Modélisation de la liaison à l'ADN et des mécanismes d'action de facteurs de transcription floraux

Arnaud Stigliani

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité Biologie Végétale

Arrêté ministériel : 25 mai 2016

Présentée par

Arnaud STIGLIANI

Thèse dirigée par **François PARCY, CNRS**

préparée au sein du **Laboratoire de Physiologie Cellulaire & Végétale**
dans **l'École Doctorale Chimie et Sciences du Vivant**

Modélisation de la liaison à l'ADN et des mécanismes d'action de facteurs de transcription floraux

Thèse soutenue publiquement le **2 octobre 2019,**
devant le jury composé de :

Dr. Sylvie COURSOL
Chargée de Recherche, INRA, Examinatrice
Dr. Thierry LAGRANGE
Directeur de Recherche, CNRS, Rapporteur
Dr. Adeline LECLERCQ-SAMSON
Professeur des Universités, Université Grenoble Alpes, Présidente
Dr. François PARCY
Directeur de Recherche, CNRS, Directeur de thèse
Dr. Magali RICHARD
Chargée de Recherche, CNRS, Examinatrice
Dr. François ROUDIER
Professeur des Universités, École Normale Supérieure de Lyon, Rapporteur

Remerciements

Si mon nom est le seul qui figure dans les auteurs de ce travail, son aboutissement est autant le produit de ceux qui m'ont entouré pendant ces trois années. Sans que je puisse leur rendre justice en quelques lignes, je vais faire de mon mieux pour leur témoigner ma gratitude.

Je souhaite d'abord remercier le laboratoire PCV pour m'avoir accueilli pendant cette thèse et le Labex GRAL pour l'avoir financée. Je remercie également François Roudier et Thierry Lagrange les rapporteurs de ce manuscrit, ainsi qu'Adeline Samson-Leclercq, Magali Richard et Sylvie Coursol pour avoir bien voulu prendre place dans mon jury.

François, merci mille fois pour m'avoir intégré à ton équipage et pour la confiance que tu m'as accordée. Ta passion et ton enthousiasme sont contagieux, les discussions avec toi palpitantes, j'en ressortais avec un désir ardent de percer les mystères. Ta manière de toujours voir le bon côté des résultats douteux et ton optimisme ont su me redonner confiance dans les moments d'incertitude et de découragement. En sciences, tu m'as appris à être exigeant avec moi-même, mes analyses, à être critique des résultats et à les discuter jusqu'à en faire fondre mon cerveau ! Enfin, merci d'avoir partagé avec moi ta passion de la montagne, de l'escalade et tes histoires qui m'ont fait rêver. Tu me l'avais dit, écrire une thèse, c'est un peu comme partir pour une longue course sauvage. On se trompe parfois de chemin, il faut faire demi-tour et on pense avoir fait le plus dur quand il faut encore gravir la chandelle. Tu étais là quand la corde s'est coincée, ma reconnaissance est immense. Ces trois années furent un voyage initiatique d'une abondante richesse. J'y ai appris autant en sciences, que sur moi-même ou sur la vie. Ce voyage, je n'aurais pas pu le faire sans guide.

Renaud, merci d'avoir répondu à mes questions d'informaticien inculte et d'avoir pris du temps pour moi quand je te l'ai demandé. Tes remarques et tes conseils, humbles et bienveillants, sont aussi plein de clairvoyance. Tes traits d'humour sont toujours drôles même si je suis parfois trop gêné pour en rire en public. J'ai passé de très bons moments avec toi, que ce soit au labo, ou pour jouer à la pétanque et boire des coups, ou quand tu essayais de m'arranger des coups tout court.

Anthony, je suis heureux que tu m'aies accueilli pendant les 3 mois que j'ai passés à Oslo, ma toute première expérience à l'étranger pour laquelle j'appréhendais. Tu m'as donné des méthodes de travail rigoureuses et tu m'as initié a des outils incontournables. Aujourd'hui, la plupart de mes analyses sont inspirées de ce que j'ai appris dans ton équipe.

Jérémy, nos discussions un peu emportées ne m'ont pas empêché de reconnaître la valeur de tes remarques et tes conseils. Quelle chance que tu fus là pour m'aider quand j'en ai eu le besoin et pour faire mon job pendant que je profitais un peu trop de mes vacances ;)

À mes amis partis vivre de nouvelles aventures ; Raquel, désolé pour ce cauchemar d'oligos :p. J'ai adoré travailler avec toi. Tu étais le soleil de mes journées pluvieuses et j'envie ceux que tu es partie éclairer de ton sourire et de ta joie. Adrien, merci d'avoir commenté tes codes, j'ai bien cru m'arracher

plusieurs fois les cheveux sur get_interdistance.py. Aujourd'hui, c'est un peu comme une Ferrari, quand il ne marche pas, je suis le seul à avoir les pièces, mais j'ai bon espoir qu'il s'impose comme n°1 dans le milieu des programmes codés au LPCV pour rechercher des interdistances. On s'est bien marré ensemble au labo et à jouer au foot à 50 sur des terrains pour 10. Younès (t'as vu, je l'ai bien écrit), t'avais un talent énorme pour m'exaspérer en 2 secondes. T'étais un peu ce mec qui te prépare à la vie en faisant tout ce qu'il peut pour te faire craquer. Malgré tout, j'ai toujours su apprécier ta franchise, dans tes remarques en sciences ou sur la vie et j'ai aimé les moments qu'on a passés ensemble. Dommage qu'on n'ait pas retrouvé ton pull...

Tiffany, Sophie et Alexandre, merci d'avoir partagé l'air frais de votre sanctuaire, j'aime bien le sauna, mais pas quand j'écris ma thèse.

Un grand merci à Franz Bruckert, sans qui je n'aurais jamais découvert la bioinformatique.

Xuelei, merci pour toute la matière que tu m'as donnée a analyser. Pas facile d'aller à ton rythme, mais au moins, je n'avais pas le temps de m'ennuyer. Merci à Chloe, sans qui ces projets n'auraient pas été réalisables. Véro, c'était toujours agréable de parler de plantes avec toi et d'écouter tes histoires à mourir de rire quand on mangeait ensemble.

Bien sûr, merci à toutes les autres personnes qui ont été proches de moi dans ce labo. Chrispi, Gabi, Manu, Louise dite "la cool", Claire, Claudius, Paulette, Vous m'avez fait me sentir chez moi durant ces trois années, on a passé des bons moments au labo, mais surtout à l'extérieur des murailles du centre. Hicham, j'ai enfin pu mettre une tête sur la personne qui a écrit ce manuscrit magnifique, source d'inspiration intarissable. Mention spéciale pour Michel, qui me rappelait (un peu trop) souvent que la montagne, c'est dangereux. Antonin, Coralie, Élodie, Loïc et Pauline B. (par ordre alphabétique), vous avez créé une ambiance "grave stylée de quelque part [1]" durant ma dernière année de thèse, un peu trop cool, même, à ce qu'il paraît ;). Coralie, quel courage d'avoir relu ce manuscrit, grâce à toi, je sais qu'il n'y a aucun rapport entre les bords du méristème et une fameuse recette aux œufs. Si vous avez besoin d'aide, adressez-vous à notre maître à tous, Philippe à qui je conseille de ne pas tenter des trucs trop aventureux en hiver ;)

Enfin, cette thèse n'aurait jamais pu voir le jour sans mes parents. Vous avez su vous effacer devant mes choix, soutenu dans tout ce que j'ai fait, vous ne m'avez jamais rien imposé ou pris trop de place dans ma vie et vous avez toujours fait ce qui semblait le mieux pour moi. Pour moi et pas pour vous. Moi qui vous montre assez peu de ce que je ressens, je vous remercie du fond de mon cœur pour la vie que vous m'avez donnée.

1. ce ne sont pas les mots exacts

Table des matières

Introduction

I.1 Généralités sur la fleur

I.1.1 Rapide histoire des plantes

Alors que les premiers fossiles de bactéries sont datés à 3.5 milliards d'années (Ga) avant notre ère, il faut attendre 1.2 Ga pour que les premières algues apparaissent dans les océans. Elles sont le produit de la fusion entre une cyanobactérie et d'un petit organisme eucaryote. Cette cyanobactérie sera à l'origine du chloroplaste, l'organelle qui permet aux plantes de produire du dioxygène à partir de dioxyde de carbone, d'eau et de lumière. Une branche des algues d'eau de mer (les charophytes) donne naissance aux algues d'eau douce il y a 850 millions d'années (Ma). Comme ces algues étaient soumises à de ponctuelles émersions, elles se sont peu à peu adaptées aux forts rayonnements, à la plus faible disponibilité en eau, à la poussée d'Archimède réduite et aux variations de température importantes pour coloniser la terre ferme il y a 450 millions d'années (figure 1). Ces plantes terrestres sont les bryophytes, essentiellement des mousses, qui ne possèdent pas de système vasculaire, ce qui leur impose une taille limitée car la totalité de la plante doit rester à l'humidité. Une famille diverge alors des bryophytes pour se protéger de la dessiccation en développant une cuticule externe étanche et un système vasculaire. Maintenant capables de transporter l'eau, les plantes peuvent croître et les premières fougères commencent à occuper la terre (420 Ma). Si les fougères ont développé de nombreux mécanismes pour se développer dans des milieux plus secs, la reproduction nécessite toujours un milieu aqueux pour que les gamètes mâles puissent se déplacer jusqu'à un gamète femelle ; aussi ces plantes ne pouvaient occuper que les terres proches des côtes et des cours d'eau. Elles s'affranchissent de cette dernière contrainte en inventant les graines (350 Ma). Ainsi, chez les gymnospermes, dont les plus connus sont les conifères (280 Ma), les gamètes mâles se développent dans des cônes et à leur maturité, ils sont transportés par le vent jusqu'aux cônes femelles d'un autre individu. Les gamètes femelles sont alors fécondés et les graines peuvent se développer. À maturité, elles sont relâchées et peuvent commencer à germer. Comme la graine protège l'embryon de la dessiccation grâce à son enveloppe, il peut survivre plusieurs années pour attendre des conditions idéales à sa germination, ce qui donne aux plantes la possibilité de coloniser des milieux hostiles, comme les montagnes. Le réservoir de nutriments contenus dans la graine assure une germination beaucoup plus rapide et un taux de réussite élevé, permettant aux gymnospermes de supplanter les fougères.

On peut dater la dernière révolution évolutive des plantes autour de 175 Ma. Cette période voit naître les premières plantes à fleurs, regroupées dans la famille des angiospermes. Nés d'une d'une origine commune, les angiospermes colonisent pourtant la planète si rapidement qu'on ne sait pas où leurs premiers représentants sont apparus. Le succès de la fleur tient à l'efficacité de la fécondation par les insectes (et quelques oiseaux ou mammifères), qui attirés par la fleur, contribuent à transporter le pollen d'une plante à une autre. Après fécondation, le fruit se forme et celui-ci peut alors être disséminé soit par le vent (c'est le cas de l'érable et du platane) ou par d'autres animaux permettant à l'espèce de de répandre. Cette

méthode de reproduction profite également aux insectes, le nectar et le pollen étant des aliments riches. Le nombre et la variété d'insectes comme de fleurs croissent de concert et aujourd'hui, 90% des espèces végétales sont des angiospermes.

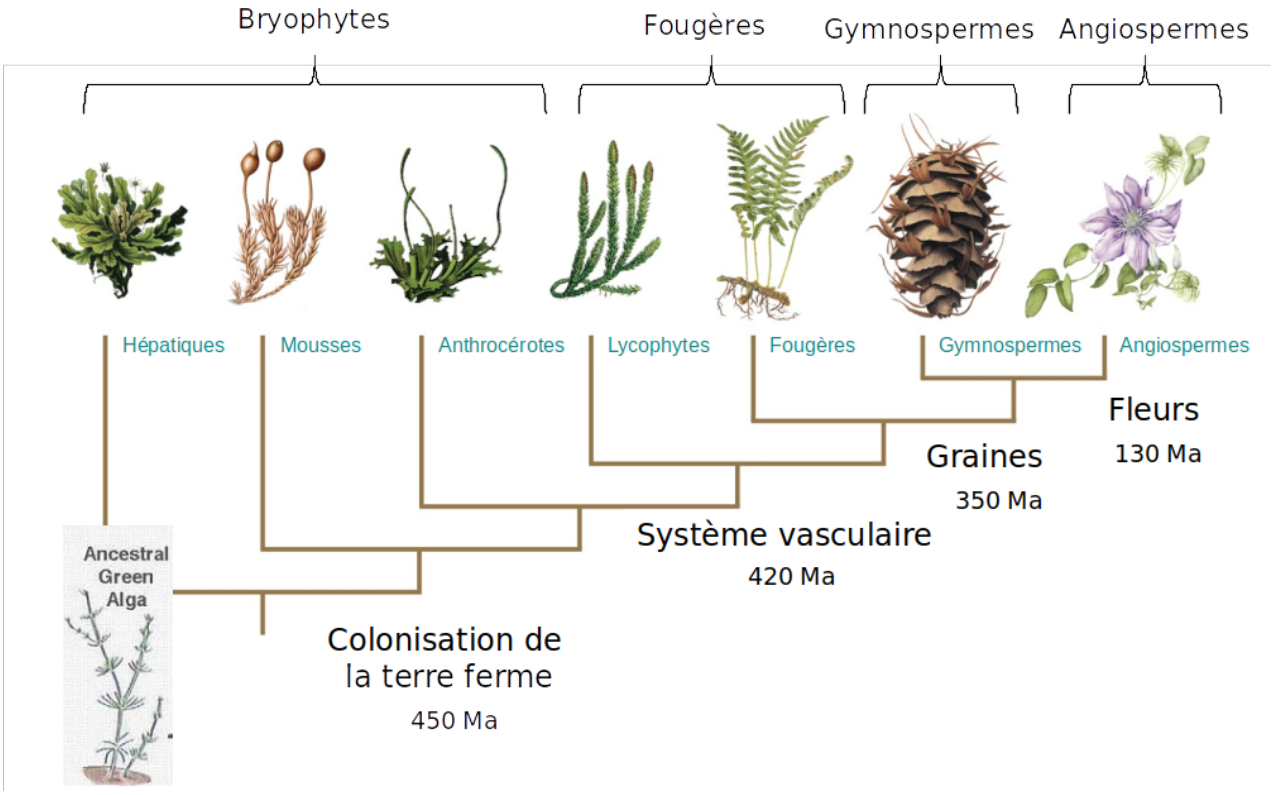


FIGURE 1: Phylogenèse des plantes terrestres

I.1.2 Anatomie de la fleur

La fleur constitue l'organe reproducteur de la plante. Au sein des espèces d'angiospermes, la structure de l'organe floral reste sensiblement la même et peut être subdivisée en 4 couronnes (figure 2).

Les sépales ressemblent à des feuilles. Ils enferment le bourgeon floral au début du développement la fleur et protègent les autres organes jusqu'à ce qu'ils arrivent à maturité.

Les pétales confèrent une grande partie de son attractivité à la fleur. Grâce à leur couleur, leur odeur et leur forme, ils attirent les insectes pour qu'ils viennent les polliniser. Ainsi, des pétales d'orchidées imitent l'insecte femelle et sécrètent des phéromones spécifiques. Croyant féconder une femelle, le mâle féconde en fait une fleur. D'une toute autre manière, les pétales de *Datura metel* sont chargées en substances qui rendent l'insecte dépendant et l'encourage à revenir.

Les étamines portent le pollen qui constitue le gamète mâle.

Enfin, les carpelles constituent le pistil, l'organe reproducteur femelle. Au sommet des carpelles, les stigmates peuvent recueillir le pollen. Le pollen produit alors un tube pollinique qui va amener les gamètes mâles le long du style (partie longue du carpelle) jusqu'aux ovules. Il est intéressant de noter que

la plupart des angiospermes sont incapables de s'autopolliniser. En effet, certains mécanismes empêchent ce processus pour favoriser le brassage génétique. Par exemple, le tube pollinique sera bloqué dans le style, incapable d'atteindre l'ovule. Dans d'autres cas, les étamines et les carpelles n'arrivent pas à maturité en même temps.

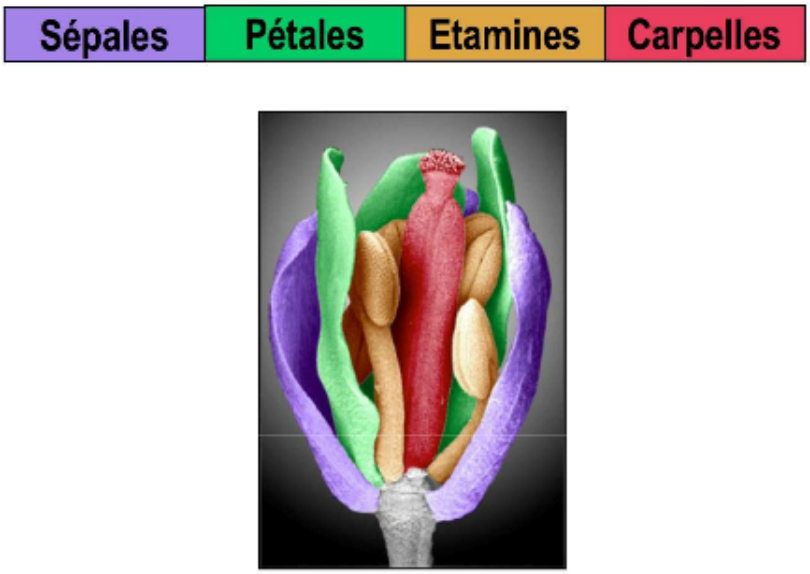


FIGURE 2: Anatomie d'une fleur

I.1.3 La fleur à l'échelle de la plante

À l'opposé des espèces animales, les végétaux entretiennent des réservoirs de cellules souches (ou indifférenciées) tout au long de leur vie appelés méristèmes. Après la germination, l'embryon de plante met en place deux méristèmes : un méristème racinaire qui permettra à toutes les racines de se former dans le sol et un méristème apical caulinaire depuis lequel se produira la formation de tous les organes émergés de la plante. La plante croît donc en taille grâce à ces deux méristèmes. La formation de méristèmes axillaires à l'aisselle des feuilles permet la formations de branches. Les méristèmes secondaires à l'extrémité de ces branches peuvent alors à leur tour induire de nouveaux méristèmes axillaires.

Lors de la transition florale, les méristèmes végétatifs se différencient en méristème d'inflorescence sur lesquels les méristèmes floraux, qui formeront les futures fleurs, vont apparaître.

Après avoir présenté les fleurs, la suite immédiate de cette introduction va d'abord introduire les éléments généraux nécessaires à la compréhension des mécanismes moléculaires pour revenir ensuite sur les acteurs spécifiques à la fleur.

I.2 Les protéines façonnent la plante

Jusqu'alors, nous avons décrit l'architecture de la fleur et les étapes développementales de la plante qui conduisent finalement à la floraison. Pour comprendre l'orchestration de ces phénomènes macroscopiques – ce qui est l'objet de cette thèse – il faut se pencher sur les mécanismes biologiques qui ont lieu

dans la cellule, en particulier au niveau de l'ADN. En effet, l'ADN est le support des gènes et ceux-ci codent pour des protéines, qui façonnent la cellule.

