tpc.112.108688>
- Meyer, P., & Saez-Rodriguez, J. (2021). Advances in systems biology modeling : 10 years of crowdsourcing DREAM challenges. *Cell Systems*, 12(6), 636-653. <https://doi.org/10.1016/j.cels.2021.05.015>
- Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., ... Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 452(7190), 7190. <https://doi.org/10.1038/nature06856>

- Minguet, E. G., Segard, S., Charavay, C., & Parcy, F. (2015). MORPHEUS, a Webtool for Transcription Factor Binding Analysis Using Position Weight Matrices with Dependency. *PLOS ONE*, 10(8), e0135586. <https://doi.org/10.1371/journal.pone.0135586>
- Morgunova, E., & Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology*, 47, 1-8. <https://doi.org/10.1016/j.sbi.2017.03.006>
- Moyroud, E., Kusters, E., Monniaux, M., Koes, R., & Parcy, F. (2010). LEAFY blossoms. *Trends in Plant Science*, 15(6), 346-352. <https://doi.org/10.1016/j.tplants.2010.03.007>
- Moyroud, E., Minguet, E. G., Ott, F., Yant, L., Posé, D., Monniaux, M., Blanchet, S., Bastien, O., Thévenon, E., Weigel, D., Schmid, M., & Parcy, F. (2011). Prediction of Regulatory Interactions from Genome Sequences Using a Biophysical Model for the *Arabidopsis* LEAFY Transcription Factor. *The Plant Cell*, 23(4), 1293-1306. <https://doi.org/10.1105/tpc.111.083329>
- Moyroud, E., Tichtinsky, G., & Parcy, F. (2009). The LEAFY Floral Regulators in Angiosperms : Conserved Proteins with Diverse Roles. *Journal of Plant Biology*, 52(3), 177-185. <https://doi.org/10.1007/s12374-009-9028-8>
- Murat, F., Peer, Y. V. de, & Salse, J. (2012). Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and Processes. *Genome Biology and Evolution*, 4(9), 917-928. <https://doi.org/10.1093/gbe/evs066>
- Narita, T., Higashijima, Y., Kilic, S., Liebner, T., Walter, J., & Choudhary, C. (2023). Acetylation of histone H2B marks active enhancers and predicts CBP/p300 target genes. *Nature Genetics*, 55(4), 4. <https://doi.org/10.1038/s41588-023-01348-4>
- Nei, M., Xu, P., & Glazko, G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences*, 98(5), 2497-2502. <https://doi.org/10.1073/pnas.051611498>
- Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. M., & Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4, e04837. <https://doi.org/10.7554/eLife.04837>
- Noor, E., Cherkaoui, S., & Sauer, U. (2019). Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15, 39-47. <https://doi.org/10.1016/j.coisb.2019.03.007>
- Oksuz, O., Henninger, J. E., Warneford-Thomson, R., Zheng, M. M., Erb, H., Vancura, A., Overholt, K. J., Hawken, S. W., Banani, S. F., Lauman, R., Reich, L. N., Robertson, A. L., Hannett, N. M., Lee, T. I., Zon, L. I., Bonasio, R., & Young, R. A. (2023). Transcription factors interact with RNA to regulate genes. *Molecular Cell*, 83(14), 2449-2463.e13. <https://doi.org/10.1016/j.molcel.2023.06.012>
- O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., & Ecker, J. R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5), 1280-1292. <https://doi.org/10.1016/j.cell.2016.04.038>
- Omidbakhshfard, M. A., Proost, S., Fujikura, U., & Mueller-Roeber, B. (2015). Growth-Regulating Factors (GRFs) : A Small Transcription Factor Family with Important Functions in Plant Biology. *Molecular Plant*, 8(7), 998-1010. <https://doi.org/10.1016/j.molp.2015.01.013>
- Otsuga, D., DeGuzman, B., Prigge, M. J., Drews, G. N., & Clark, S. E. (2001). REVOLUTA regulates meristem initiation at lateral positions. *The Plant Journal*, 25(2), 223-236. <https://doi.org/10.1111/j.1365-313X.2001.00959.x>
- Pajoro, A., Madrigal, P., Muiño, J. M., Matus, J. T., Jin, J., Mecchia, M. A., Debernardi, J. M., Palatnik, J. F., Balazadeh, S., Arif, M., Ó'Maoiléidigh, D. S., Wellmer, F., Krajewski, P., Riechmann, J.-L., Angenent, G. C., & Kaufmann, K. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biology*, 15(3), R41. <https://doi.org/10.1186/gb-2014-15-3-r41>
- Palma, K., Zhao, Q., Cheng, Y. T., Bi, D., Monaghan, J., Cheng, W., Zhang, Y., & Li, X. (2007). Regulation of plant innate immunity by three proteins in a complex conserved across the plant and

- animal kingdoms. *Genes & Development*, 21(12), 1484-1493.