Au sein d'un même organisme, l'unicité de l'ADN dans les cellules somatiques impose une régulation complexe des gènes exprimés et réprimés : la plupart des gènes régulateurs de l'architecture florale ne doivent pas être exprimés dans les racines, par exemple.

I.2.1 Généralités sur la régulation de la transcription

I.2.1.1 Les facteurs de transcription

La régulation de l'expression des gènes est orchestrée par les facteurs de transcription (TF). Ces protéines se lient sur des sites de liaison spécifiques sur le génome (Wasserman and Sandelin, 2004) à proximité des gènes pour recruter ou empêcher le recrutement du complexe d'initiation de la transcription.

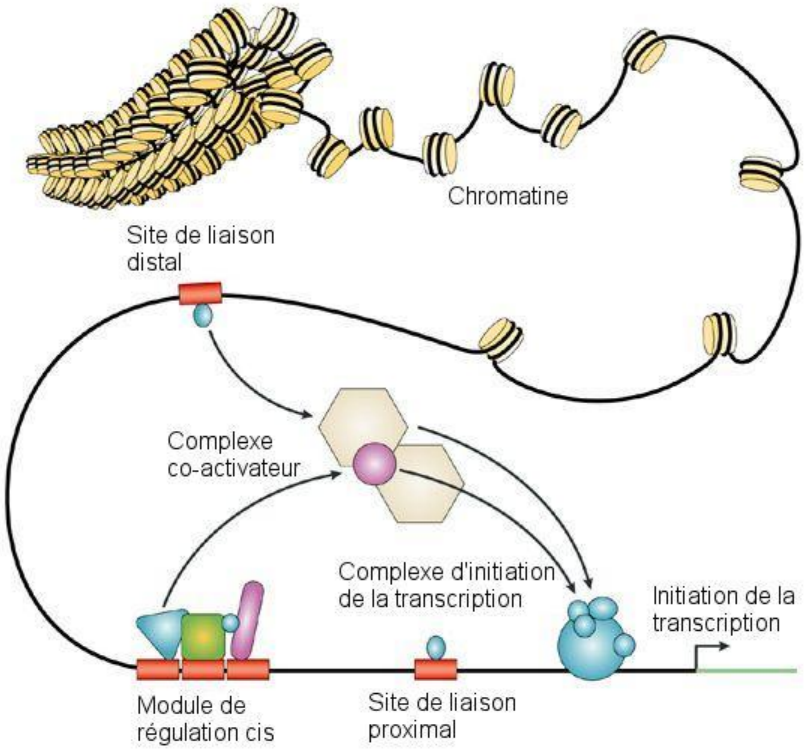


FIGURE 3: Les TF se lient à des sites spécifiques qui peuvent être à proximité (site de liaison proximal) ou à distance (site de liaison distant) du site d'initiation de la transcription. Des ensembles de TF peuvent agir sur des modules de régulation cis pour réguler la transcription. Les interactions entre les TF liés et des co-facteurs stabilisent le complexe d'initiation de la transcription, permettant l'expression du gène. Les propriétés de régulation conférées par les facteurs de transcription liés sont très dépendantes de la structure tridimensionnelle de la chromatine. D'après Wasserman and Sandelin (2004).

À l'échelle d'un TF, le noyau de la cellule et le génome sont immenses. Pour faciliter le déplacement

du TF vers ses sites spécifiques, celui-ci diffuse d'abord vers l'ADN. Une fois lié à l'ADN, des interactions aspécifiques permettent vraissemblablement au TF de glisser le long de la double hélice (Raccaud et al., 2019; Marklund et al., 2013). Le contexte génomique aux alentours des sites spécifiques (le contenu en nucléotides GC, la chromatine...) sont donc des éléments pouvant aider ou empêcher le TF d'atteindre ses sites de liaison spécifiques.

Le contact du TF à l'ADN est assuré par une séquence d'acide aminés appelé domaine de liaison à l'ADN. Comme ils assurent la spécificité de liaison du TF, ces acides aminés sont généralement très conservés au cours de l'évolution au sein d'une famille de TF partageant une même spécificité. Pour comprendre les interactions spécifiques entre le TF et l'ADN, il nous faut détailler la structure de la double hélice. Celle-ci forme deux sillons : le petit sillon (*Minor groove*) (figure 14) et le grand sillon (*Major groove*). La majorité des TF se lie dans le grand sillon : celui-ci affiche un plus grand nombre de configurations de donneurs et d'accepteurs de liaisons hydrogène que le petit sillon et permet donc une meilleure spécificité de liaison (figure 4).

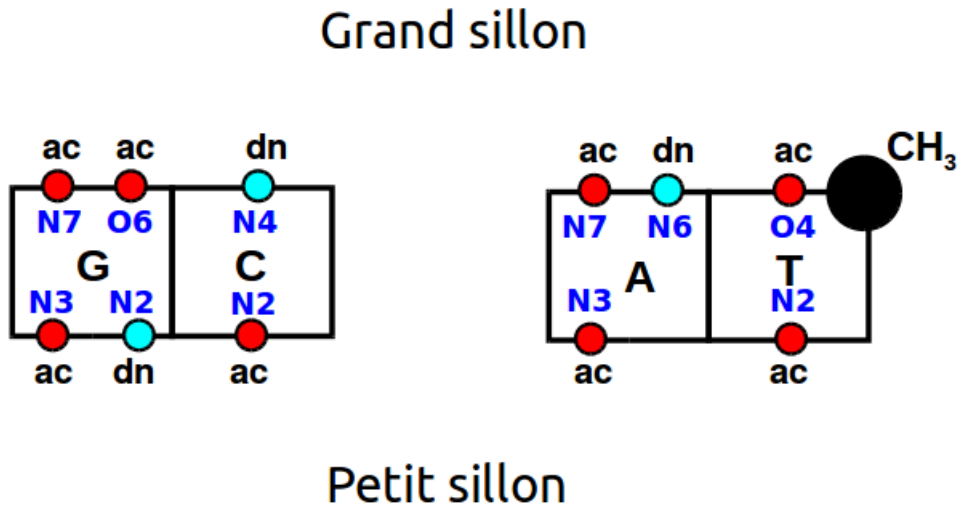


FIGURE 4: Donneurs (en bleu) et accepteurs (en rouge) dans le grand sillon et le petit sillon de la double hélice d'ADN pour les paires de bases (G : C) et (A : T). Dans le petit sillon, les configurations de donneurs et d'accepteurs ne permettent pas de différencier (A : T) de (T : A) et (G : C) de (C : G). Dans le grand sillon, on peut discerner (G : C) de (C : G) et le groupement méthyle permet de différencier (A : T) de (T : A).

Parmi les structures protéiques, les hélices α et les feuillets β sont celles qui embrassent le mieux la double hélice d'ADN en s'insérant dans le grand et le petit sillon (Yamasaki et al., 2008; Liu et al., 1999). Les *basic leucine zipper* (figure 5.a) illustrent la liaison à l'ADN dans le grand sillon par le biais d'hélices α. Les domaines de liaison sont néanmoins souvent plus complexes en combinant hélices α et feuillets β ; dans la famille des protéines à doigt de zinc, l'ion Zn^{2+} joue le rôle de ligand pour assurer la conformation des 2 feuillets β et de l'hélice α qui constituent le domaine de liaison (figure 5.b). Plusieurs TF agissent souvent ensemble en formant des complexes fonctionnels (figures 3 et 6).

Des ensembles définis de TF seront exprimés dans des types cellulaires donnés, assurant ainsi l'activation et la répression des gènes nécessaires à ce type cellulaire. Dans la plante *Arabidopsis thaliana* (modèle d'étude pour la biologie végétale), on compte plus de 2 000 facteurs de transcription assurant la

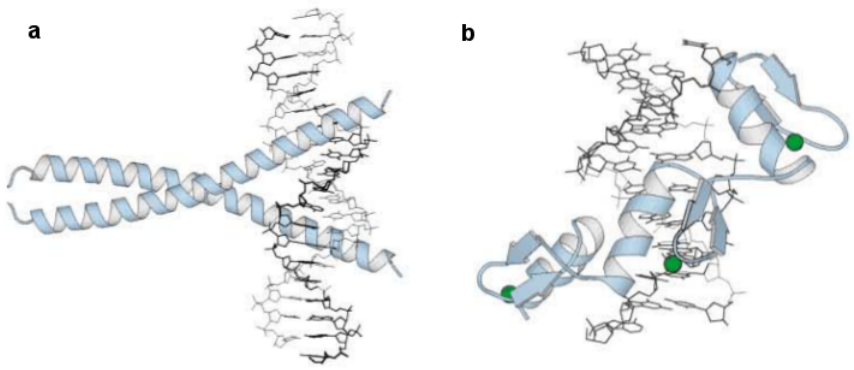


FIGURE 5: (A) Le domaine de liaison est constitué de deux *basic leucine zipper*. (B) Les trois doigts de zinc assurent le contact avec l'ADN. D'après Yamasaki et al. (2008).

régulation des 25 000 gènes contenus dans le génome (Mitsuda and Ohme-Takagi, 2009).

I.2.1.2 Les facteurs épigénétiques

L'ADN s'enroule autour des histones pour former des nucléosomes qui constituent la chromatine (figure 3) Les nucléosomes sont constitués des histones de cœur H2A, H2B, H3 et H4. Les différents variant d'histones (H3.1, H3.2, H3.3, H2A.Z, H2A.X...), leur présence ou leur absence et les modifications post-traductionnelles présentes sur les celles-ci engendrent différents degrés de compacité. Par conséquent, si un promoteur suffisamment fermé contient un site de liaison pour un TF donné, ce TF ne pourra pas se lier (Klemm et al., 2019). De ce fait, il ne pourra pas exercer son action régulatrice. L'état des histones, leur présence ou leur absence agissent donc comme un deuxième niveau de régulation. Ainsi, des groupes de TF et des profils chromatiniens contribuent à définir des types cellulaires particuliers (Pikaard and Scheid, 2014).

L'état de la chromatine n'est pas figé, il peut être altéré par l'ajout, le retrait de nucléosomes, ou certaines modifications des histones qui le composent (Xiao et al., 2017) : certaines protéines peuvent déposer des marques, les lire ou les retirer (ces notions sont traitées plus en détail dans le paragraphe I.2.2). Ces protéines doivent être adressées à des régions données du génome alors qu'elles ne sont souvent pas spécifiques de séquences d'ADN définies. Certains facteurs de transcription sont alors capables d'interagir avec les remodeleurs chromatiniens, leurs permettant de cibler des régions précises (Li et al., 2016a).

De même que les histones, l'ADN est sujet à des modifications : les cytosines peuvent être respectivement méthylées ou déméthylées par des protéines appelées ADN-méthyltransférases et ADN-déméthylases (He et al., 2011). Comme les remodeleurs chromatiniens, ces protéines ont besoin de partenaires pour cibler des régions spécifiques (Zhu et al., 2016). La méthylation des cytosines est généralement considérée comme une marque empêchant la liaison des TF et favorisant la liaison des *methyl binding proteins* (MBP) se liant sur l'ADN méthylé. Ces protéines agissent comme des compétiteurs des TF et sont connues pour recruter des remodeleurs qui compactent la chromatine (Zhu et al., 2016). Cependant, de nouvelles données montrent que certains TF ont une affinité plus importante pour les régions où l'ADN est méthylé, suggérant que certaines régions méthylées peuvent être activement transcrites (Zuo et al., 2017).

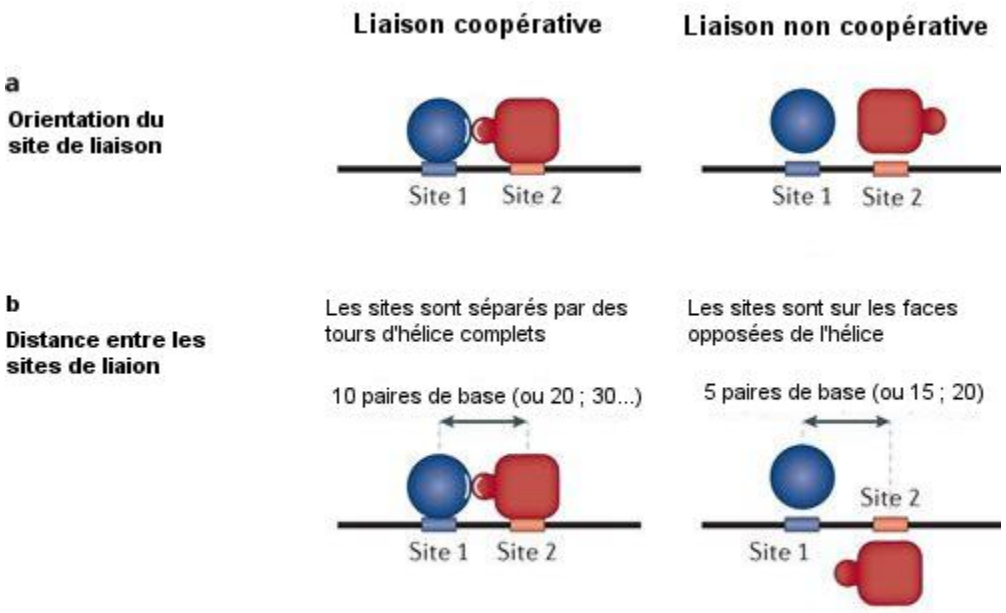


FIGURE 6: (A) L'orientation des sites de liaison est importante pour assurer la coopérativité des TF. (B) L'espacement entre les TF est également important. S'ils se situent de part et d'autre de la double hélice d'ADN, ils ne pourront pas interagir. D'après Spitz and Furlong (2012).

I.2.1.3 Ouvrir la chromatine fermée

Il convient d'ajouter que la différenciation en un type cellulaire donné fait souvent intervenir des modifications du paysage chromatinien très importantes (Pikaard and Scheid, 2014). Par exemple, certains gènes doivent demeurer complètement inaccessibles aux TF et aux remodeleurs chromatiniens avant la différentiation du tissu méristématique en tissu floral. Après la transition florale, il faut à l'inverse qu'ils puissent être régulés avec beaucoup plus de finesse. Ouvrir la chromatine au niveau des régions du génome fermées aux TF communs est un rôle qui appartient à un type de TF très particuliers appelé *facteurs pionniers*. Si de nombreux facteurs pionniers ont été identifiés dans les génomes du règne animal, la propriété pionnière n'a pour le moment été démontrée que pour le TF LEAFY COTYLEDON 1 chez *A. thaliana* (Lai et al., 2018).

Bilan

Même si un grand nombre de paramètres entre en jeu dans la régulation des gènes, le paragraphe précédent montre que les TF sont les véritables maîtres de ce processus. En effet, eux seuls "savent" exactement où se fixer et qui recruter pour que la machinerie transcriptionnelle agisse de façon adéquate.

Comprendre et être capable de prédire où les facteurs de transcription se lient n'est assurément pas équivalent à comprendre la régulation. Cela n'en reste pas moins une étape absolument nécessaire et un fossé tout aussi formidable à franchir quand on veut décrypter le grand mystère que constitue la régulation des gènes dans chaque type cellulaire au cours d'un processus développemental.

I.2.2 Les modifications post traductionnelles des histones

Ici, nous détaillerons les phénomènes chromatiniens évoqués dans le paragraphe I.2.1.2.

I.2.2.1 Hétérochromatine et euchromatine

On distingue deux types de chromatine : l'hétérochromatine et l'euchromatine. La première forme comprend les régions de l'ADN qui ne sont pas destinées à être transcrites et qui sont donc très compactes. Dans ces régions, on compte d'une part les centromères et les télomères, qui ne contiennent pas de gène, et d'autre part les transposons (Sequeira-Mendes et al., 2014). À l'inverse, l'euchromatine est beaucoup plus riche en séquences codantes et moins compacte que l'hétérochromatine. La compacité de l'hétérochromatine est notamment assurée par la di-méthylation de la lysine 9 de l'histone 3 (H3K9me2). On constate que l'euchromatine affiche une plus grande variété de profils lui permettant de moduler l'expression des gènes : par exemple, H3K4me3, H3K27ac ou H3K36me3 peuvent marquer les gènes actifs alors que H3K27me3 est trouvé chez les gènes réprimés (Sequeira-Mendes et al., 2014; Pikaard and Scheid, 2014).

I.2.2.2 Les acteurs des modifications

Les remodeleurs chromatiniens peuvent être subdivisés en trois catégories : les protéines qui reconnaissent les marques (*readers*), celles qui positionnent ces marques (*writers*) ou celles qui les enlèvent (*erasers*). Le complexe PRC2 (polycomb repressive complex) qui dépose la marque H3K27me3 est un *writer* alors que les protéines qui contiennent un domaine "Jumonji" peuvent enlever cette marque et sont des *erasers*. Les *readers* joue également un rôle important en reconnaissant certaines marques et en recrutant des *writers* ou des *erasers* (Liu et al., 2010).

On peut illustrer leurs rôles à travers le phénomène de vernalisation, qui se traduit par une induction de la floraison d'une plante préalablement exposée au froid. Si la plante n'a pas subi de froid prolongé, le gène *FLOWERING LOCUS C* (*FLC*) est actif et réprime la floraison. Lors d'une exposition au froid suffisamment longue, le complexe PRC2 ajoute les marques répressives H3K27me3 sur le locus du gène *FLC* (Bastow et al., 2004; De Lucia et al., 2008). La protéine TERMINAL FLOWER 2 se lie à ces marques et peut induire le dépot des marques H3K9me2 sur le gène (Gaudin et al., 2001; Mylne et al., 2006; Turck et al., 2007; Zhang et al., 2007). La chromatine est alors fermée de manière irréversible et le gène *FLC* est définitivement inactif. Notons que les *readers* sont également sensibles à la méthylation des cytosines. Par exemple, la protéine KRYPTONITE est une *methyl binding protein* qui recrute des H3K9 méthyltransferases pour déposer la marque H3K9me2 (Jackson et al., 2002, 2004).

I.2.2.3 Définir des états chromatiniens

Les différents variants d'histones et les nombreuses modifications qu'elles peuvent subir donnent un très grand nombre de combinaisons, ce qui rend leur interprétation difficile.

En prenant en compte plusieurs marques post-traductionnelles, plusieurs variants d'histones et la méthylation de l'ADN, des travaux ont mis en évidence des profils précis de combinaisons appelés états chromatiniens. Le nombre plus réduit d'états chromatiniens peut alors faciliter la compréhension de certains phénomènes (Roudier et al., 2011; Sequeira-Mendes et al., 2014).

I.3 Comprendre où les TF se lient pour comprendre la régulation

I.3.1 Les apports de la génomique

Le début du XXI^{e} siècle a coïncidé avec des avancées prodigieuses dans le domaine de la génomique. Si séquencer les premiers génomes était un véritable exploit, c'est devenu aujourd'hui une opération courante. De nombreuses méthodes ont ainsi vu le jour : elles donnent accès au transcriptome, à la compacité de l'ADN et aux régions du génome liées par une protéine donnée. Elles sont donc des outils précieux pour comprendre les réseaux de régulation entre les gènes et les TF.

Ici, nous allons détailler les principales méthodes utilisées. Dans ce cadre, vous trouverez des informations supplémentaire dans une revue que j'ai co-écrite, placée en annexe V.1.