<https://doi.org/10.1101/gad.1559607>
- Parcy, F., Nilsson, O., Busch, M. A., Lee, I., & Weigel, D. (1998). A genetic framework for floral patterning. *Nature*, 395(6702), 561-566. <https://doi.org/10.1038/26903>
- Park, D., Lee, Y., Bhupindersingh, G., & Iyer, V. R. (2013). Widespread Misinterpretable ChIP-seq Bias in Yeast. *PLOS ONE*, 8(12), e83506. <https://doi.org/10.1371/journal.pone.0083506>
- Park, P. J. (2009). ChIP-seq : Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10), 10. <https://doi.org/10.1038/nrg2641>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 4.
<https://doi.org/10.1038/nmeth.4197>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825-2830.
- Pohl, A., & Beato, M. (2014). bwtool : A tool for bigWig files. *Bioinformatics*, 30(11), 1618-1619.
<https://doi.org/10.1093/bioinformatics/btu056>
- Pollard, D. A., Bergman, C. M., Stoye, J., Celniker, S. E., & Eisen, M. B. (2004). Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, 5(1), 6.
<https://doi.org/10.1186/1471-2105-5-6>
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1), 110-121.
<https://doi.org/10.1101/gr.097857.109>
- Prigge, M. J., Otsuga, D., Alonso, J. M., Ecker, J. R., Drews, G. N., & Clark, S. E. (2005). Class III Homeodomain-Leucine Zipper Gene Family Members Have Overlapping, Antagonistic, and Distinct Roles in Arabidopsis Development. *The Plant Cell*, 17(1), 61-76.
<https://doi.org/10.1105/tpc.104.026161>
- Puranik, S., Acajjaoui, S., Conn, S., Costa, L., Conn, V., Vial, A., Marcellin, R., Melzer, R., Brown, E., Hart, D., Theißen, G., Silva, C. S., Parcy, F., Dumas, R., Nanao, M., & Zubieta, C. (2014). Structural Basis for the Oligomerization of the MADS Domain Transcription Factor SEPALLATA3 in Arabidopsis. *The Plant Cell*, 26(9), 3603-3615.
<https://doi.org/10.1105/tpc.114.127910>
- Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., & Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*, 38(10), 2943-2945.
<https://doi.org/10.1093/bioinformatics/btac166>
- Qiao, P., Bourgault, R., Mohammadi, M., Matschi, S., Philippe, G., Smith, L. G., Gore, M. A., Molina, I., & Scanlon, M. J. (2020). Transcriptomic network analyses shed light on the regulation of cuticle development in maize leaves. *Proceedings of the National Academy of Sciences*, 117(22), 12464-12471. <https://doi.org/10.1073/pnas.2004945117>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools : A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842. <https://doi.org/10.1093/bioinformatics/btq033>
- Rajeev, L., Luning, E. G., & Mukhopadhyay, A. (2014). DNA-affinity-purified Chip (DAP-chip) Method to Determine Gene Targets for Bacterial Two component Regulatory Systems. *JoVE (Journal of Visualized Experiments)*, 89, e51715. <https://doi.org/10.3791/51715>
- Ravel, C., Fiquet, S., Boudet, J., Dardevet, M., Vincent, J., Merlino, M., Michard, R., & Martre, P. (2014). Conserved cis-regulatory modules in promoters of genes encoding wheat high-molecular-weight glutenin subunits. *Frontiers in Plant Science*, 5.