I.3.1.1 Étudier l'expression des gènes grâce au Séquençage de l'ARN (RNA-Seq)

Le RNA-Seq donne accès aux transcrits présents dans un échantillon donnant des indications sur les gènes exprimés. Des techniques avancées permettent même le RNA-Seq sur cellules uniques ! La méthode est décrite dans la figure 7. Précisons qu'on ne séquence en général pas les fragments en entier mais leurs extrémités : ceux-ci sont trop longs pour les technologies couramment utilisées. Connaissant la taille approximative des fragments, on est capable de les replacer précisément sur le génome pour en déduire le fragment.

Remarquons que le RNA-Seq donne accès à la quantité de transcrits et non de protéine traduite à partir de l'ARN.

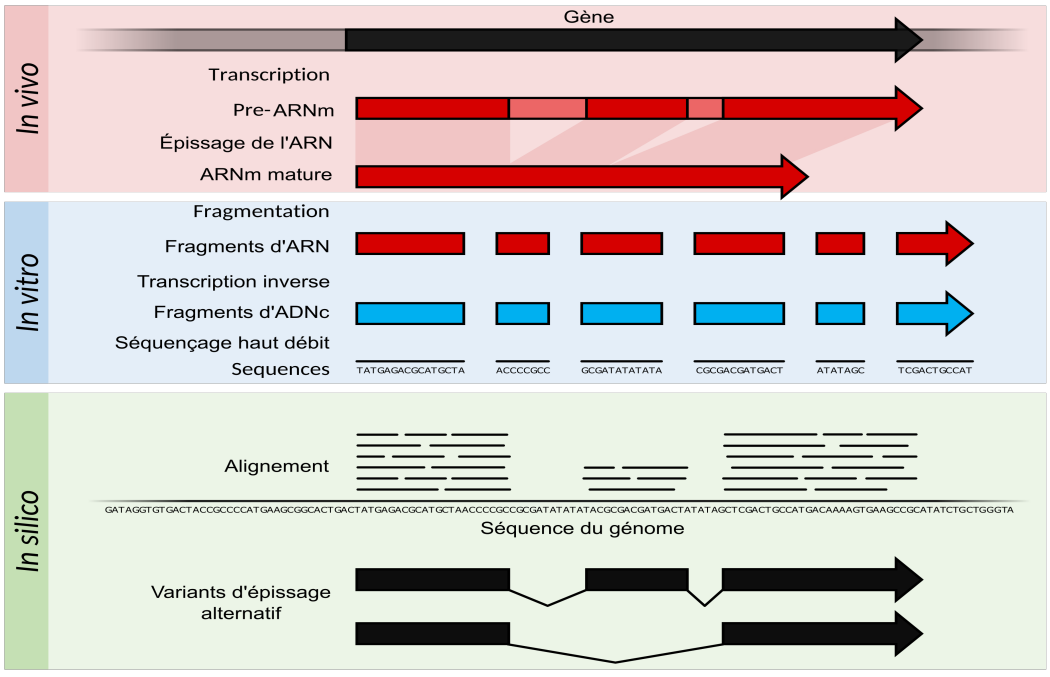


FIGURE 7: Principe du RNA-Seq d'après *Wikipedia*

I.3.1.2 Étudier les régions du génome liées par un TF *in vivo* grâce à l'immunoprécipitation de chromatine suivie du séquençage (ChIP-Seq) (1-2p)

À l'aide d'un anticorps qui cible une protéine, on peut connaître les régions liées par cette protéine dans un tissu donné *in vivo*. Le ChIP-Seq (figure 8) peut aussi bien cibler des TF, des remodeleurs chromatiniens ou bien certaines modifications post traductionnelles des histones.

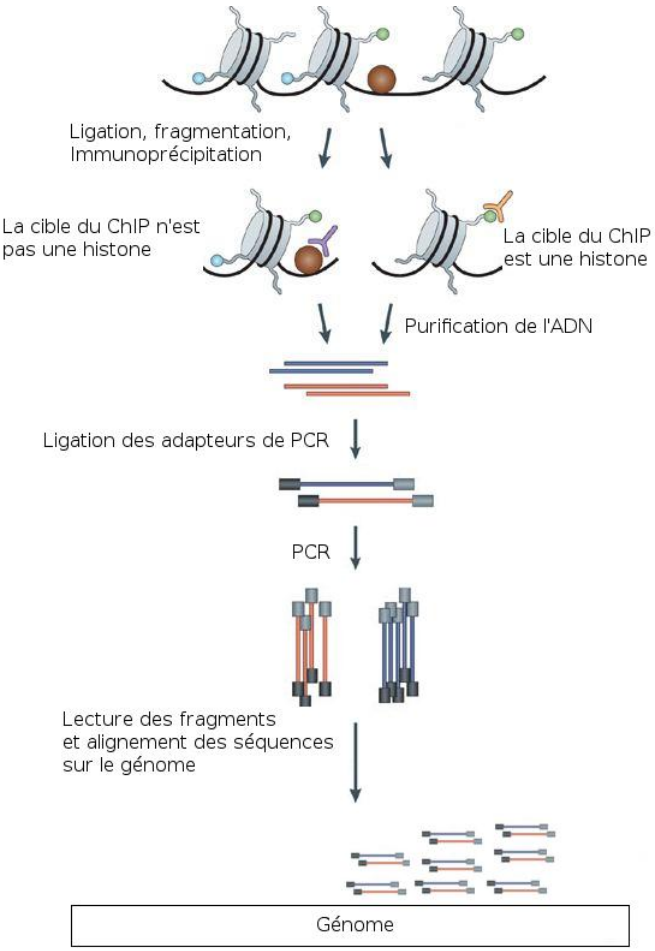


FIGURE 8: Principe du ChIP-Seq d'après Kidder et al. (2011)

Si le ChIP-Seq donne des informations précieuses, il possède plusieurs limites. Certaines sont techniques : il faut d'abord avoir réussi à isoler suffisamment de tissu, ce qui peut être délicat lorsque le nombre de cellules disponibles est faible (dans les premiers stades des méristèmes floraux par exemple). Ensuite, les résultats de ChIP-Seq sont parfois difficiles à interpréter et à modéliser. Par exemple, on ne peut pas toujours savoir si un TF d'intérêt lie directement l'ADN ou s'il a formé un complexe avec une protéine qui est, elle, liée directement à l'ADN. Le contexte chromatinien, qui reste souvent inconnu, rend la tâche encore plus ardue. Le nombre de paramètres qui peuvent influencer la liaison est donc souvent trop important pour qu'elle soit modélisée.

I.3.1.3 Étudier les régions de l'ADN liés par un TF *in vitro* grâce au *DNA Affinity Purification Sequencing* (DAP-Seq)

Pour étudier les propriétés d'un TF dans un contexte plus simple que la cellule, il est possible de déterminer ses préférences *in vitro* : il existe pour cela plusieurs méthodes biochimiques comme le *Protein Binding Microarray* (PBM) ou l'enrichissement exponentiel de ligand (SELEX) (Cf revue en annexe V.1). Ces méthodes présentent en plus d'autres avantages, comme leur facilité de mise en œuvre (l'ADN synthétique et les protéines sont recombinantes). Le fait que l'ADN soit synthétique présente néanmoins un inconvénient si on veut connaître avec certitude le comportement du TF en présence d'ADN génomique.

Une nouvelle technique permet aujourd'hui de franchir cette barrière.

Le DAP-Seq a été inventé par O'Malley et al. (2016) (figure 9). Il peut être vu comme une amélioration du DIP-ChIP (Liu et al., 2005). La méthode consiste à isoler l'ADN d'un organisme, le débarrasser de ses nucléosomes et de toutes les protéines qui y sont liées et le fragmenter. On exprime ensuite une protéine avec un étiquettes et cette protéine est mise en présence de l'ADN. Les fragments d'ADN liés sont récupérés grâce à des billes qui ont une affinité particulière pour l'étiquette et comme pour le ChIP-Seq, ces fragments sont amplifiés par PCR, puis alignés sur le génome. Remarquons que les cytosines sont toujours potentiellement méthylées au moment où le facteur de transcription se lie sur les fragments. La méthylation peut avoir une influence sur la liaison du TF à l'ADN, il est donc possible de réaliser une PCR sur les fragments avant de les mettre en présence avec le TF, la PCR ne conservant pas la méthylation (Amp-DAP).

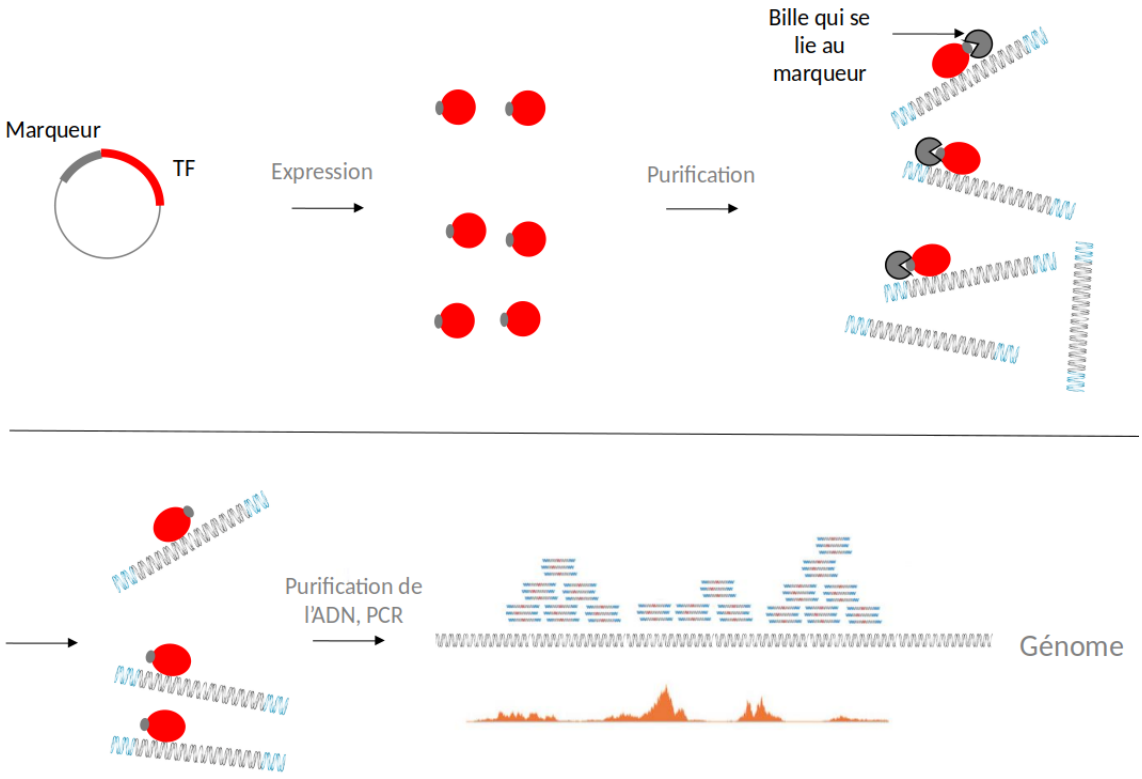


FIGURE 9: Principe du DAP-Seq. Les extrémités bleues des fragments représentent les adaptateurs utilisés pour préparer les librairies.

I.3.1.4 Positionner les nucléosomes et quantifier l'ouverture de la chromatine

Dans la section I.2, nous avons évoqué l'importance du contexte chromatinien sur la régulation. Savoir où se positionnent les nucléosomes *in vivo* et connaître l'état de compacité de la chromatine en tout point du génome s'avère donc parfois indispensable. Les méthodes suivantes, résumées dans la figure 10, permettent d'avoir ces informations.

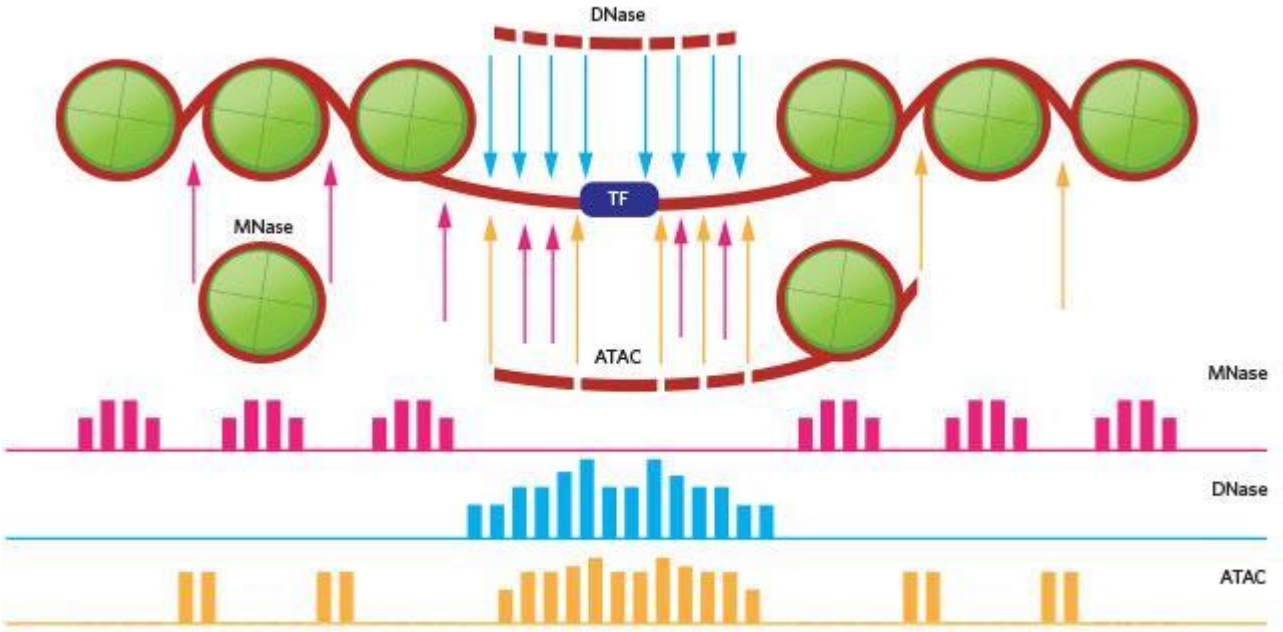


FIGURE 10: Le MNase-Seq donne accès à la position des nucléosomes. Le DNase-Seq et l'ATAC-Seq donnent la position des régions ouvertes de la chromatine. D'après Tsompana and Buck (2014).

MNase-Seq Le *Micrococcal Nuclease Sequencing* (MNase-Seq) est une méthode qui donne accès à la position des nucléosomes. L'ADN est mis en présence de nucléase micrococcale, une enzyme qui digère celui-ci s'il n'est pas enroulé autour des nucléosomes. Il suffit ensuite de séquencer les fragments qui n'ont pas été digérés pour avoir ces positions.

DNase-Seq Le *DNase I hypersensitive sites sequencing* (DNase-Seq) utilise le même principe que le MNase-Seq. On utilise une autre enzyme, la DnaseI, qui coupe l'ADN lorsqu'il est accessible. Les extrémités des fragments coupés sont ensuite séquencés, un fort signal DNase-Seq indique qu'une région est accessible.

ATAC-Seq L'*Assay for Transposase-Accessible Chromatin with highthroughput Sequencing* (ATAC-Seq) donne sensiblement les mêmes résultats que le DNase-Seq. Elle présente cependant l'avantage d'une plus grande facilité d'exécution et un principe de mise en œuvre différent (Buenrostro et al., 2013, 2015). On utilise la transposase Tn5, une enzyme qui se pose sur l'ADN et induit l'insertion de transposons dans les régions les plus accessibles. Les transposons utilisés possèdent des adaptateurs pour séquençage, si bien que

les régions cibles de la Tn5 peuvent être amplifiées par PCR. Les fragments amplifiés sont ensuite recueillis et séquencés. La Tn5 ayant accès aux régions ouvertes, les fragments séquencés donnent leur position sur le génome. Comme la transposase Tn5 utilisée est une version très active, elle donne la position de régions plus compactées que la DNaseI.

Remarquons que la DNaseI et la Tn5 ne peuvent agir à une position de l'ADN protégée par un TF. On peut donc déduire la position des TF grâce au DNase-Seq et à l'ATAC-Seq (figure 10) si la couverture en *reads* est suffisamment forte.

I.3.2 Prédire la liaison d'un TF à l'ADN à partir de données génomiques

Nous venons de détailler les outils qui permettent d'observer ce qui se passe à l'échelle du génome. Comprendre ces phénomènes, c'est être capable de les expliquer, de les modéliser. Sans oublier notre quête, qui est d'expliquer les phénomènes macroscopiques qui conduisent à la formation de la fleur par les mécanismes moléculaires qui se produisent à l'échelle du génome, rappelons (Bilan de la section I.2) que les protéines qui induisent ces phénomènes sont les TF ; comprendre où ils se lient est donc capital.

Dans cette section nous détaillerons les principales méthodes utilisées pour modéliser leurs préférences, un inventaire plus général et descriptif étant présent dans la revue en annexe V.1.

I.3.2.1 Un modèle simple : La PWM

La Matrice Poids Position (PWM), introduite la première fois par Stormo et al. (1982), est probablement l'outil le plus indiqué compte tenu de son rapport simplicité/efficacité. L'idée est de donner des scores à des séquences d'ADN possédantla taille de la région reconnue par le TF. Ici, nous allons détailler comment fabriquer une PWM, comment l'utiliser et ses limites.

Fabriquer une PWM La méthode pour obtenir une PWM est détaillée dans la figure 11 et se base sur l'article de Wasserman and Sandelin (2004). Les formules pour passer d'une étape à la suivante sont exposées dans le paragraphe suivant. La méthode pour aligner les séquences (Figure 11.a) dépend de la technique utilisée pur obtenir les sites de liaison du facteur de transcription. Pour le DAP-Seq, on peut utiliser un programme comme *meme-suite* (Bailey et al., 2009) qui cherche à aligner les régions liées autour d'un motif commun.

Formules liées à la PWM Le score S d'une séquence de N nucléotides est calculé en ajoutant les poids respectifs de chaque nucléotide de cette séquence.

$$S = \sum_{i}^{N} W(b_i, i) \tag{1}$$

$W(b, i)$: Poids de la base b à la position i

Les poids de la PWM représentent des sortes de pénalités et correspondent au ratio entre chaque fréquence observée et une probabilité attendue $p_{exp}(b)$, converti dans une échelle logarithmique.

$$W_{b,i} = \ln \frac{f(b,i)}{p_{exp}(b)} \tag{2}$$

$f(b,i)$: Fréquence de la base b à la position i
$p_{exp}(b)$: Fréquence attendue de la base b

Comme un nombre d'occurrences nul (figure 11.b) empêche l'obtention de la PWM, On calcule une probabilité observée corrigée de présence de la base b à la position i (figure 11.c).

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in [A,T,G,C]} s(b')} \tag{3}$$

$f_{b,i}$: Nombre d'occurrences de la base b à la position i
N : Nombre de sites
$s(b)$: Nombre de *pseudocounts* de la base b (nombre choisi artificiellement)

Pour obtenir le logo (figure 11.f), on calcule d'abord le contenu en information à une position donnée :

$$R_i = 2 + \sum_{b=A,C,G,T} p_{b,i} \log_2(p_{b,i}) \tag{4}$$

On calcule ensuite la taille d'un nucléotide dans le logo en multipliant le contenu en information par la probabilité du nucléotide :

$$L_{b,i} = R_i p_{b,i} \tag{5}$$

$L_{b,i}$: taille du nucléotide b à la position i dans le logo

---

Remarque :

Si on considère que $p_{exp}(b) = 0.25$ quel que soit le nucléotide b, il a été montré que l'on peut établir un lien entre l'énergie de contribution à la liaison de chaque nucléotide et le score de la PWM, ***ie***, les énergies de contribution de la liaison sont proportionnelles à $-\ln p(b,i)$ (Berg and von Hippel, 1987; Stormo, 2013). L'énergie totale de liaison vaut donc $E = -S$. On peut décider de poser $S' = E - E_{max}$, où E_{max} correspond à la plus forte énergie de liaison possible entre le TF et la séquence d'ADN, de sorte que le meilleur score $S'_{max} = 0$. On calcule les poids de la façon suivante :

$$W'(b,i) = \ln \frac{p(b,i)}{max(p(i))} \Rightarrow S' = \sum_{i}^{N} W'(b_i, i) \tag{6}$$

$W'(b,i)$: Poids de la base b à la position i
$max(p(i))$: Probabilité maximale à la position i

---

a

Site 1	T	T	T	G	T	C	G	C	A	T
Site 2	T	T	T	G	T	C	G	A	A	T
Site 3	C	A	T	G	T	C	G	G	T	C
Site 4	C	A	T	G	T	C	G	G	T	C
Site 5	C	G	T	G	T	C	G	G	T	T
Site 6	C	T	T	G	T	C	C	T	G	A
Site 7	G	T	T	G	T	C	T	G	G	A
Site 8	G	T	T	G	T	C	T	G	G	A
Site 9	T	T	G	G	T	C	G	G	G	T
Site 10	T	T	G	G	T	C	G	G	G	T
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

b

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	0	2	0	0	0	0	0	1	2	3
C	4	0	0	0	0	10	1	1	0	2
G	2	1	2	10	0	0	8	7	5	0
T	4	7	8	0	10	0	2	1	3	5

c

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	0,14	0,23	0,14	0,14	0,14	0,14	0,14	0,18	0,23	0,27
C	0,32	0,14	0,14	0,14	0,14	0,58	0,18	0,18	0,14	0,23
G	0,23	0,18	0,23	0,58	0,14	0,14	0,47	0,45	0,36	0,14
T	0,32	0,45	0,49	0,14	0,58	0,14	0,22	0,18	0,27	0,36

d

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	-1,5	-1	-1,5	-1,5	-1,5	-1,5	-1,5	-1,2	-1	-0,9
C	-0,1	-0,9	-0,9	-0,9	-0,9	0,5	-0,7	-0,7	-0,9	-0,5
G	-0,5	-0,7	-0,5	0,5	-0,9	-0,9	0,3	0,2	0	-0,9
T	-0,7	-0,4	-0,3	-1,5	-0,1	-1,5	-1,1	-1,3	-0,9	-0,6

e

A	T	T	A	G	G	C	C	C	G
-1,52	-0,36	-0,26	-1,52	-0,95	-0,95	-0,7	-0,67	-0,95	-0,95

$\sum$ = - 8,83

f

FIGURE 11: Ici, les séquences liées par le TF ARF5 sont déduites du DAP-Seq. Les séquences de 10 sites de liaison sont alignées (A). La fréquence des nucléotides à chaque position est calculée pour produire la Matrice d'occurrences (B), qui est ensuite convertie en matrice de Fréquences (C), puis en PWM (D). On obtient le score d'une séquence donnée en sommant la poids de ses nucléotides à chaque position (E). Le logo (F) montre les préférences du facteur de transcription.

Limites de la PWM Si les PWM possèdent en général un pouvoir prédictif, elles ne prennent pas en compte les successions de nucléotides, qui confèrent à l'ADN des propriétés particulières. Les modèles suivants essayent au mieux de pallier ces limitations.

I.3.2.2 Prendre en compte les dépendances

Le modèle MORPHEUS (Minguet et al., 2015) Ce modèle permet d'établir des relations de dépendances entre nucléotides. La PWM de LEAFY (figure 12.a), un facteur de transcription responsable de la transition florale, peut être améliorée en prenant en compte des dépendances entre nucléotides (figure 12.b). En effet, le logo montre des particularités pour les positions (4-5-6), (9-10-11) et (14-15-16) (Moyroud et al., 2011b).

Si le modèle MORPHEUS permet d'améliorer efficacement certains modèles, les pénalités affectées à chaque dépendance doivent être déterminées manuellement.

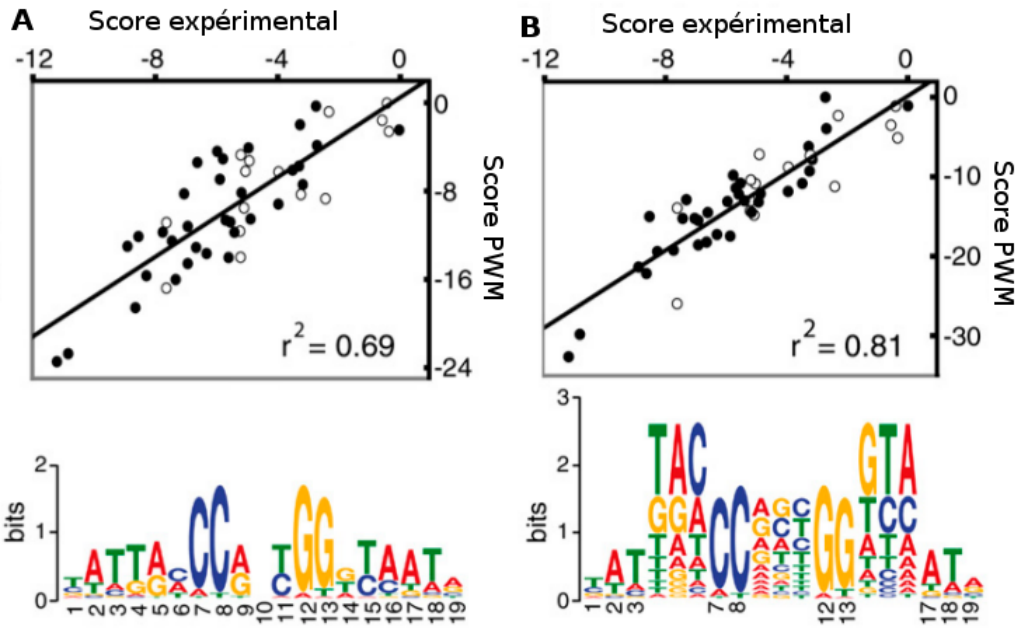


FIGURE 12: En (A), les scores de la PWM de LFY sont tracés en fonction de l'énergie de liaison expérimentale des séquences correspondantes. La prise en compte des dépendances (B) améliore la corrélation entre scores et énergie de liaison. Les cercles blancs décrivent les séquences qui contiennent le consensus CCANTG[G/T]. D'après Moyroud et al. (2011a).

Le *Transcription Factor Flexible Model* (TFFM) Il apparaît que la plupart des dépendances ont lieu entre nucléotides successifs. Plusieurs travaux (voir revue en annexe V.1) permettent de construire des modèles qui prennent ces éléments en compte. La TFFM (Mathelier and Wasserman, 2013) est l'un des plus aboutis parmi ces modèles. En effet, et elle tient compte du fait que certains TF peuvent lier des séquences incluant des gaps et prend en compte le contexte nucléotidique autour du site de liaison. Ses performances dépassent celles de la *Dinucleotide Weight Matrix* (Siddharthan, 2010), le modèle classiquement utilisé pour prendre en compte les dépendances entre nucléotides. Contrairement au modèle MORPHEUS, ce

modèle est capable d'apprendre par lui-même les dépendances si on lui fournit les séquences liées par le TF.

La figure 13, qui représente les préférences de liaison d'un TF à boîte MADS, montre que la succession de nucléotides entre les positions 6 et 11 est importante. Les nucléotides A et T sont tout autant préférés mais il semble qu'il faille éviter la transition T-A.

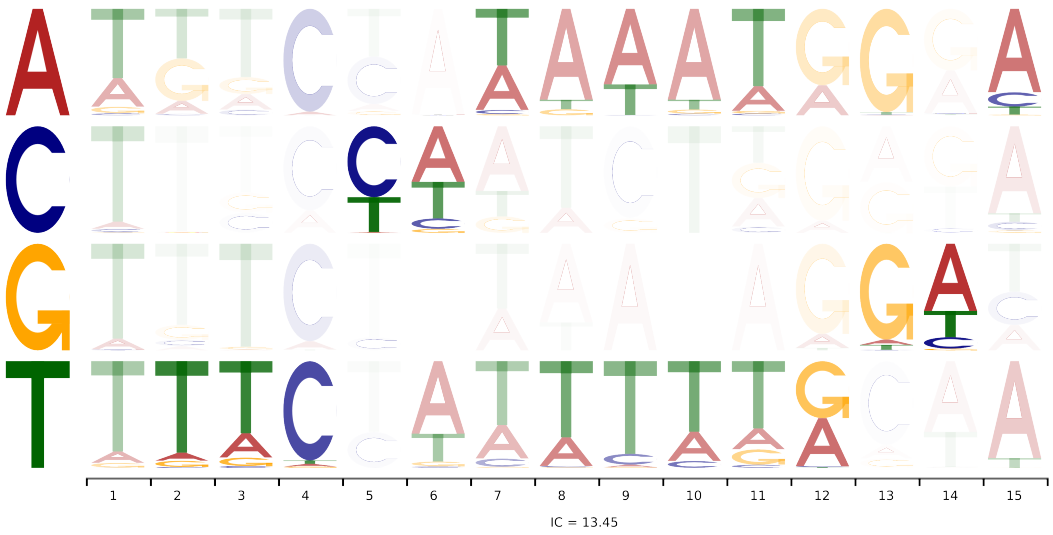


FIGURE 13: Logo de la TFFM d'un TF à boîte MADS. La taille de chaque lettre donne la probabilité d'obtenir ce nucléotide à la position donnée si le nucléotide précédent est celui marqué à l'entrée de la ligne. Par exemple, en position 5, il y a de fortes chances d'obtenir un C ou un T si le nucléotide précédent est un C. La transparence marque la probabilité que le nucléotide précédent soit celui à l'entrée de la ligne. Par exemple, la forte opacité à la ligne G de la position 14 montre que la lettre précédente est très probablement un G.

I.3.2.3 Prendre en compte la structure de l'ADN

Des études ont montré que les successions précises de nucléotides influencent la structure de la double hélice d'ADN (voir revue en annexe V.1). En utilisant des modèles qui prennent en compte les dépendances comme les TFFM, on capture en partie cette information, ce qui explique que la TFFM soit plus efficace que la PWM. Par exemple, les TF à boîte MADS ont besoin que le *propeller twist* (figure 14) reste très faible, ce qui empêche la transition T-A dans la région riche en A/T du motif (figure 13) (Mathelier et al., 2016).

Mais se baser sur une séquence de di-nucléotides – comme le fait la TFFM – n'est pas suffisant pour capturer de façon fiable la totalité des informations de structure de l'ADN (Mathelier et al., 2016). De plus la TFFM n'a pas l'avantage d'expliquer en quoi une succession de di-nucléotides spécifique peut avoir une importance pour un TF donné. Alors que des programmes de modélisation 3D permettent d'obtenir les différents éléments qui caractérisent cette structure, leur utilisation nécessite une grande puissance de calcul qui interdit leur utilisation à l'échelle d'un génome. Aujourd'hui, des laboratoires sont parvenus à obtenir des modèles efficaces capables de modéliser les éléments qui caractérisent cette structure en se basant sur des séquences penta-nucléotidiques (figure 14) (Li et al., 2017; Chiu et al., 2015).

Certains modèles sont ainsi capables d'apprendre les préférences de structure que peut avoir un TF et donner une estimation de la probabilité de liaison de ce TF sur une séquence donnée en se basant sur ces données (Mathelier et al., 2016).

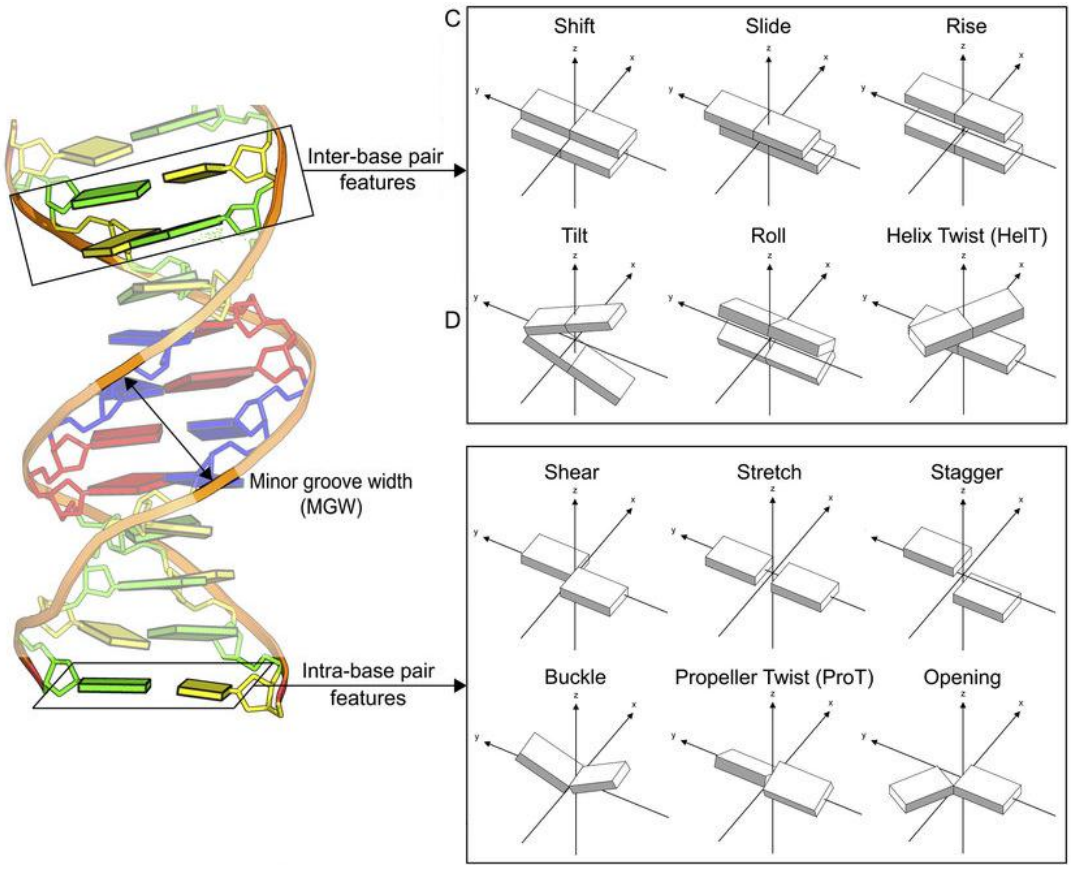


FIGURE 14: Exemple d'éléments qui définissent la structure de l'ADN. D'après Li et al. (2017).

I.3.2.4 Améliorer les modèles en prenant en compte le contexte génomique

Nous avons détaillé l'influence du contexte chromatinien sur la liaison des TF dans le paragraphe I.2.2. Il est évident que la prise en compte de ces paramètres compte énormément et la revue en annexe V.1 fait un inventaire des modèles qui ont été développés.

I.3.2.5 Base de données JASPAR (Khan et al., 2017)

Nous venons de présenter les principaux outils qui permettent de prédire les sites de liaison d'un TF. Il existe des bases de données qui recensent ces profils de liaison et l'une des plus populaire est JASPAR. Réunissant des PWM et des TFFM, elle présente l'avantage d'être en libre-accès et d'être facile d'utilisation. Enfin chaque profil de liaison introduit dans la base de données a au préalablement été vérifié manuellement en étant confronté à une source qui confirme le motif. Cette base de données est mise à jour tous les 2 ans et j'ai participé à la dernière mise à jour de JASPAR *Plants* en introduisant 262 PWM et 218 TFFM. Ces travaux ont abouti à la parution d'un article dont je suis co-premier auteur, présenté en

annexe V.2.

Maintenant que les connaissances de bases sur les fleurs et sur la régulation génique sont exposées, nous allons nous focaliser sur les acteurs moléculaires qui régulent l'expression des gènes au cours du développement floral.

I.4 Le déclenchement de la floraison

Comme nous l'avons mentionné dans le paragraphe I.1.3, les organes aériens de la plante se forment à partir du méristème apical caulinaire. L'architecture de la plante est donc établie à partir des organes qui émergent de ce méristème. La position et l'émergence de nouveaux organes suit des règles précises qui sont déterminées génétiquement. Le principal régulateur de ce processus est une hormone appelée auxine ou acide indole acétique (AIA). Cette hormone est essentiellement fabriquée dans les méristèmes et dans les jeunes feuilles, depuis où elle peut être transportée dans les autres organes de la plante (Paque and Weijers, 2016; Rosquete et al., 2012). Le mutant du transporteur de l'auxine PINFORMED (PIN1) ne forme pas d'organe latéral (Okada et al., 1991; Reinhardt et al., 2000) (figure 15). Il a été montré que les méristèmes axillaires émergent aux positions des maxima d'auxine en induisant l'activité du TF AUXIN RESPONSE FACTOR 5 (ARF5)/MONOPTEROS (MP) (Cole et al., 2009), un mutant faible *mp* présentant les mêmes défauts que le mutant *pin1* (figure 15).

FIGURE 15: Le mutant *pin1* ne forme pas d'organe latéral ou de fleur. Certains allèles du mutant *mp* présente les mêmes défauts. D'après Cheng and Zhao (2007).

Après une première phase au cours de la croissance de la plante où le méristème apical est dans l'état végétatif, des signaux lui indiquent de passer au stade reproducteur en se différenciant en méristème d'inflorescence. Pour cette étape, l'expression de la protéine FLOWERING LOCUS T (FT) est cruciale

(Chahtane, 2014). Celle-ci est contrôlée à de nombreux niveaux, tant par la durée du jour (Song et al., 2013; Valverde et al., 2004; Jang et al., 2008; Laubinger et al., 2006; Zuo et al., 2011) que par la température (Deng et al., 2011; Lee et al., 2007; Li et al., 2008) ou que par l'âge de la plante (Yant et al., 2010). Exprimée dans les feuilles, FT se déplace jusqu'au méristème pour former un complexe avec la protéine FD. Celui-ci va induire l'expression des protéines APETALA1 (AP1) et SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (Wigge et al., 2005). Avec MP, elles vont participer à l'activation d'un facteur de transcription essentiel à la floraison, LEAFY (LFY) (Parcy et al., 1998; Nilsson et al., 1998; Parcy, 2004). Le rôle de MP est primordial, le mutant *mp* ne fait pas de fleur (figure 15), et il a été montré que MP était directement impliqué dans l'induction de *LFY* (Yamaguchi et al., 2013).