<https://www.frontiersin.org/articles/10.3389/fpls.2014.00621>
- Reddy, G. V. (2008). Live-imaging stem-cell homeostasis in the Arabidopsis shoot apex. *Current Opinion in Plant Biology*, 11(1), 88-93. <https://doi.org/10.1016/j.pbi.2007.10.012>

- Reineke, A. R., Bornberg-Bauer, E., & Gu, J. (2011). Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. *Nucleic Acids Research*, 39(14), 6029-6043. <https://doi.org/10.1093/nar/gkr179>
- Reneker, J., Lyons, E., Conant, G. C., Pires, J. C., Freeling, M., Shyu, C.-R., & Korkin, D. (2012). Long identical multispecies elements in plant and animal genomes. *Proceedings of the National Academy of Sciences*, 109(19), E1183-E1191. <https://doi.org/10.1073/pnas.1121356109>
- Rhee, H. S., & Pugh, B. F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6), 1408-1419. <https://doi.org/10.1016/j.cell.2011.11.013>
- Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C.-Z., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O. J., Samaha, R. R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J. Z., Ghandehari, D., Sherman, B. K., & Yu, G.-L. (2000). Arabidopsis Transcription Factors : Genome-Wide Comparative Analysis Among Eukaryotes. *Science*, 290(5499), 2105-2110. <https://doi.org/10.1126/science.290.5499.2105>
- Rieu, P., Turchi, L., Thévenon, E., Zarkadas, E., Nanao, M., Chahtane, H., Tichtinsky, G., Lucas, J., Blanc-Mathieu, R., Zubieta, C., Schoehn, G., & Parcy, F. (2023). The F-box protein UFO controls flower development by redirecting the master transcription factor LEAFY to new cis-elements. *Nature Plants*, 9(2), 2. <https://doi.org/10.1038/s41477-022-01336-2>
- Risseuw, E., Venglat, P., Xiang, D., Komendant, K., Daskalchuk, T., Babic, V., Crosby, W., & Datla, R. (2013). An Activated Form of UFO Alters Leaf Development and Produces Ectopic Floral and Inflorescence Meristems. *PLOS ONE*, 8(12), e83807. <https://doi.org/10.1371/journal.pone.0083807>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47-e47. <https://doi.org/10.1093/nar/gkv007>
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., & Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8), 651-657. <https://doi.org/10.1038/nmeth1068>
- Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., & Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268), 7268. <https://doi.org/10.1038/nature08473>
- Roider, H. G., Kanhere, A., Manke, T., & Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2), 134-141. <https://doi.org/10.1093/bioinformatics/btl565>
- Rozière, J., Guichard, C., Brunaud, V., Martin, M.-L., & Coursol, S. (2022). A comprehensive map of preferentially located motifs reveals distinct proximal cis-regulatory sequences in plants. *Frontiers in Plant Science*, 13. <https://www.frontiersin.org/articles/10.3389/fpls.2022.976371>
- RStudio Team. (2020). *RStudio : Integrated Development Environment for R*. <http://www.rstudio.com/>
- Samee, Md. A. H. (2023). Noncanonical binding of transcription factors : Time to revisit specificity? *Molecular Biology of the Cell*, 34(9), pe4. <https://doi.org/10.1091/mbc.E22-08-0325>
- Sandelin, A. (2004). JASPAR : An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001), 91D - 94. <https://doi.org/10.1093/nar/gkh012>
- Sayou, C., Monniaux, M., Nanao, M. H., Moyroud, E., Brockington, S. F., Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y., Wong, G. K.-S., Weigel, D., Parcy, F., & Dumas, R. (2014). A Promiscuous Intermediate Underlies the Evolution of LEAFY DNA Binding Specificity. *Science*, 343(6171), 645-648. <https://doi.org/10.1126/science.1248229>
- Sayou, C., Nanao, M. H., Jamin, M., Posé, D., Thévenon, E., Grégoire, L., Tichtinsky, G., Denay, G., Ott, F., Peirats Llobet, M., Schmid, M., Dumas, R., & Parcy, F. (2016). A SAM oligomerization domain shapes the genomic binding landscape of the LEAFY transcription factor. *Nature Communications*, 7(1), 1. <https://doi.org/10.1038/ncomms11222>

- Schmid, M., Henz, S., Davison, T., Pape, U., Vingron, M., Schölkopf, B., Weigel, D., & Lohmann, U. (2004). *AtGenExpress : Expression atlas of Arabidopsis Development*. 485.