Le méristème d'inflorescence a maintenant tous les éléments pour créer des fleurs, trois acteurs étant en balance pour permettre à la fois la formation des fleurs et empêcher que le méristème d'inflorescence ne se transforme en méristème floral (figure 16). Ce sont *AP1*, *LFY* et *TERMINAL FLOWER 1* (*TFL1*), un répresseur de la formation des fleurs qui refrène l'expression de *LFY* et d'*AP1* (Shannon and Meeks-Wagner, 1993). En plus de donner son identité au méristème floral, *LFY* va réguler la formation des organes floraux en activant les gènes homéotiques A, B, C et E (voir paragraphe I.4.4.1), des gènes codant pour des TF à boîte MADS. Ces gènes sont les architectes de la fleur et leurs patrons d'expression définissent l'identité des sépales, des pétales, des étamines et des carpelles.

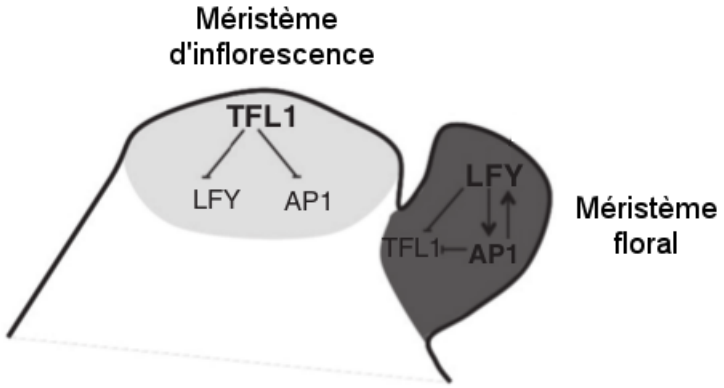


FIGURE 16: Dans le méristème d'inflorescence, l'expression de TFL1 est forte et réprime celle d'AP1 et de LFY. Cette expression est moins forte sur les flancs du méristème d'inflorescence, ce qui permet l'augmentation progressive du niveau d'expression de LFY et d'AP1 et la diminution conjointe de celui de TFL1. La position des méristème floraux sur les flancs du méristème d'inflorescence est donnée par les maxima d'auxine, qui régulent l'activité de *MP*, qui induit l'expression de *LFY*. D'après Chahtane et al. (2014).

Dans cette partie, nous allons d'abord nous pencher sur les mécanismes moléculaires de régulation par l'auxine. Nous examinerons ensuite le fonctionnement de LFY pour nous focaliser enfin sur la différenciation des organes floraux à partir du méristème floral.

I.4.1 Le rôle de l'auxine

I.4.1.1 Présentation de l'auxine

Charles Darwin est aujourd'hui le plus célèbre de tous les naturalistes, universellement connu pour sa théorie de l'évolution. Il est cependant moins connu pour ses travaux sur le coléoptile, le premier organe (qui ressemble à une petite tige) émergeant de terre après la germination de certaines espèces d'angiospermes. En 1881, il observe que l'exposition du coléoptile à la lumière provoque sa courbure. En couvrant différentes parties de la plante, il s'aperçoit que le coléoptile ne se plie que lorsque la lumière peut atteindre son sommet. Il en déduit qu'un signal est produit dans le sommet et voyage dans le reste du coléoptile pour induire sa courbure ; même s'il ne l'a pas encore identifiée, Darwin vient de découvrir une hormone : l'auxine.

Aujourd'hui, l'auxine est une hormone végétale qui a été extrêmement étudiée. Ses effets sont remarquables dans presque tous les processus développementaux de la plante. Au niveau cellulaire, elle contrôle par exemple l'élongation, la différenciation et la réplication. Au niveau de la plante, on peut évoquer son importance sur la dominance du méristème apical, sur les tropismes (comme dans l'expérience de Darwin) ou sur l'organogenèse (des méristèmes axillaires et des fleurs, par exemple) (Li et al., 2016b; Paque and Weijers, 2016). Connaissant cela, il est légitime de se demander comment une seule molécule est capable d'orchestrer une telle quantité de phénomènes cruciaux et pourtant de natures si différentes. Cette question fera office de guide dans notre cheminement mais il faut auparavant introduire certains principes.

L'auxine, dont la grande majorité est synthétisée dans le méristème apical caulinaire (Paque and Weijers, 2016), se déplace dans la plante par un mécanisme connu comme le "transport polaire de l'auxine" (Zažímalová et al., 2010).

$$\mathrm{AIAH} \rightleftharpoons \mathrm{AIA}^- + \mathrm{H}^+$$

En dehors de la cellule, l'auxine est dans sa forme acide AIAH, le pH étant suffisamment faible. L'auxine est alors apolaire et peut traverser la membrane lipidique de la cellule. À l'intérieur de la cellule, le pH est plus alcalin et l'AIA se trouve dans sa forme basique AIA^-, polaire. Elle est donc piégée et ne peut plus sortir de la cellule par elle-même. Les protéines PIN jouent le rôle de transporteurs actifs de l'auxine pour permettre à l'hormone de quitter la cellule. Localisées de manière asymétrique dans la cellule, elles peuvent ainsi orienter la direction de son transport.

I.4.1.2 La voie de signalisation nucléaire par l'auxine

La multitude des processus dans lesquels l'auxine est impliquée, les différentes réponses qu'elle induit dans des organes distincts impose une signalisation complexe qui met de nombreux acteurs à contribution. Il a été montré que l'auxine agit dans plusieurs voies de signalisation. La plus étudiée, et celle que nous détaillerons ici, met en œuvre le complexe TRANSPORT INHIBITOR RESISTANT 1/AUXIN SIGNALING F-BOX (TIR1/AFB), les co-régulateurs transcriptionnels AUXIN/INDOLE-3-ACETIC ACID (Aux/IAA) et les AUXIN RESPONSE FACTORS (ARF), des TF ayant des spécificités de liaison à l'ADN particulières.

Les ARF et l'ADN Dans la plante *A. thaliana*, 23 ARF ont étés identifiés. Deux groupes ont été faits : les activateurs (ARF5-8 ARF21) et les répresseurs, même si l'action répressive de ces ARF ainsi que les mécanismes qui mettent en œuvre une telle répression n'ont pas rigoureusement été mis en évidence (Peer, 2013). Le domaine de liaison des ARF à l'ADN dispose d'un domaine de dimérisation (Boer et al., 2014) (figure 17) qui permet aux ARF de former des dimères en se liant sur des sites appelés *Auxin Responsive*

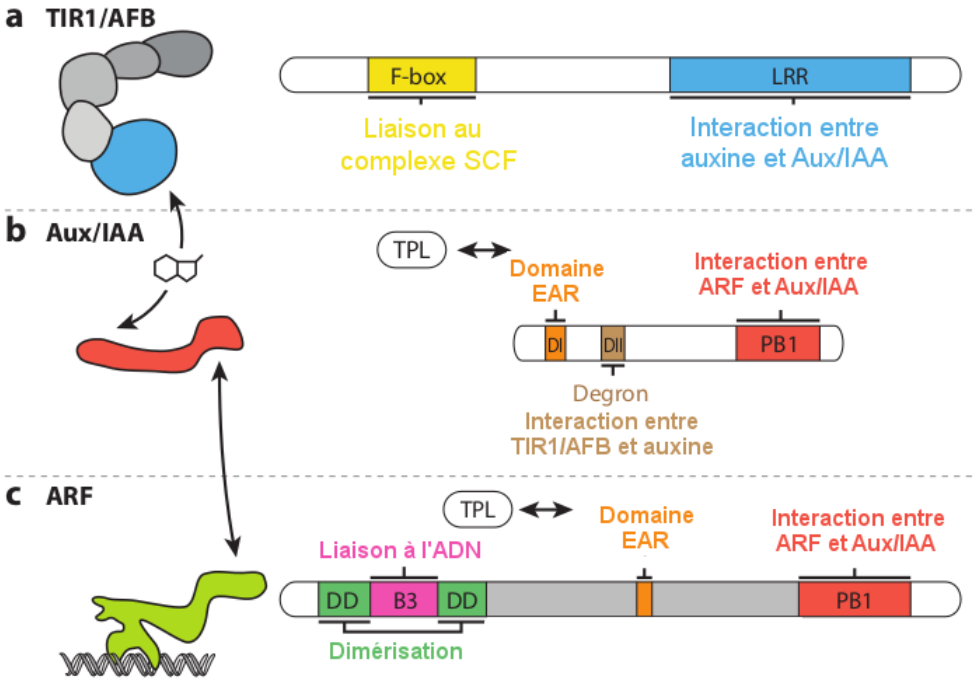


FIGURE 17: En présence d'auxine, TIR1/AFB (A) peut se lier à SCF, une ubiquitinase grâce à sa F-box et aux IAA grâce à leur domaine Degron (B). Les Aux/IAA peuvent se lier aux ARF (C) grâce au domaine PB1 et recruter le co-répresseur TPL par le biais du domaine EAR, ce qui induit le compactage de la chromatine et la répression des gènes liés par les ARF. D'après Weijers and Wagner (2016).

Elements (AuxRE). Nous ne nous attardons pas sur la spécificité de liaison des ARF puisque ce sujet sera traité dans l'introduction du chapitre 1.

Les ARF et les Aux/IAA Pour comprendre l'action des ARF répresseurs et des Aux/IAA, il nous faut détailler très rapidement leur structure, deux domaines en particulier.

Les ARF comme les Aux/IAA ont un domaine Phox/Bem1p (PB1) qui peut permettre une hétéro-oligomérisation ou une homo-oligomérisation entre ARF et/ou Aux/IAA (figure 17). Les ARF activateurs interagissent beaucoup mieux avec les Aux/IAA que les ARF répresseurs (Guilfoyle, 2015).

Un second domaine, ETHYLENE-RESPONSIVE ELEMENT BINDING FACTOR–ASSOCIATED REPRESSOR (EAR), est présent chez les Aux/IAA et les ARF répresseurs. Ce domaine permet de lier le répresseur TOPLESS (TPL) (Causier et al., 2012) qui peut recruter des protéines de désacétylation d'histones (HDAC), induisant le compactage de la chromatine et donc l'inactivation des gènes (figure 17). Remarquons que la structure de TPL d'*A. thaliana* a été résolue dans mon équipe d'accueil par Martin-Arevalillo et al. (2017).

Le complexe TIR1/AFB et les Aux/IAA Les Aux/IAA possèdent domaine Degron, qui a une affinité forte pour l'auxine (figure 17). Grâce à sa région riche en leucine (LRR), le complexe TIR1/AFB

possède également une affinité naturelle pour l'auxine et le domaine Degron des Aux/IAA. L'auxine agit ainsi comme une colle entre le complexe TIR1/AFB et les Aux/IAA. Le complexe TIR1/AFB peut également se lier à une ubiquitinase grâce à sa F-box. Ainsi, lorsqu'il y a de l'auxine, TIR1/AFB induit l'ubiquitination des Aux/IAA, qui sont dégradés par le protéasome (Weijers and Wagner, 2016).

I.4.2 Bilan et réflexion sur la voie de signalisation nucléaire par l'auxine

Ainsi, la présence d'auxine permet d'induire la dégradation des Aux/IAA par le protéasome et l'induction des gènes liés par les ARF : lorsqu'ils ne sont pas liés aux Aux/IAA, les ARF se conduisent comme des activateurs en recrutant des remodeleurs chromatiniens (SWI/SNF) qui éjectent les nucléosomes et rendent l'ADN libres d'accès aux TF (figure 18.b) (Wu et al., 2015; Weijers and Wagner, 2016). À l'inverse, lorsque les Aux/IAA ne sont pas dégradés (faible concentration d'auxine), ils se lient au domaine PB1 des ARF recrutant TPL et HDAC (figure 18.a), réprimant l'expression des gènes liés.

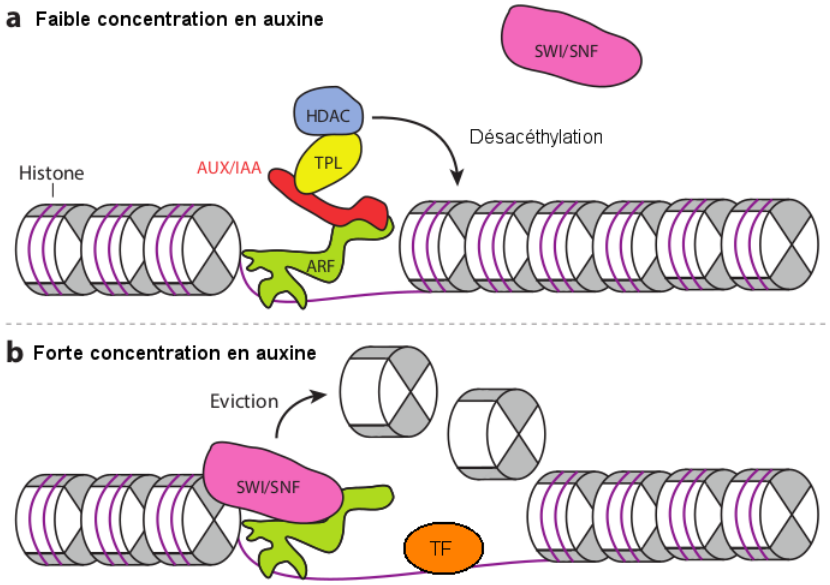


FIGURE 18: À faible concentration d'auxine (A), les Aux/IAA se lient aux ARF, recrutent TPL qui peut se lier à de HDAC qui compactent l'ADN, et désactivent donc l'expression des gènes liés par les ARF. À forte concentration d'auxine (B), les ARF recrutent des remodeleurs chromatiniens qui éjectent les nucléosomes activant l'expression des gènes. D'après Weijers and Wagner (2016).

Qu'en est-il des ARF répresseurs ? Leur interaction avec les Aux/IAA est faible et ils peuvent à eux seuls recruter TPL, mais le lien qu'ils ont avec l'auxine est moins clair. On peut cependant imaginer qu'ils sont exprimés en même temps que les ARF activateurs et qu'ils entrent en compétition avec les mêmes sites de liaison sur l'ADN.

Connaissant maintenant les mécanismes moléculaires orchestrés par l'auxine, la figure 19 propose un mécanisme qui réponde à la question posée dans le paragraphe I.4.1.1 : comment l'auxine orchestre autant de mécanismes différents dans des types cellulaires tout aussi différents.

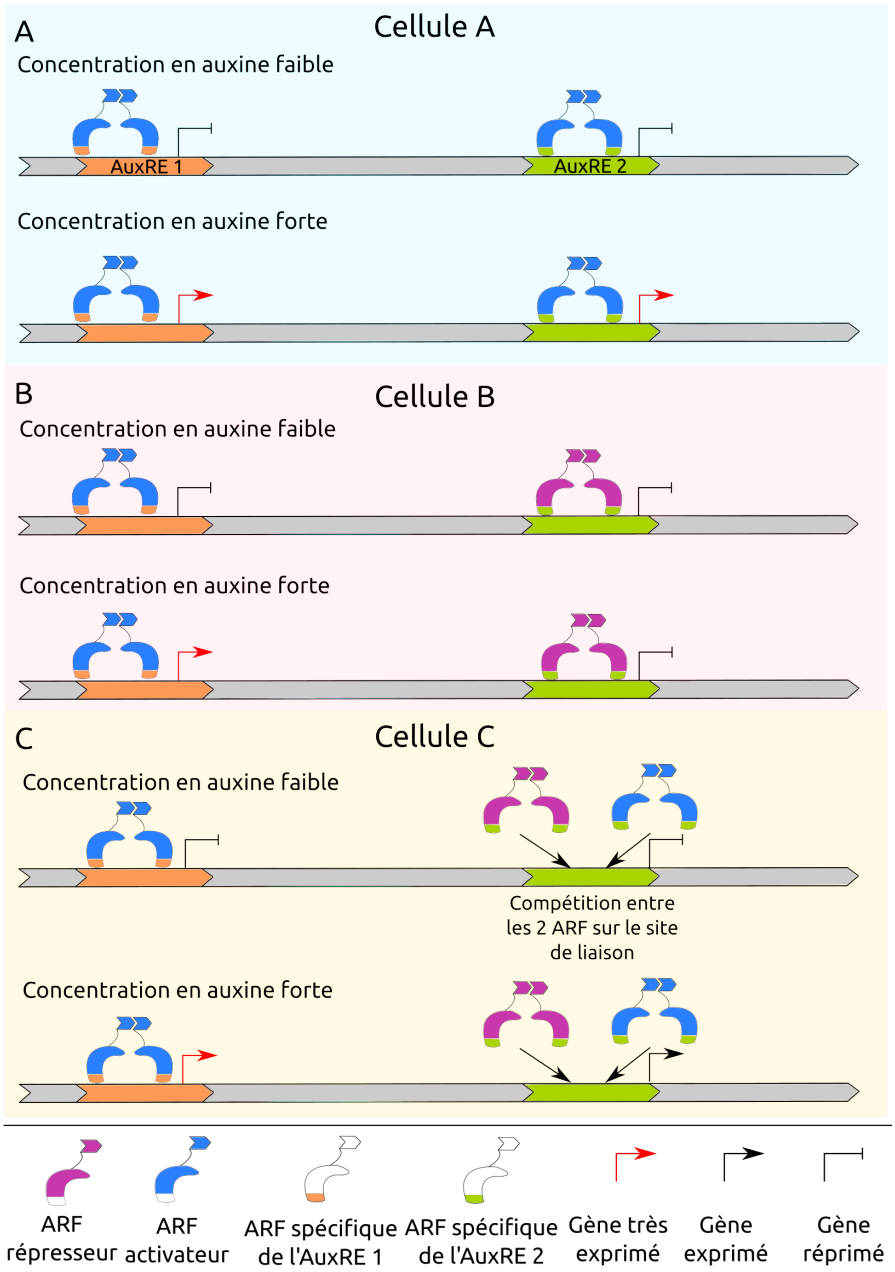


FIGURE 19: Deux gènes contiennent des AuxRE différents dans leurs promoteurs. En (A), deux ARF activateurs sont présents et sont respectivement spécifiques de l'un et de l'autre. En présence d'auxine, la transcription des gènes est activée. En (B), deux ARF sont présents, l'un activateur (spécifique du premier AuxRE) et l'autre répresseur (spécifique du second AuxRE). Le premier gène est activé en présence d'auxine alors que le second reste inactif quelle que soit la concentration en auxine. En (C), un ARF activateur est spécifique du premier AuxRE alors que deux ARF, l'un activateur et l'autre répresseur, sont en compétition pour se lier sur le deuxième AuxRE. Le niveau de transcrit du gène correspondant en présence d'auxine dépendra de la quantité de chacun de ces ARF dans le noyau et de leur affinité respective pour l'AuxRE.

I.4.3 LEAFY, un gène maître du développement floral

I.4.3.1 D'où vient *LEAFY* ?