- Schmid, M., Uhlenhaut, N. H., Godard, F., Demar, M., Bressan, R., Weigel, D., & Lohmann, J. U. (2003). Dissection of floral induction pathways using global expression analysis. *Development*, 130(24), 6001-6012. <https://doi.org/10.1242/dev.00842>
- Schneider, T. D., & Stephens, R. M. (1990). Sequence logos : A new way to display consensus sequences. *Nucleic Acids Research*, 18(20), 6097-6100. <https://doi.org/10.1093/nar/18.20.6097>
- Schultz, E. A., & Haughn, G. W. (1991). *LEAFY, a Homeotic Gene That Regulates Inflorescence Development in Arabidopsis*.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shen, W., Pan, J., Wang, G., & Li, X. (2021). Deep learning-based prediction of TFBSs in plants. *Trends in Plant Science*, 0(0). <https://doi.org/10.1016/j.tplants.2021.06.016>
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8), 1034-1050. <https://doi.org/10.1101/gr.3715005>
- Singh, G., Mullany, S., Moorthy, S. D., Zhang, R., Mehdi, T., Tian, R., Duncan, A. G., Moses, A. M., & Mitchell, J. A. (2021). A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells. *Genome Research*, 31(4), 564-575. <https://doi.org/10.1101/gr.272468.120>
- Siriwardana, N. S., & Lamb, R. S. (2012). The poetry of reproduction : The role of LEAFY in Arabidopsis thaliana flower formation. *International Journal of Developmental Biology*, 56(4), 4. <https://doi.org/10.1387/ijdb.113450ns>
- Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6, e21856. <https://doi.org/10.7554/eLife.21856>
- Slyper, M., Porter, C. B. M., Ashenberg, O., Waldman, J., Drokhyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., Vigneau, S., Jané-Valbuena, J., Tickle, T. L., Napolitano, S., Su, M.-J., Patel, A. G., Karlstrom, A., Gritsch, S., Nomura, M., ... Regev, A. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature Medicine*, 26(5), 5. <https://doi.org/10.1038/s41591-020-0844-1>
- Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R., Busscher, M., Busscher-Lange, J., Dinh, Q. D. (Peter), Liu, S., Westphal, A. H., Boeren, S., Parcy, F., Xu, L., Carles, C. C., Angenent, G. C., & Kaufmann, K. (2012). Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proceedings of the National Academy of Sciences*, 109(5), 1560-1565. <https://doi.org/10.1073/pnas.1112871109>
- Smet, D., Opdebeeck, H., & Vandepoele, K. (2023). Predicting transcriptional responses to heat and drought stress from genomic features using a machine learning approach in rice. *Frontiers in Plant Science*, 14. <https://www.frontiersin.org/articles/10.3389/fpls.2023.1212073>
- Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools : Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3), 491-499. <https://doi.org/10.1101/gr.209601.116>
- Snetkova, V., Ypsilanti, A. R., Akiyama, J. A., Mannion, B. J., Plajzer-Frick, I., Novak, C. S., Harrington, A. N., Pham, Q. T., Kato, M., Zhu, Y., Godoy, J., Meky, E., Hunter, R. D., Shi, M., Kvon, E. Z., Afzal, V., Tran, S., Rubenstein, J. L. R., Visel, A., ... Dickel, D. E. (2021). Ultraconserved enhancer function does not require perfect sequence conservation. *Nature Genetics*, 53(4), 521-528. <https://doi.org/10.1038/s41588-021-00812-3>
- Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq : Transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 1521. <https://doi.org/10.12688/f1000research.7563.1>

- Song, L., & Crawford, G. E. (2010). DNase-seq : A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, 2010(2), pdb.prot5384. <https://doi.org/10.1101/pdb.prot5384>
- Souer, E., Rebocho, A. B., Bliet, M., Kusters, E., de Bruin, R. A. M., & 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. *The Plant Cell*, 20(8), 2033-2048. <https://doi.org/10.1105/tpc.108.060871>
- Soufi, A., Donahue, G., & Zaret, K. S. (2012). Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell*, 151(5), 994-1004. <https://doi.org/10.1016/j.cell.2012.09.045>
- Spiegel, J., Cuesta, S. M., Adhikari, S., Hänsel-Hertsch, R., Tannahill, D., & Balasubramanian, S. (2021). G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biology*, 22(1), 117. <https://doi.org/10.1186/s13059-021-02324-z>
- Spitz, F., & Furlong, E. E. M. (2012). Transcription factors : From enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9), 9. <https://doi.org/10.1038/nrg3207>
- Staller, M. V. (2022). Transcription factors perform a 2-step search of the nucleus. *Genetics*, 222(2), iyac111. <https://doi.org/10.1093/genetics/iyac111>
- Stigliani, A., Martin-Arevalillo, R., Lucas, J., Bessy, A., Vinos-Poyo, T., Mironova, V., Vernoux, T., Dumas, R., & Parcy, F. (2019). Capturing Auxin Response Factors Syntax Using DNA Binding Models. *Molecular Plant*, 12(6), 822-832. <https://doi.org/10.1016/j.molp.2018.09.010>
- Stormo, G. D. (2015). DNA Motif Databases and Their Uses. *Current Protocols in Bioinformatics*, 51(1), 2.15.1-2.15.6. <https://doi.org/10.1002/0471250953.bi0215s51>
- Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyannopoulos, J. A., & Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology*, 3-4, 40-47. <https://doi.org/10.1016/j.cpb.2015.10.001>
- Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S., & Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Research*, 21(7), 1139-1149. <https://doi.org/10.1101/gr.119016.110>
- Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21), e140. <https://doi.org/10.1093/nar/gkv711>
- Teytelman, L., Thurtle, D. M., Rine, J., & van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, 110(46), 18602-18607. <https://doi.org/10.1073/pnas.1316064110>
- The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), 6814. <https://doi.org/10.1038/35048692>
- Theißen, G., Melzer, R., & Rümpler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development : Linking plant development and evolution. *Development*, 143(18), 3259-3271. <https://doi.org/10.1242/dev.134080>
- Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., Schwartz, M. S., Weber, R. J., Kent, W. J., Karolchik, D., Bruen, T. C., Bevan, R., Cutler, D. J., ... Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950), 6950. <https://doi.org/10.1038/nature01858>
- Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., & Gao, G. (2020). PlantRegMap : Charting functional regulatory maps in plants. *Nucleic Acids Research*, 48(D1), D1104-D1113. <https://doi.org/10.1093/nar/gkz1020>
- Tu, X., Ren, S., Shen, W., Li, J., Li, Y., Li, C., Li, Y., Zong, Z., Xie, W., Grierson, D., Fei, Z., Giovannoni, J., Li, P., & Zhong, S. (2022). Limited conservation in cross-species comparison of GLK

- transcription factor binding suggested wide-spread cistrome divergence. *Nature Communications*, 13(1), 1. <https://doi.org/10.1038/s41467-022-35438-4>
- van Dijk, A. D. J., Kootstra, G., Kruijer, W., & de Ridder, D. (2021). Machine learning in plant science and plant breeding. *iScience*, 24(1), 101890. <https://doi.org/10.1016/j.isci.2020.101890>