LFY est l'un des gènes les plus importants à la fois pour l'émergence des méristèmes floraux et leur plan d'organisation avec leurs couronnes de natures différentes. Même si chez les angiospermes, *LFY* agit principalement sur la formation des fleurs, ce gène est apparu très tôt au cours de l'évolution des plantes. On retrouve en effet ses homologues chez les bryophytes – les premières plantes terrestres, apparues bien avant les angiospermes – et chez certaines algues charophytes (Moyroud et al., 2010). Si les organes reproducteurs des mousses telles que *Physcomitrella patens* sont très différents des fleurs, il a été observé que *PpLFY* y joue un rôle dans le développement précoce de la mousse. *PpLFY* est exprimé dans les spores et sa perte de fonction bloque la division cellulaire immédiatement après la fécondation (Tanahashi et al., 2005). Il semble que la protéine LFY soit un facteur de transcription aussi bien dans les angiospermes que dans les mousses et les fougères ; même si les cibles de CrLFY chez la fougère *Ceratopteris richardii* demeurent inconnues, celui-ci est capable de complémenter un mutant *lfy* dans *A. thaliana* ; notons cependant que *PpLFY* n'y parvient pas (Maizel et al., 2005).

Chez les gymnospermes, qui ne forment pas de fleurs mais des cônes mâles et femelles, on trouve également deux homologues de *LFY*. Leur forte expression dans les cônes laisse penser qu'il est impliqué dans leur développement. Des homologues des gènes B et C existent chez les gymnospermes et s'il n'y a pas de preuve formelle qu'ils soient régulés par *LFY*, son expression précède celle des gènes B et C (Theißen and Becker, 2004; Moyroud et al., 2010). La préexistence d'un réseau régulateur entre *LFY* et les gènes B et C dans la famille des gymnospermes est une information importante. Cela laisse penser que de très petites modifications aient pu suffire pour permettre l'apparition des fleurs, puisqu'une partie du réseau génétique était déjà en place.

I.4.3.2 À propos de la protéine LFY

La région C-terminale La région C-terminale de LFY présente 2 caractéristiques notables : la première est le domaine de liaison à l'ADN et la seconde est un domaine d'homodimérisation. Il a été observé que le dimère constitué de deux protéines LFY est plus stable sur l'ADN que le monomère, la rupture du domaine de dimérisation affaiblissant l'affinité de liaison (Hamès et al., 2008). C'est pourquoi la PWM de LFY de *A. thaliana* (donnée en figure 12) est celle du dimère (Hamès et al., 2008; Moyroud et al., 2011b). La spécificité de LFY reste presque la même chez les fougères, les gymnospermes et les angiospermes mais change chez les bryophytes (Sayou et al., 2014), ce qui explique qu'un *LFY* de fougères parvienne à complémenter un mutant *lfy* d'*A. thaliana* alors qu'un *LFY* de mousse n'y parvient pas.

La région N-terminale Des analyses phylogéniques ont montré qu'une partie de la région N-terminale de LFY était conservée au cours de l'évolution, attestant de son rôle important. Cette région s'est révélée être un domaine d'oligomérisation du type *sterile alpha motif* (SAM). Il a été montré que l'oligomérisation améliore la liaison à l'ADN dans les régions compactes de la chromatine, ce qui laisse penser que LFY pourrait avoir un rôle de facteur de transcription pionnier, capable de reconnaître ses cibles même dans les régions fermées du génome et peut-être d'induire leur ouverture (Sayou et al., 2016).

I.4.3.3 LEAFY dans *Arabidopsis thaliana*

LFY, gène maître de la floraison Des expériences ont montré le rôle indispensable de *LFY* dans la floraison. Chez *A. thaliana*, le mutant *lfy* ne fait pas de fleur fertile, des tiges ou des fleurs mal différenciées poussant à leur place (figure 20.A). À l'inverse, son expression constitutive dans les premières semaines de la plante entraîne une floraison précoce et les fleurs sont alors ectopiques, apparaissant à la base des feuilles (figure 20.B) (Weigel and Nilsson, 1995).

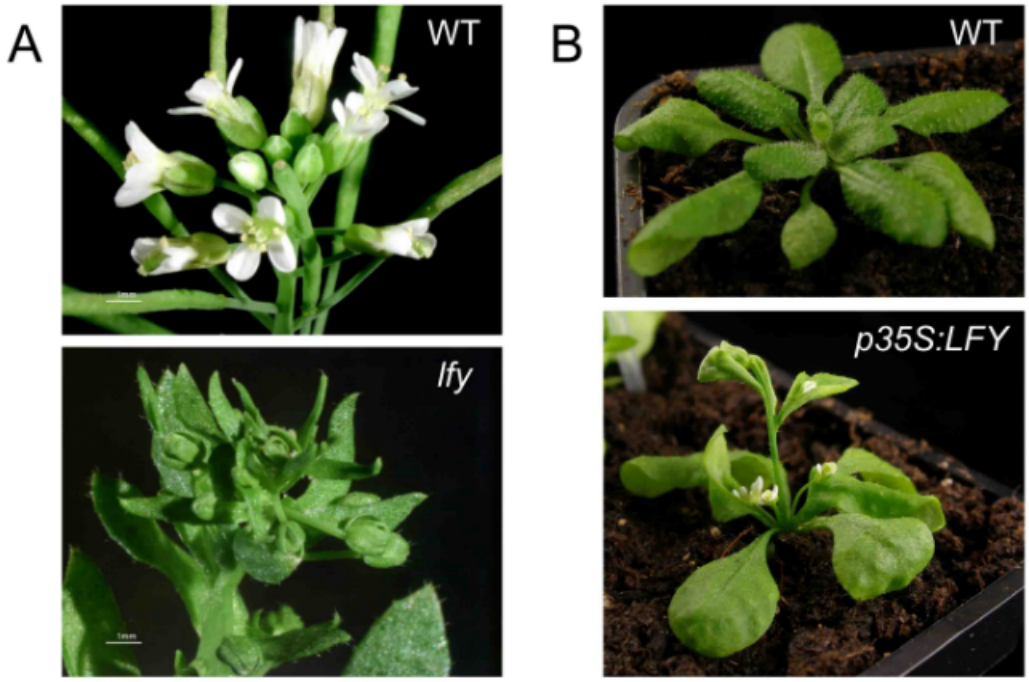


FIGURE 20: (A), *LFY* est nécessaire à la présence de fleur. (B), l'expression constitutive de *LFY* fait apparaître des fleurs à la base des feuilles. D'après Sayou (2013).

LFY définit l'identité du méristème floral Nous avons évoqué ce rôle dans l'introduction du paragraphe I.4 et dans la figure 16. Ajoutons que le mutant *tfl1*, alors incapable de réprimer *LFY* dans les méristèmes d'inflorescence, voit ces derniers se différencier en méristèmes floraux, produisant des fleurs terminales (Shannon and Meeks-Wagner, 1993).

LFY et l'identité des organes floraux Une fois que *LFY* est exprimé dans le méristème floral, son expression persiste pour induire celle des gènes A, B, C et E, qui détermineront la nature des organes floraux. Le rôle de *LFY* et de ces gènes sera détaillé dans la section suivante.

I.4.4 Les gènes A, B, C et E contrôlent l'identité des organes floraux

I.4.4.1 Présentation des gènes du modèle ABCE

Le modèle ABC (figure 21) a été proposé par Coen et Meyerowitz en 1991. Avant de devenir un spécialiste d'*A. thaliana*, Elliot Meyerowitz s'intéressait à l'organisation des organes chez la drosophile, chez

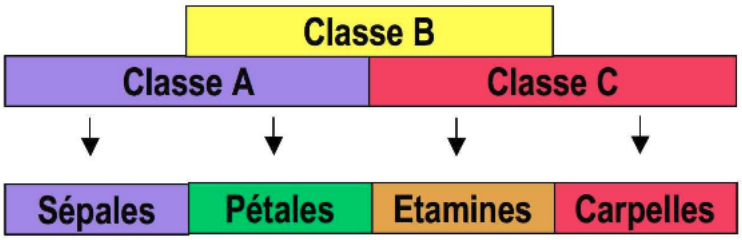


FIGURE 21: Les gènes A contrôlent la formation des sépales. La genèse des pétales est orchestrée par la mise en commun des gènes A et B. Les gènes B et C gouvernent la formation des étamines ensemble et les gènes C régulent celle du carpelle

laquelle une famille de gènes, dits homéotiques, déterminent l'identité de certains organes de la mouche. Il se tourne vers les fleurs, et avec Coen, ils observent que chez certains mutants d'Arabidopsis ou de muflier, des pétales sont convertis en sépales ou que des étamines sont converties en pétales (Coen and Meyerowitz, 1991). Ils déduisent un modèle (figure 21) où des combinaisons de gènes contrôlent l'identité des organes dans les différentes couronnes.

Chez *A. thaliana*, les fonctions A, B et C sont respectivement assurées par :

— *AP1* (en plus de son rôle que dans l'identité du méristème floral) et *APETALA2* (*AP2*)

— *PISTILLATA* (*PI*) et *APETALA3* (*AP3*)

— *AGAMOUS* (*AG*)

Les gènes E, qui participent à l'identité de toutes les couronnes, ont ensuite été rajoutés au modèle (Pelaz et al., 2000). Chez *A. thaliana*, ces gènes sont *SEPALLATA1-4* (*SEP1-4*), quatre gènes codants pour des protéines assez semblables.

L'un des rôles de LFY est d'activer directement l'expression des gènes *AP1*, *AP3*, *AG* et *SEP3* (Weigel and Nilsson, 1995; Parcy et al., 1998; Moyroud et al., 2011b; Winter et al., 2011). LFY étant exprimé dans tout le méristème floral, il a besoin de partenaires pour spécifier les patrons d'expression des gènes A,B, C et E. Ainsi, LFY interagit avec la protéine UNUSUAL FLORAL ORGANS (Lee et al., 1997) pour activer le gène B (*AP3*). Enfin, l'expression d'*AG* est assurée par le complexe LFY-WUSCHEL (Lohmann et al., 2001).

Tous les gènes du modèle ABCE codent pour des TF et mis à part AP2, ils appartiennent tous à la famille des TF à boîte MADS.

I.4.4.2 Spécificité des TF à boîte MADS

De nombreuses études se sont focalisées sur la spécificité de liaison des TF à boîte MADS sur leurs sites, appelés CArGbox. Huang et al. (1993) montrent que AG lie la séquence TT(A/T)**CC(A/T)$_3$(T/A)-NNGG**(G)(A/T)$_2$ alors que le TF MEF2A préfère CTCGG**CTATTAATAG**CCGAG (Huang et al., 2000). Chez l'homme, les TF à boîte MADS SRF et MEF1 ont des préférences respectives pour TGC**CC(A/T)T-ATA(T/A)GG**(T/A)NNT et C**CC(T/C)AA(T/A)NNGG**TAA (Pollock and Treisman, 1990; Wynne and Treisman, 1992). Il apparaît cependant que ces deux facteurs possèdent une plus grande diversité de sites de liaison ; MEF1 peut lier les sites de SRF alors que MEF1 peut lier ceux de SRF (Passmore et al., 1989; Hayes et al., 1988). En raison des grandes similitudes dans les préférences de liaison des TF à boîtes MADS, la séquence consensus $CC(A/T)_{n=6}GG$ (qui correspond à la partie en gras dans les séquences

détaillées plus haut) est généralement utilisé pour décrire le site de liaison de ces TF (Egea-Cortines et al., 1999; Honma and Goto, 2001; Theissen and Saedler, 2001). Un logo (voir I.3.2.1 sur les PWM), obtenu par ChIP-Seq, est employé pour la première fois par Kaufmann et al. (2009) afin de décrire le site de liaison de la protéine SEP3.

Dans les plantes, les gènes du modèle ABC font partie de la sous-famille des gènes MIKC, n'existant que le règne végétal (Kaufmann et al., 2005). Au domaine M (MADS), s'ajoutent les domaines *intervening* (I), *keratin* (K), et C-terminal (C) (Theißen et al., 1996; Kaufmann et al., 2005). Alors que le domaine M contient le domaine de liaison à l'ADN, les autres domaines permettent l'interaction entre protéines, ce qui permet aux TF du modèle ABC de former les complexes régulateurs des différentes couronnes ou d'intéragir avec d'autres protéines. Cependant, alors que ces combinaisons de TF régulent des gènes différents, les complexes formés par les TF A,B et C semblent tous partager la même préférence pour la séquence $CC(A/T)_{n=6}GG$ (Riechmann et al., 1996).

Même si les combinaisons possibles de TF A,B,ou C sont différentes suivant les couronnes de la fleur, il est difficile d'expliquer que les gènes qu'ils activent soient également différents si leur site de liaison est identique. Une étude récente basée sur du SELEX a révélé que le nombre n pourrait varier en fonction des combinaisons de dimères. Ainsi, AG-AG pourrait préférer $n = 4$ (Smaczniak et al., 2017). Des données *in vitro* qui montrent que les dimères peuvent interagir pour former des tétramères ont donné naissance au modèle des quartets (Melzer et al., 2008; Melzer and Theißen, 2009) (figure 22). Même si des quartets de TF liés sur l'ADN *in vivo* n'ont pas été observés, le mutant *sep1 sep2 sep3* est complémenté par SEP3 mais seulement partiellement par $SEP3_{del}$, dont l'interface de tétramérisation a été détériorée (Hugouvieux et al., 2018) [2] (figure 23). Le modèle du quartet offre une nouvelle perspective quant à la spécificité des TF à boîte MADS : on peut imaginer que les différents tétramères aient des préférences spécifiques en terme d'espacement entre deux CArGbox.

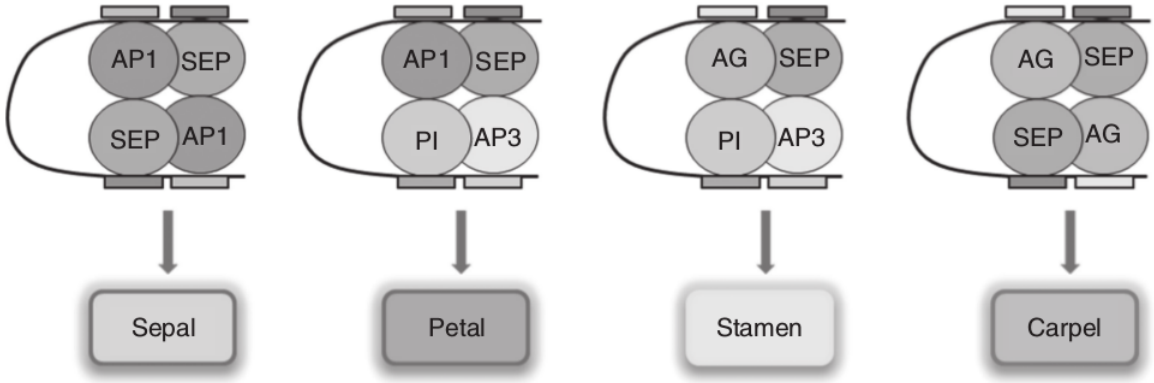


FIGURE 22: Modèle des quartets. Les TF peuvent former des tétramères en se liant sur des sites de liaison suffisamment espacés pour que l'ADN fasse des boucles. D'après Chahtane et al. (2014).

2. Ces travaux ont été réalisés dans notre laboratoire et je suis l'un des co-auteurs. Cependant, en raison de ma contribution mineure, cet article ne figure ni dans le corps ni dans les annexes de ce manuscrit

FIGURE 23: (I) Fleur sauvage. (J) Le mutant *sep1 sep2 sep3* ne forme que des sépales. Ce mutant est complémenté par SEP3 (K) alors que les organes de *sep1 sep2 sep3* $SEP3_{del}$ montrent des défauts (l).

Objectifs

Pour des raisons fondamentales ou plus appliquées, la floraison et la formation des fleurs ont été largement étudiés et aujourd'hui, les réseaux de gènes et de TF qui contrôlent ce processus sont assez bien établis. Cependant, le fonctionnement de ces TF reste encore relativement mystérieux : qu'il s'agisse des ARF, de LFY ou des protéines à boîte MADS, la liste de leurs cibles sont souvent établies expérimentalement sans être comprises au point de pouvoir les prédire. Plus tôt dans l'introduction, nous avions déclaré qu'être capable de prédire où les TF se lient est un grand pas dans la compréhension des processus de régulation. Par conséquent, cette thèse a pour ambition d'éclairer le mécanisme de formation des fleurs depuis l'établissement du méristème floral jusqu'à la formation des organes floraux grâce à des modèles bioinformatiques prédictifs de la liaison des TF clés.

Le premier chapitre sera une étude sur la spécificité des ARF. Quelles sont leurs motifs reconnus ? Diffèrent-ils d'une protéine ARF à l'autre ? Parmi ces gènes cibles, lesquels sont régulés ? Peut-on établir des règles ?

Le deuxième chapitre sera consacré à mettre en évidence la nature des éléments qui influencent la liaison d'un TF *in vivo* à travers l'étude de *LFY*.

Enfin, le troisième chapitre portera sur les gènes ABCE. La tétramérisation influe-t-elle sur spécificité de liaison aux gènes à boîte MADS ? Auquel cas, observe-t-on cette spécificité *in vivo* ?.

Les méthodes utilisées pour obtenir les résultats seront présentées dans la partie suivante.

Méthodes

II.1 Pré-traitement des données brutes de DAP-Seq

Cette section s'applique essentiellement au chapitre 3. Les données brutes de DAP-Seq sont sous la forme de *reads* appariés (*paired reads*) de 150 paires de bases. Cela signifie que les deux extrémités $5'$ des fragments liés par le TF d'intérêt ont été séquencées.

II.1.1 Traitement des *reads*

II.1.1.1 Format des *reads*

Le format utilisé est appelé FastQ. Pour chaque librairie, les *reads* sont appariés et stockés dans deux fichiers : les séquences de la première extrémité de chaque fragment sont dans un premier fichier, les séquences de l'autre extrémité sont dans le second fichier. Le format FastQ se présente sous la forme suivante :

```
@GWNJ-0901:397:GW1903011914:5:1101:13250:1309 1:N:0:NTCACG
NAGCCCATTTATTTAGCAACTCAAGATTGGTTTCATCGAGAGATGACCAACCGTTACCACGGGTTTTCTCTGA
+
#AAFFJJAAFJ-J<JJJFJ<JJJJJJAFJJJFJJJJJJFJJJJJJAJJJJJJJFJJJJJFJJJ-AFJFFJFFA
```

La première ligne, qui commence par "@" contient le nom du fragment. Lors de l'alignement sur le génome, cela permet de trouver le *read* du brin complémentaire dans le second fichier. La deuxième ligne est la séquence du *read*. Chaque nucléotide est associé à une qualité, reportée sur la dernière ligne au format ASCII. La qualité Q donne la probabilité p que le nucléotide soit mal identifié : $Q = -10 \log p$. Cette qualité est ensuite convertie au format ASCII. Sur la séquence ci-dessus, on constate que la première base affiche une qualité de "#", qui correspond presque à la plus mauvaise qualité possible. Ceci explique que le nucléotide (ici N) n'ait pas été déterminé correctement.

II.1.1.2 Qualité des *reads*

La qualité des *reads* a été évaluée grâce au logiciel FastQc. Les rapports ont montré que les extrémités $3'$ étaient enrichies en adaptateurs (figure 9). Ceux-ci doivent être coupés avant que l'on puisse aligner les *reads* sur le génome. En effet, la présence d'adaptateurs peut d'une part empêcher l'alignement et peut d'autre part provoquer l'alignement de ces *reads* à des positions du génome qui contiennent des séquences proches des adaptateurs. L'observation des *reads* a également montré que la première base en $5'$ avait une qualité très faible et nous avons décidé de l'enlever.