- Vandepoele, K., Casneuf, T., & Van de Peer, Y. (2006). Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biology*, 7(11), R103. <https://doi.org/10.1186/gb-2006-7-11-r103>
- Velde, J. V. de, Bel, M. V., Vaneechoutte, D., & Vandepoele, K. (2016). A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. *Plant Physiology*, 171(4), 2586-2598. <https://doi.org/10.1104/pp.16.00821>
- Velde, J. V. de, Heyndrickx, K. S., & Vandepoele, K. (2014). Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis. *The Plant Cell*, 26(7), 2729-2745. <https://doi.org/10.1105/tpc.114.127001>
- Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., & Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature*, 583(7818), 7818. <https://doi.org/10.1038/s41586-020-2528-x>
- Wagh, K., Garcia, D. A., & Upadhyaya, A. (2021). Phase separation in transcription factor dynamics and chromatin organization. *Current Opinion in Structural Biology*, 71, 148-155. <https://doi.org/10.1016/j.sbi.2021.06.009>
- Wagner, D., Sablowski, R. W. M., & Meyerowitz, E. M. (1999). Transcriptional Activation of APETALA1 by LEAFY. *Science*, 285(5427), 582-584. <https://doi.org/10.1126/science.285.5427.582>
- Wang, M., Li, Q., & Liu, L. (2023). Factors and Methods for the Detection of Gene Expression Regulation. *Biomolecules*, 13(2), 2. <https://doi.org/10.3390/biom13020304>
- Wang, M., & Zhang, Y. (2021). *Tn5 transposase-based epigenomic profiling methods are prone to open chromatin bias* [Preprint]. Genomics. <https://doi.org/10.1101/2021.07.09.451758>
- Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4), 4. <https://doi.org/10.1038/nrg1315>
- Watson, D. S. (2022). Interpretable machine learning for genomics. *Human Genetics*, 141(9), 1499-1513. <https://doi.org/10.1007/s00439-021-02387-9>
- Weigel, D., Alvarez, J., Smyth, D. R., Yanofsky, M. F., & Meyerowitz, E. M. (1992). LEAFY controls floral meristem identity in Arabidopsis. *Cell*, 69(5), 843-859. [https://doi.org/10.1016/0092-8674\(92\)90295-N](https://doi.org/10.1016/0092-8674(92)90295-N)
- Weigel, D., & Nilsson, O. (1995). A developmental switch sufficient for flower initiation in diverse plants. *NATURE*, 377.
- Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., ... Hughes, T. R. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6), 1431-1443. <https://doi.org/10.1016/j.cell.2014.08.009>
- Wellmer, F., Graciet, E., & Riechmann, J. L. (2014). Specification of floral organs in Arabidopsis. *Journal of Experimental Botany*, 65(1), 1-9. <https://doi.org/10.1093/jxb/ert385>
- Wigge, P. A. (2005). Integration of Spatial and Temporal Information During Floral Induction in Arabidopsis. *Science*, 309(5737), 1056-1059. <https://doi.org/10.1126/science.1114358>
- William, D. A., Su, Y., Smith, M. R., Lu, M., Baldwin, D. A., & Wagner, D. (2004). Genomic identification of direct target genes of LEAFY. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6), 1775-1780. <https://doi.org/10.1073/pnas.0307842100>
- Wingender, E., Schoeps, T., Haubrock, M., & Dönitz, J. (2015). TFClass : A classification of human transcription factors and their rodent orthologs. *Nucleic Acids Research*, 43(Database issue), D97-102. <https://doi.org/10.1093/nar/gku1064>

- Winter, C. M., Austin, R. S., Blanvillain-Baufumé, S., Reback, M. A., Monniaux, M., Wu, M.-F., Sang, Y., Yamaguchi, A., Yamaguchi, N., Parker, J. E., Parcy, F., Jensen, S. T., Li, H., & Wagner, D. (2011). LEAFY Target Genes Reveal Floral Regulatory Logic, cis Motifs, and a Link to Biotic Stimulus Response. *Developmental Cell*, 20(4), 430-443. <https://doi.org/10.1016/j.devcel.2011.03.019>
- Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements : Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1), 1. <https://doi.org/10.1038/nrg3095>
- Wong, E. S., Zheng, D., Tan, S. Z., Bower, N. I., Garside, V., Vanwalleghem, G., Gaiti, F., Scott, E., Hogan, B. M., Kikuchi, K., McGlinn, E., Francois, M., & Degnan, B. M. (2020). Deep conservation of the enhancer regulatory code in animals. *Science*, 370(6517), eaax8137. <https://doi.org/10.1126/science.aax8137>
- Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., & Elgar, G. (2004). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biology*, 3(1), e7. <https://doi.org/10.1371/journal.pbio.0030007>
- Workman, C. T., Yin, Y., Corcoran, D. L., Ideker, T., Stormo, G. D., & Benos, P. V. (2005). enoLOGOS : A versatile web tool for energy normalized sequence logos. *Nucleic Acids Research*, 33(suppl_2), W389-W392. <https://doi.org/10.1093/nar/gki439>
- Wu, M.-F., Sang, Y., Bezhani, S., Yamaguchi, N., Han, S.-K., Li, Z., Su, Y., Slewinski, T. L., & Wagner, D. (2012). SWI2/SNF2 chromatin remodeling ATPases overcome polycomb repression and control floral organ identity with the LEAFY and SEPALLATA3 transcription factors. *Proceedings of the National Academy of Sciences*, 109(9), 3576-3581. <https://doi.org/10.1073/pnas.1113409109>
- Wu, X., Dinneny, J. R., Crawford, K. M., Rhee, Y., Citovsky, V., Zambryski, P. C., & Weigel, D. (2003). Modes of intercellular transcription factor movement in the Arabidopsis apex. *Development*, 130(16), 3735-3745. <https://doi.org/10.1242/dev.00577>
- Wu, Z., & Irizarry, R. (2022). *gcrma : Background Adjustment Using Sequence Information*.
- Yamaguchi, N., Jeong, C. W., Nole-Wilson, S., Krizek, B. A., & Wagner, D. (2016). AINTEGUMENTA and AINTEGUMENTA-LIKE6/PLETHORA3 Induce LEAFY Expression in Response to Auxin to Promote the Onset of Flower Formation in Arabidopsis. *Plant Physiology*, 170(1), 283-293. <https://doi.org/10.1104/pp.15.00969>
- Yamaguchi, N., Wu, M.-F., Winter, C. M., Berns, M. C., Nole-Wilson, S., Yamaguchi, A., Coupland, G., Krizek, B. A., & Wagner, D. (2013). A Molecular Framework for Auxin-Mediated Initiation of Flower Primordia. *Developmental Cell*, 24(3), 271-282. <https://doi.org/10.1016/j.devcel.2012.12.017>
- Yan, W., Chen, D., Schumacher, J., Durantini, D., Engelhorn, J., Chen, M., Carles, C. C., & Kaufmann, K. (2019). Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nature Communications*, 10(1), 1. <https://doi.org/10.1038/s41467-019-09513-2>
- Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C., & Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337), eaaj2239. <https://doi.org/10.1126/science.aaj2239>
- Yocca, A. E., Lu, Z., Schmitz, R. J., Freeling, M., & Edger, P. P. (2021). Evolution of Conserved Noncoding Sequences in Arabidopsis thaliana. *Molecular Biology and Evolution*, 38(7), 2692-2703. <https://doi.org/10.1093/molbev/msab042>
- Zhang, S., Xie, M., Ren, G., & Yu, B. (2013). CDC5, a DNA binding protein, positively regulates posttranscriptional processing and/or transcription of primary microRNA transcripts. *Proceedings of the National Academy of Sciences*, 110(43), 17588-17593. <https://doi.org/10.1073/pnas.1310644110>

- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), R137. <https://doi.org/10.1186/gb-2008-9-9-r137>
- Zhang, Y., Wang, Z., Zeng, Y., Zhou, J., & Zou, Q. (2021). *High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method*. 12.