II.1.1.3 Suppression des adaptateurs

Les adaptateurs ont été supprimés grâce au logiciel NGmerge avec l'option -a, qui indique au programme de supprimer les adaptateurs et l'option -q 35 qui met un seuil sur la qualité des séquences. Nous n'avons pas compris l'influence du seuil sur le comportement du programme. Le programme s'arrêtant subitement et sans raison apparente avec l'option par défaut -q 33, nous avons choisi -q 35. Le logiciel FastQc a montré que les adaptateurs ont été correctement supprimés par NGmerge. La base de mauvaise qualité en 5′ a été supprimée grâce au programme Trimmomatic avec l'option HEADCROP :1.

II.1.1.4 Alignement des *reads* sur le génome

Le logiciel bowtie2 (Langmead and Salzberg, 2012) a été choisi pour cette étape, qui nécessite un soin particulier. En effet, avant de supprimer les adaptateurs et l'extrémité en 5′, les fragments peuvent être représentés de cette façon :

```
5′ ----N====================++++ 3′ R1
3′ ++++====================N---- 5′ R2
```

```
---- : adaptateur en 5′
++++ : adaptateur en 3′
==== : séquence génomique à aligner
N : Base de mauvaise qualité en 5′
```

D'après FastQc, les adaptateurs en 5′ ont déjà été enlevés, la partie séquencée est donc en gras ci-dessous (dans les DAP-Seq que nous avons traité, 150 paires de bases en 5' des fragments ont été séquencées) :

```
5′      N==================++++ 3′ R1
3′ ++++====================N     5′ R2
```

La présence d'adaptateurs en 3′ indique que les fragments sont trop courts, ce qui conduit à en séquencer une partie :

```
5′      N===++++ 3′ R1
3′ +++===N       5′ R2
```

Après que les adaptateurs sont supprimés avec NGmerge, on se retrouve avec les *reads* suivants :

```
5′ N=== 3′ R1
3′ ===N 5′ R2
```

Comme décrit dans le paragraphe précédent, le **N** est enlevé :

```
5′  === 3′ R1
3′ ===  5′ R2
```

Par défaut, bowtie2 n'aligne pas ces *reads* sur le génome car l'extrémité 5′ du premier fragment commence après l'extrémité 3′ du second fragment. Ce problème est résolu en utilisant l'option --dovetail.

II.1.1.5 Filtrer les *reads* alignés par bowtie2

En observant les *reads* sur le navigateur IGB (Freese et al., 2016), nous avons constaté que toutes les librairies séquencées, y compris le contrôle (l'ADN nu amplifié par PCR), affichaient une grande quantité de *reads* de mauvaise qualité dans les régions centromériques. Ce phénomène a notamment été décrit par *ENCODE* qui met des listes noires de régions à disposition (Consortium et al., 2012) et par Carroll et al. (2014). Il est d'usage de supprimer les *reads* qui s'alignent dans ces régions avant de rechercher celles liées par le TF d'intérêt. Cependant, l'étude de Cheneby et al. (2018) propose de filtrer les reads avec les commandes suivantes :

```
$ grep -e "^@" -e "XM:i:[012][^0-9]"
$ grep -v "XS:i:"
```

Ceci a pour effet de supprimer les reads indésirables sans avoir au préalable besoin d'une liste noire de régions.

La première ligne supprime les *reads* alignés dont au moins 3 bases ne concordent pas avec la séquence de référence grâce à l'option -e "XM:i:[012][^0-9]". Pour chaque *read* dans le fichier ".sam", bowtie2 ajoute le *flag* "XM :i :n" où n est le nombre de nucléotides qui ne concordent pas avec la séquence de référence. Chaque *read* étant écrit sur une ligne, on décide donc de ne conserver que les lignes où n commence par 0, 1 ou 2 ([012]) et où le caractère suivant n'est pas un chiffre ([^O-9]). On garde l'entête de la sortie, qui est au format ".sam" [3] grâce à l'option -e "^@", qui conserve toute les lignes commençant par "@" (nécessaire pour compresser le fichier au format ".bam").

Le *flag* "XS :i :k" donne le nombre de positions k où le *read* peut être aligné. Ce *flag* n'apparaît pas si $k = 1$. L'option "-v" indique au programme de ne conserver que les lignes qui ne contiennent pas "XS :i :". La seconde ligne supprime donc tous les *reads* qui peuvent être alignés à plusieurs positions du génome.

Note : À l'issue du filtrage, une régions très enrichie en *reads* subsiste dans le chromosome 2. Cette région a été décrite par Lin et al. (1999) comme l'insertion d'une partie du génome mitochondrial. Celui-ci étant présent dans l'ADN utilisé pour le DAP-Seq, les fragments sont alignés par erreur sur le chromosome 2. Pour les cas étudiés dans le chapitre 3, ces régions ne sont pas retenues comme étant liées par le *peak caller* (voir II.1.2.1 dans la suite des méthodes), certainement en raison de l'enrichissement fort en *reads* dans le contrôle. Nous avons donc décidé de ne pas filtrer cette région. Cependant, en utilisant les mêmes paramètre de *peak calling* pour l'étude du DAP-Seq de LFY (chapitre 2), des pics sont présents dans cette région et nous les supprimons manuellement.

Les fichiers de *reads* alignés, alors au format ".sam", sont compressés puis triés grâce aux commandes de la suite samtools (Li et al., 2009) :

```
$ samtools view -b fichier.sam | samtools sort - > fichier.bam
```

Le tri rend l'exploitation des fichiers beaucoup plus aisée, celui-ci étant nécessaire pour de nombreuses opérations.

Certain fragments étant parfois excessivement amplifiés par PCR, nous avons fait le choix d'écarter les duplicats en utilisant la commande :

```
$ samtools rmdup fichier.bam
```

3. format qui donne entre autre la position de chaque *read* aligné sur le génome

II.1.2 Déterminer les régions liées par le TF

À ce stade, il nous faut préciser que nous avions entre 2 et 3 réplicats pour chacune de nos librairies. Ce paragraphe détaille d'une part les méthodes utilisées pour évaluer la reproductibilité des expériences et d'autre part les méthodes utilisées pour sélectionner les régions liées par un TF donné à partir des réplicats.

II.1.2.1 Évaluer la qualité des réplicats

Un bruit de fond est toujours inhérent aux expériences de séquençage : en DAP-Seq de nombreuses régions ne sont pas liées par le TF et sont néanmoins capturées, puis séquencées. Le bruit est intrinsèquement lié au hasard et ne permet donc pas de nous assurer de la reproductibilité des expériences. La première étape est donc de déterminer les régions probablement liées par le TF au sein de chaque réplicat ; sur ces régions les *reads* ne sont pas capturés par hasard et peuvent donner lieu à une comparaison. À cette fin, on utilise le programme macs2 (Zhang et al., 2008) ; macs2 est un *peak caller* car il trouve ces régions, appelées "pics". On lance macs2 en ligne de commande de la manière suivante :

```
$ macs2 callpeak -t fichier.bam -c temoin.bam -f BAMPE -g 120000000\
--nomodel -B
```

-f BAMPE indique que les *reads* sont pairés
-g 120000000 indique la taille du génome (donnée sur le site `https://www.arabidopsis.org/`)
-B pour obtenir le signal DAP-Seq au format ".bedgraph" [4]. Précisons que le fichier est normalisé par le nombre de *reads* total dans "fichier.bam", de sorte que la profondeur de séquençage n'affecte pas les résultats.
--nomodel car les reads pairés donnent la position exacte du fragment (dans le cas de *reads* non pairés, macs2 créé un modèle pour élargir les *reads* en $3'$ afin de déterminer la position exacte des fragments, les *reads* donnant seulement leur extrémité)
-t fichier.bam : reads capturés par DAP-Seq
-c temoin.bam : On utilise de l'ADN génomique amplifié par PCR, puis séquencé. Cela permet de réduire le biais de séquence (contenu en nucléotides GC, structure de l'ADN, ...)

Une fois les pics déterminés, on compare les réplicats deux à deux. Un premier script que j'ai écrit en python permet de distinguer les pics communs aux 2 réplicats de ceux qui ne le sont pas. On décide que deux pics sont communs si le plus grand recouvre 80% du plus petit ; auquel cas, le pic commun est défini comme l'intersection stricte des 2 pics. Si deux pics sont définis communs mais que plus de 50% du plus grand pic demeure non couvert par le plus petit, un pic non commun correspondant à la région non couverte du plus grand pic est créé (figure 24).

Le programme bedtools intersect (Quinlan and Hall, 2010) associé a un second script personnel écrit en python permettent de calculer le nombre de *reads* sous chaque pic, dans chaque réplicat, à partir de leur position et du fichier ".bedgraph" fourni en sortie de macs2. Nous rappelons que les fichiers ".bedgraph" sont normalisés par la profondeur du séquençage. La taille des pics pouvant influencer les résultats, on divise le nombre de *reads* sous chaque pic par celle-ci. À partir d'ici, ce nombre sera appelé "couverture normalisée". Un script écrit en R permet enfin de comparer les 2 réplicats (figure 25). Ainsi, en comparant la couverture normalisée sous les pics, dans les deux réplicats, nous pouvons évaluer la reproductibilité des expériences (figure 25).

4. format qui donne le nombre de *reads* en tout point du génome

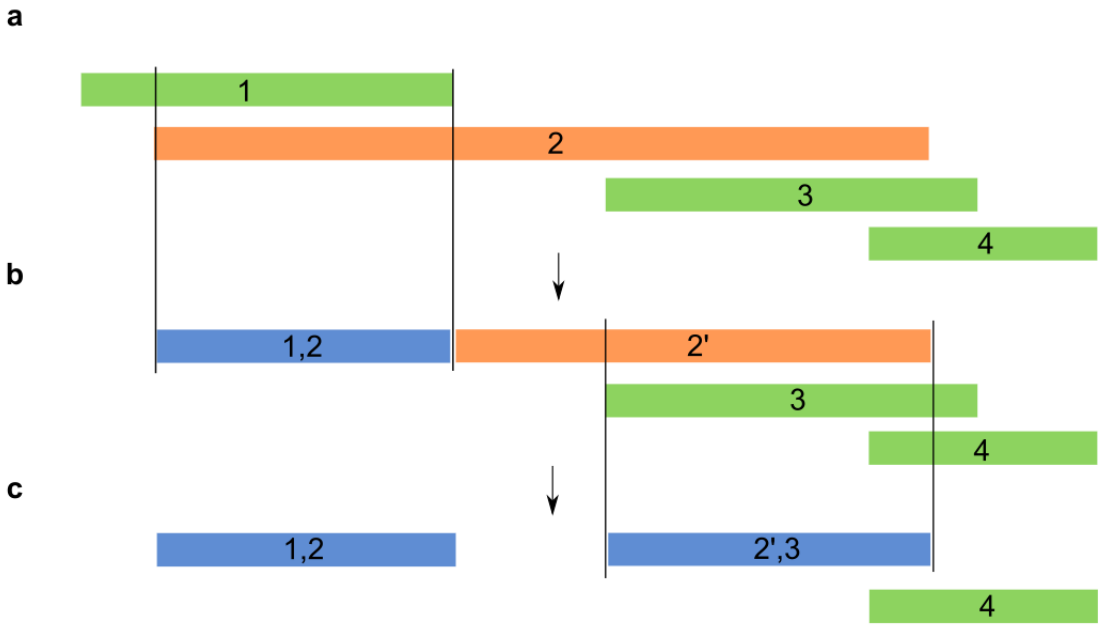


FIGURE 24: Sélection de pics communs à deux échantillons. Les couleurs donnent les pics du réplicat 1 (en vert), du réplicat 2 (en orange) et les pics communs (en bleu). (A) Le pic 2 superpose 80% d'un pic plus petit, le pic commun $1, 2$ est donc créé (B). La portion restante du pic 2 est supérieure à 50% du pic total (A), on créé donc le pic $2'$ (B). Le pic 3 recouvre plus de 80% du pic $2'$. On créé donc le pic $2', 3$ (C). Le pic 4 ne convient pas aux critères pour créer un pic commun et reste donc unique au réplicat 1.

II.1.2.2 Déterminer les pics significatifs pour un TF donné

Après nous être assurés que les expériences sont reproductibles, il nous faut décider quels pics utiliser pour les analyses futures. Une première étape consiste à prendre uniquement les pics communs aux réplicats (ce qui revient à supprimer les pic rouges et verts sur la figure 25). Ce filtrage ne nous a pas semblé suffisant. En effet, nous nous sommes aperçus que la grande variabilité entre réplicats d'une même expérience était souvent forte pour les pics de faible intensité. Cela pose des questions sur la robustesse des conclusions basées sur ces pics lorsque nous comparons une expérience à une autre. Nous avons donc fait le choix d'appliquer un autre filtre utilisé par *ENCODE* appelé *Irreproducible discovery rate* (IDR) (Li et al., 2011). L'IDR se calcule à partir du rang de chaque pic dans les deux réplicats ; lorsque les rangs diffèrent trop dans les deux réplicats, cela signifie que les pics correspondent probablement à du bruit et l'IDR est alors élevé. La couverture normalisée à permis d'attribuer un rang à chaque pic dans chaque réplicat. Plusieurs essais nous ont conduit à conserver les pics tel que $IDR < 5 \cdot 10^{-3}$ (figure 26).

Il convient d'émettre des réserves quant à cette procédure. Comme nous l'avons mentionné plus tôt, nous disposons parfois de 3 réplicats pour certaines expériences et les méthodes utilisées ici ne permettent d'en traiter que deux. En effet, nous ne sommes parvenus à aucune méthode satisfaisante prenant en compte les 3 réplicats. La simple question sur la méthode pour déterminer les pics communs aux 3 réplicats ne trouve pas de réponse simple si on souhaite imposer des critères restrictifs. Nous avons donc fait le choix de définir les pics significatifs à partir de deux réplicats seulement.

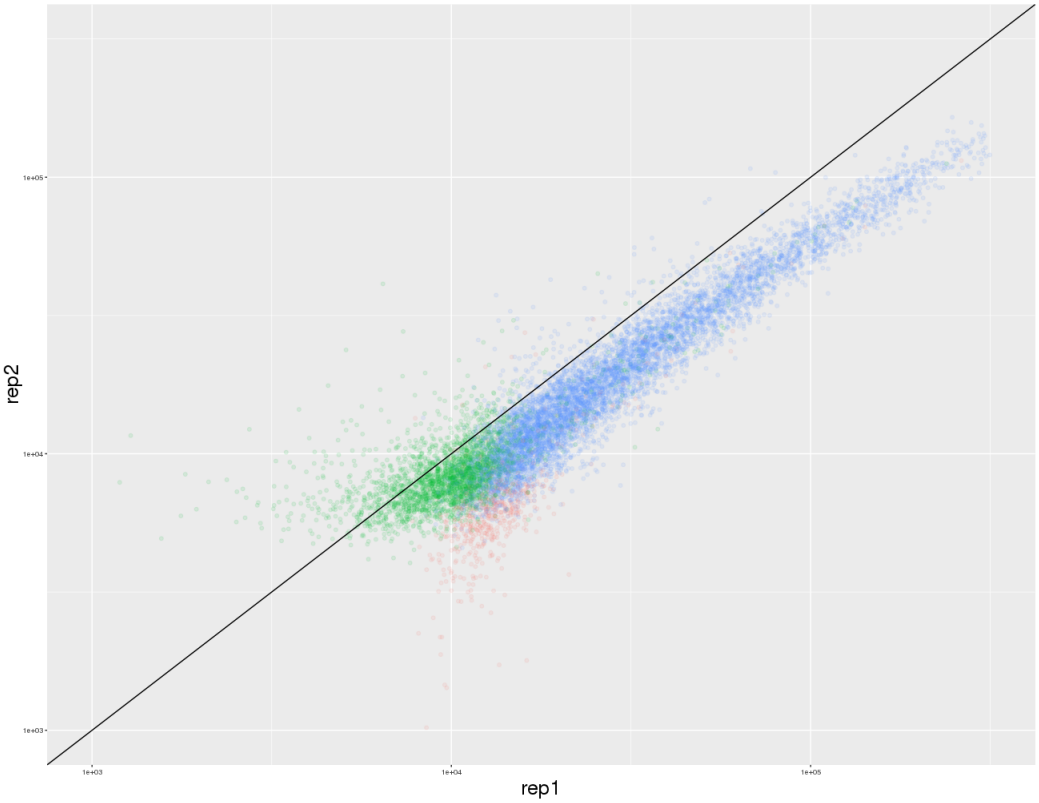
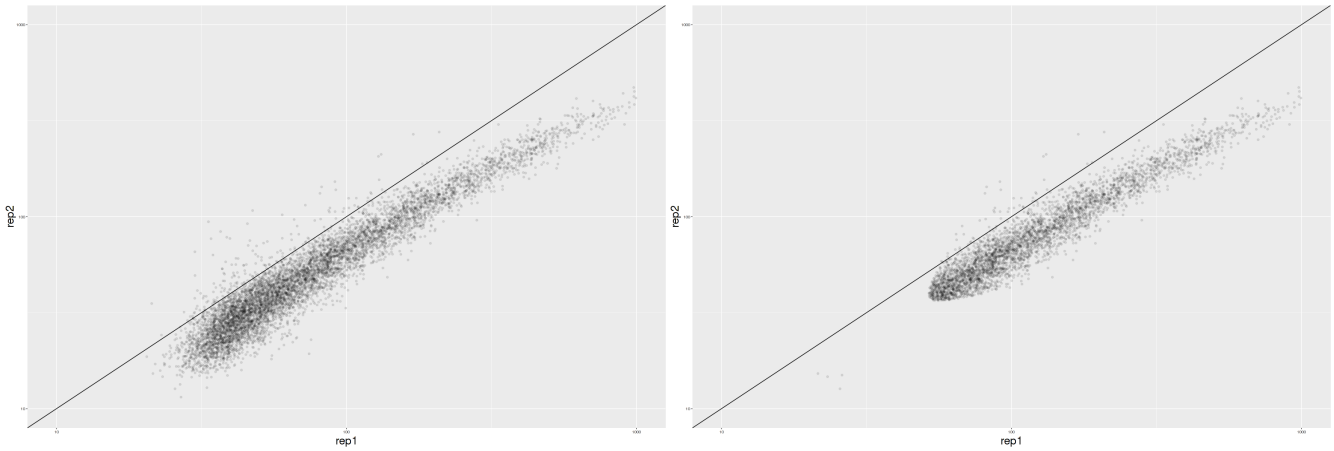


FIGURE 25: Chaque point correspond à un pic (commun aux deux réplicats en bleu ou appartenant à l'un ou à l'autre en rouge ou en vert). Les axes x et y donnent respectivement la couverture normalisée dans le premier et le second réplicat. Ici, les points s'alignent presque sur la diagonale, ce qui indique que les 2 réplicats sont presque identiques et exploitables.