- Zhao, H., Tu, Z., Liu, Y., Zong, Z., Li, J., Liu, H., Xiong, F., Zhan, J., Hu, X., & Xie, W. (2021). PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*, 49(W1), W523-W529. <https://doi.org/10.1093/nar/gkab383>
- Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P., & Taipale, J. (2018). The interaction landscape between transcription factors and the nucleosome. *Nature*, 562(7725), 7725. <https://doi.org/10.1038/s41586-018-0549-5>
- Zhu, Y., Klasfeld, S., Jeong, C. W., Jin, R., Goto, K., Yamaguchi, N., & Wagner, D. (2020). TERMINAL FLOWER 1-FD complex target genes and competition with FLOWERING LOCUS T. *Nature Communications*, 11(1), 5118. <https://doi.org/10.1038/s41467-020-18782-1>

Résumé

Le contrôle de l'expression des gènes est essentiel pour les organismes vivants, et sa perturbation peut compromettre la survie. Les facteurs de transcription (TFs) régulent l'expression génique en se liant à des séquences d'ADN spécifiques appelées sites de liaison des facteurs de transcription (TFBS). LEAFY (LFY) est un TF spécifique aux plantes qui joue un rôle crucial dans le développement floral. Il est fortement conservé en termes de séquence et de spécificité de liaison tout au long de l'évolution des plantes.

Le rôle central de LFY dans la floraison a été étudié pendant des décennies, avec pourtant d'importantes zones d'ombres qui subsistent en ce qui concerne l'identification des cibles (gènes régulés) : pourquoi régule-t-il certaines régions génomiques *in vivo* et pas d'autres ? Pour élucider ce point, dans la première partie de ce manuscrit, je présente une approche permettant de prédire la régulation transcriptionnelle des TFBS de LFY dans la plante modèle *Arabidopsis thaliana*. J'ai utilisé des modèles classiques de liaison de LFY à l'ADN ainsi que le contexte génomique des sites LFY pour construire un modèle capable de distinguer les sites LFY fonctionnels (c'est-à-dire ceux liés par LFY et ayant effet sur l'expression génique *in vivo*) des sites non fonctionnels. Mes résultats suggèrent que la présence de TFBS de LFY environnants et, dans une moindre mesure, le niveau de diversité des TFBS d'autres TFs autour des sites LFY, sont importants pour distinguer les sites LFY fonctionnels des non fonctionnels. De plus, cette approche révèle plusieurs TFs qui co-occurrent avec LFY et qui contribuent à distinguer les sites régulés par LFY des sites LFY non fonctionnels. Malgré des preuves antérieures de l'importance fonctionnelle des régions conservées dans la régulation génique, l'inclusion de la conservation des sites LFY dans notre modèle n'a pas amélioré les prédictions, et je discute de raisons possibles derrière ce résultat. Dans l'ensemble, cette approche m'a permis de mieux caractériser la liaison de LFY à l'ADN, et elle peut être utilisée sur de nouvelles séquences génomiques pour prédire la régulation transcriptionnelle des sites par LFY, ainsi que par de nouveaux facteurs.

En plus de son action indépendante, LFY interagit avec UNUSUAL FLORAL ORGANS (UFO), une protéine F-box, pour garantir le développement correct des pétales et des étamines. Bien que l'interaction LFY-UFO et leur implication dans le développement floral soient déjà connues, le rôle exact d'UFO dans ce processus devait encore être déterminé. Dans la seconde partie de ce manuscrit, j'inclus un article récemment publié sur le rôle transcriptionnel du complexe LFY-UFO dans le développement floral, permettant à LFY de se lier à des régions génomiques distinctes de LFY seul. De plus, je présente quelques résultats supplémentaires suggérant l'implication de LFY et UFO dans l'établissement du méristème floral aux premiers stades du développement floral, élargissant ainsi leur importance dans ce processus développemental crucial.