(a) (b)

FIGURE 26: Pics communs dans 2 réplicats avant (A) et après (B) filtrage IDR

II.1.2.3 Fusionner les réplicats pour obtenir le signal sous les pics

Même si les pics sont déterminés à partir de deux réplicats, l'idée reste de n'en laisser aucun de côté. On choisit donc de concaténer les reads d'une même expérience. Cela est réalisé par la commande "macs2 callpeak −t rep1.bam rep2.bam rep3.bam −c control.bam −B" puisque macs2 donne un bedgraph normalisé grâce à l'option -B. La procédure est globalement la même que dans le paragraphe II.1.2.1 : il suffit ici de calculer la couverture normalisée sous chaque pic à partir du fichier bedgraph nouvellement obtenu.

II.1.3 Pré-traitement des données brutes de DAP-Seq en résumé

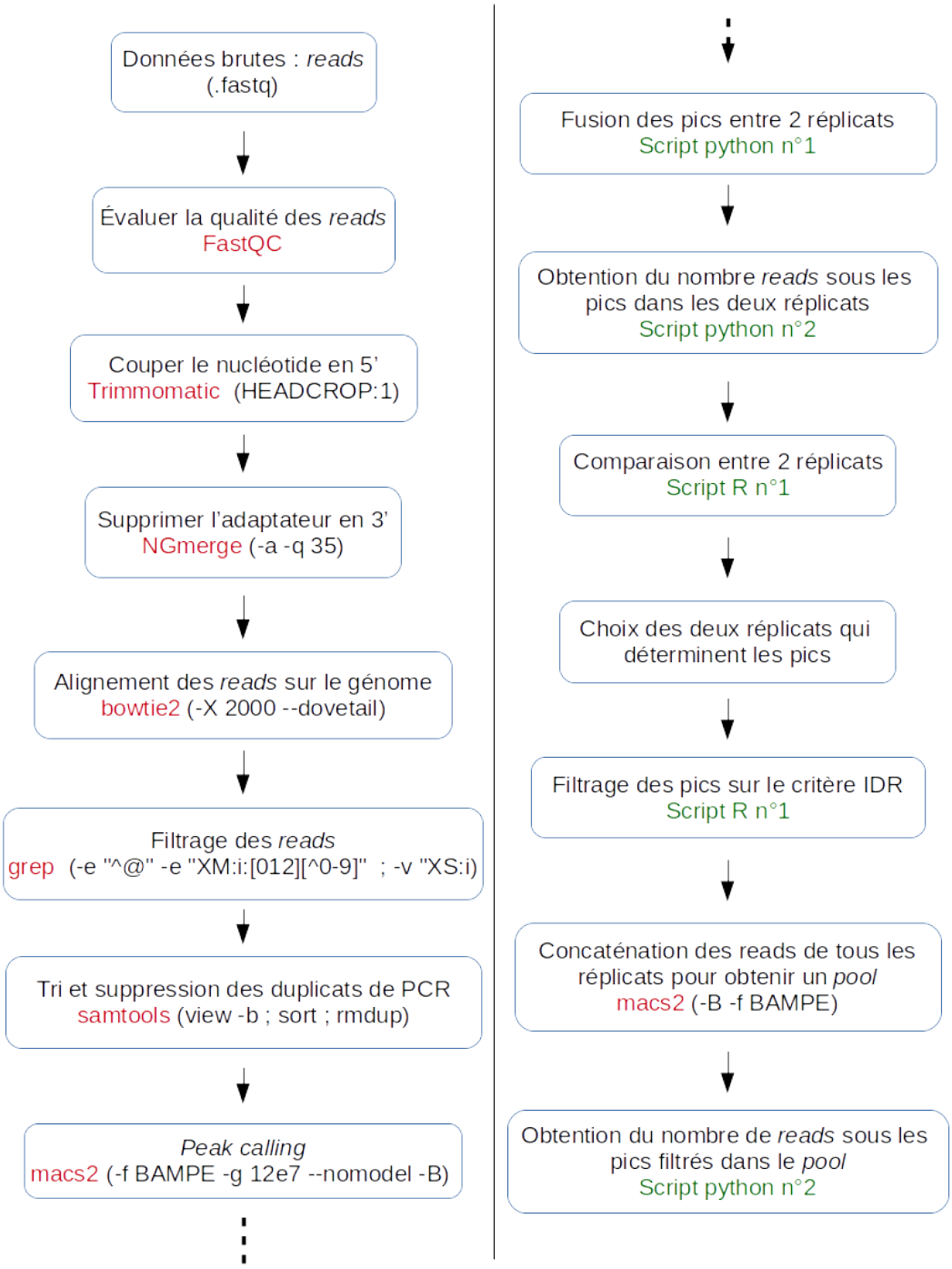


FIGURE 27: Récapitulatif des étapes pour traiter les données de DAP-Seq. En rouge, les programmes utilisés avec les différentes options entre parenthèses. En vert, les programmes que j'ai développés dans le cadre de cette procédure.

II.1.4 Discussions

II.1.4.1 À propos des *reads*

L'utilisation de *reads* de 150 paires de bases a d'abord semblé un choix intéressant pour obtenir un séquençage de bonne précision. Cependant, il est apparu que le contrôle de la taille des fragments séquencés n'est pas évident, et que la taille de ces fragments était souvent inférieure à 250 paires de bases. L'usage de long *reads* est alors inutile, d'autant que le séquençage coûte plus cher, que la mise au point des analyses prend plus de temps, et que la quantité d'erreurs humaines augmente potentiellement.

II.1.4.2 À propos de l'alignement

Nous n'avons pas effectué de comparaison entre différents programmes d'alignement. Il semble que bowtie2 et BWA (Li and Durbin, 2009) sont les plus utilisés dans la communauté et notre choix s'est naturellement tourné sur l'un des deux. Compte tenu du grand nombre de librairies que nous avons alignées ($\simeq$ 100) et du nombre d'essais effectués, la rapidité de bowtie2 est extrêmement satisfaisante. Après alignement, la profondeur de séquençage varie entre 20 et 30 suivant les échantillons. Le filtrage des *reads* décrit dans le paragraphe II.1.1.5 divise généralement ce nombre par 5. Étant donné la faible sélectivité de bowtie, il semble que le filtrage soit vraiment nécessaire.

II.1.4.3 À propos des pics

L'obtention correcte des régions liés par un TF est une étape déterminante dont le plus important chaînon est la détermination des pics. Le DAP-Seq étant une technique très récente, aucune étude comparative n'existe quant au choix du *peak caller* sur de tels jeux de données. Alors que O'Malley et al. (2016); Galli et al. (2018) utilisent le *peak caller* GEM (Guo et al., 2012), celui-ci introduit un biais. En effet, lors du *peak calling*, les pics sont centrés autour de séquences d'ADN fréquemment retrouvées dans ces derniers, ces séquences correspondant potentiellement au site de liaison du TF. Nous avons donc préféré macs2, très utilisé par la communauté sur tout type de jeux de données. La documentation exhaustive, l'investissement de la communauté et de l'auteur dans la correction des erreurs et dans l'aide apportée aux utilisateurs, le fait qu'il supporte les jeux de données pairées en font un outil indiqué.

Lors de son utilisation, nous avons remarqué que la fermeté du filtrage influence considérablement la sensibilité de macs2. En effet, si quelques pics aberrants associés à des *reads* de mauvaise qualité surpassent les pics aux positions liées par le TF, alors seuls ces pics aberrants sont retournés, le reste du signal étant considéré comme du bruit par macs2. Cela conforte l'importance de la mise au point d'un filtrage efficace. Un rapide examen des tailles de pics renvoyés par macs2 donne généralement une idée de la fiabilité des pics (figure 28).

Remarquons enfin que nous avons observé un *bug* lorsque plusieurs instances de macs2 sont lancées sur la même machine. Certaines instances s'arrêtent brutalement ce qui rend la recherche de pics dans plusieurs librairies difficile à paralléliser.

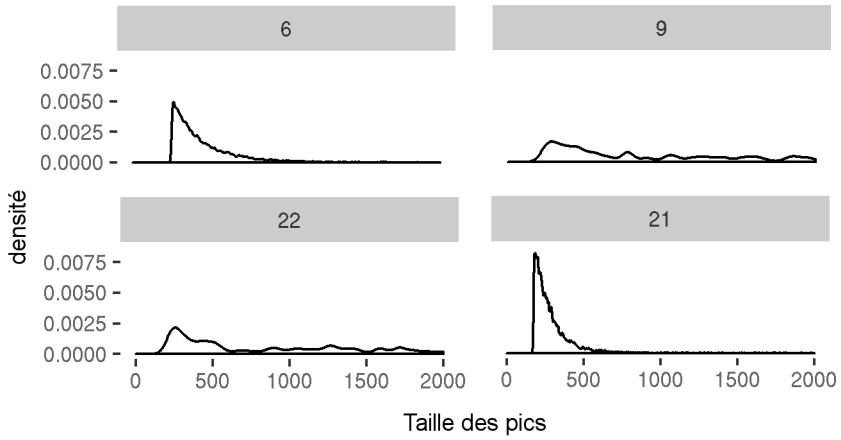


FIGURE 28: Les analyses ont montré que dans les cas où les pics sont fiables, leur taille est généralement comprise entre 200 et 800 paires de bases. Ainsi, les pics des librairies 9 et 22 ne sont pas exploitables.

II.2 Analyse des données brutes de ChIP-Seq

Cette section s'applique aux donnée de ChIP-Seq étudiées dans le chapitre 3.

Nous avons été amenés à traiter des données publiées dans différentes études. À la différence des données de DAP-Seq que nous avons étudiées, les expériences de ChIP-Seq ont été faites sur des *reads* non pairés. Cela engendre des différences mineures sur le paramétrage de l'alignement des *reads* sur le génome et de la recherche des pics.

Ainsi, l'alignement a été effectué à l'aide de bowtie 2 avec les paramètres par défauts. Le filtrage et le traitement des *reads* alignés est identique à celui effectué pour le DAP-Seq.

La recherche de pics a été faite avec macs2, l'option -f BAMPE étant remplacée par -f BAM (pour indiquer que les *reads* ne sont pas pairés) et l'option --nomodel ayant été supprimée. En effet, comme seules les extrémités des *reads* sont alignées, macs2 doit pouvoir estimer la taille de ces fragments afin de situer correctement les pics.

La suite des analyses (évaluation des réplicats, choix des pics, calcul de la couverture normalisée) est effectuée de la même manière que pour les données de DAP-Seq.

II.3 Analyse des sites de liaison

Cette section explique les méthodes utilisées à la fois dans les chapitres 1 (qui porte sur les ARF) et le chapitre 3 (qui porte sur les MADS).

II.3.1 Recherche de motifs

Plusieurs programmes permettent de trouver des motifs redondants à partir d'un jeu de séquences. Parmi eux, on peut citer RSAT (Nguyen et al., 2018) et meme-suite (Bailey et al., 2009). Ces deux suites de programmes peuvent être lancées sur des serveurs distants à partir de services web graphiques.

Néanmoins, la disponibilité du code source de meme-suite nous a fait choisir cette option afin que nous puissions l'installer localement sur notre serveur de calcul. Comme nous n'avons pas généré les jeux de données du chapitre 1, les méthodologies des chapitre 1 , 2 et 3 diffèrent légèrement.

II.3.1.1 Données DAP-Seq sur les ARF

Ce paragraphe s'applique au chapitre 1. Seuls les pics ont été publiés (O'Malley et al., 2016), sans mention de réplicats potentiels. Aucun contrôle qualité n'a donc été fait sur les données brutes. Chaque pic est donné avec une q-value qui permet d'évaluer sa qualité. À la fois parce que l'algorithme de meme [5] est plus efficace sur un petit nombre de séquence, et aussi parce que les pics avec une q-value élevée sont peu enrichis en site de liaison, nous avons sélectionné les 600 meilleurs pics d'après le critère de la q-value. La taille des pics n'importe pas étant donné que meme ne considère que les 100 bases centrales de la séquence. Afin d'obtenir des motifs comparables pour MP et pour ARF2, nous avons forcé leur taille à 10 paires de bases dans les options de meme. Aidé de la suite bedtools et du programme awk, disponible sous les suites linux, nous avons centré les motifs pour obtenir : NN**TGTC**NNNN.

II.3.1.2 Données DAP-Seq et ChIP-Seq sur LFY

Ce paragraphe s'applique au chapitre 2.

PWM Les pics ont été classés par couverture normalisée, nous avons utilisé le programme meme-chip (de la suite meme-suite) sur les 300 meilleurs pics pour déterminer les motifs. La taille des motifs a été fixée à 19 (options -meme-minw 19 -meme-maxw 19) et nous avons obtenu des motifs palindromiques grâce à l'option -pal.

Confronter les motifs à des bases de données Afin de déterminer la présence éventuelle de co-facteurs, nous avons élargi la taille de la fenêtre de recherche à 200 paires de bases (contre 100 par défaut) grâce à l'option -ccut 200. Nous avons demandé 10 motifs (option -meme-nmotif 10) et nous les avons confrontés à la base de données JASPAR Plants (Khan et al., 2017) disponible à l'adresse `http://jaspar.genereg.net/download/CORE/JASPAR2018_CORE_plants_non-redundant_pfms_meme.zip` grâce à l'option -db, qui lance le programme tomtom (Gupta et al., 2007).

TFFM Les TFFM ont été entraînées à partir des PWM obtenues et des 300 pics utilisés grâce au paquet disponible sur le site `https://github.com/wassermanlab/TFFM` (Mathelier and Wasserman, 2013).

II.3.1.3 Données DAP-Seq sur les gènes à boîte MADS

Ce paragraphe s'applique au chapitre 3.

PWM Les pics ont été traités comme décrit dans le paragraphe II.1.2.2. Classés par couverture normalisée, nous avons utilisé le programme meme-chip sur les 300 meilleurs pics pour déterminer les motifs. La taille des motifs a été fixée à 16 (options -meme-minw 16 -meme-maxw 16) et nous avons obtenu des motifs palindromiques grâce à l'option -pal.

5. programme de la suite meme-suite qui cherche des motifs redondants

TFFM Les TFFM ont étés entraînées à partir des PWM obtenues et des 300 pics utilisés grâce au paquet disponible sur le site `https://github.com/wassermanlab/TFFM` (Mathelier and Wasserman, 2013).

II.3.2 Contrôle des motifs

Les motifs, déterminés à partir des meilleurs pics, sont ensuite testés sur la totalité des régions liées. L'idée est d'évaluer le pouvoir discriminant de la PWM entre les régions liées (déterminées par DAP-Seq) et des régions non liées.

Le choix des régions non liées demande une attention particulière et nos essais ont montré que celui-ci affecte énormément les performances de la PWM. L'outil BiasAway (Hunt et al., 2014) permet de créer des région non liées respectant le contenu nucléotidique ou di-nucléotidique des régions liées semble une alternative acceptable. Cet outil permet également de créer des sets de régions *génomiques* dont le contenu nucléotidique est proche à celui des régions liées. Nous avons souhaité pousser les exigences plus loin en demandant une origine identique : pour toute région liée qui est un promoteur, un intron, un exon, ou un intergénique, nous demandons une région non liée dont les caractéristiques sont similaires. Ainsi, nous avons écrit notre propre programme en python qui à chaque région liée associe une région non liée de contenu nucléotidique et d'origine identique.

Les régions liées et non liées sont ensuite respectivement scannées par la PWM/TFFM grâce au paquet R Bioconductor-Biostrings ou le package TFFM (`https://github.com/wassermanlab/TFFM`) (Mathelier and Wasserman, 2013), le meilleur score de chaque région étant retenu. On évalue ensuite le pouvoir discriminant du modèle entre les régions liées et celles non-liées grâce au critère de l'aire sous la courbe *Receiver operating characteristics* (AUROC), tracé et calculé grâce au paquets python pROC et matplotlib.

II.3.3 Calcul des espacements entre les sites de liaison

Nous avons écrit un programme python accessible à l'adresse `https://github.com/Bioinfo-LPCV-RDF/get_interdistances`.

L'idée est la suivante : les régions liées – (pos) dans les calculs – sont scannées avec une matrice (TFFM ou PWM). Appelons $N(pos)$ le nombre de sites total, qui correspond à $N(pos) = \sum_{k=0}^{j}(l_k - lm + 1)$ où l_k vaut la taille de la séquence k, lm la taille de la matrice et j le nombre de séquences. Un seuil est choisi si bien que seuls les sites dont le score est au dessus du seuil sont conservés.

Étant donné ce seuil, on calcule les effectifs $N_{C_i}(pos)$ pour chaque configuration $C_i(pos) \in \{DR_n(pos) \cup ER_n(pos) \cup IR_n(pos), n \in [0, S_{max}]\}$ (définies dans l'introduction du chapitre 1) où S_{max} est l'espacement maximal entre 2 sites de liaison considérés.

La démarche est de reproduire les mêmes calculs dans un set de régions non liées – (neg) dans les calculs – et de calculer l'enrichissement E_{C_i} (formule 7) pour chaque C_i.

$$E_{C_i} = \frac{N_{C_i}(pos)N(neg)}{N_{C_i}(neg)N(pos)} \tag{7}$$

Ici, nous avons été confronté à un problème. En effet, le nombre de bons sites de liaison est plus grand dans le set de régions liées que dans le set de régions non liées. La conséquence est que pour tout $i \in [1, n]$, $E_i > 1$. Cela donne l'impression que la protéine d'intérêt a une bonne affinité pour chaque configuration. Pour pallier ce biais, on calcule le nombre total de configurations pour le seuil considéré dans le set de régions liées $N_{C_{tot}}(pos)$ et dans le set de régions non liées $N_{C_{tot}}(neg)$. Ainsi, on obtient

l'enrichissement normalisé NE_{C_i} (formule 8) (*Normalized Enrichment* dans les figures), dont la moyenne sur toutes les configurations vaut 1.

$$NE_{C_i} = \frac{N_{C_i}(pos)N(neg)N_{C_{tot}}(pos)}{N_{C_i}(neg)N(pos)N_{C_{tot}}(neg)} \tag{8}$$

Néanmoins, nous avons souhaité garder l'information donnée par l'enrichissement E_{C_i}, qui permet de savoir si pour un seuil donné, le nombre de configurations est plus important dans les régions liées que dans les régions non liées. Nous avons donc décidé de résumer chaque valeur E_i en sommant tous les effectifs d'un même type de configuration (*ie.* (DR, ER, IR)). L'enrichissement absolu AE (ou *Absolute Enrichment* dans les figures) se calcule de la manière suivante :

$$AE_{C=DR,ER,IR} = \frac{\sum_{i=0}^{S_{max}} N_{C_i}(pos)N(neg)}{\sum_{i=0}^{S_{max}} N_{C_i}(neg)N(pos)} \tag{9}$$

II.4 Discussion sur les choix programmation

À l'issue de ma thèse, à l'exception des notions scientifiques, je pense avoir été confronté à 3 problèmes de natures différentes.

Le premier concerne la représentation des données. Il s'agit souvent de l'étape finale et la force des messages aussi bien que leur clarté passent par des représentations graphiques soignées et réfléchies. Les langages R et python offrent tous deux les moyens pour produire des graphiques d'excellentes qualité grâce aux librairies respectives ggplot2 et matplotlib. Les tableaux de données peuvent être aussi bien traités par l'environnement natif de R que grâce aux librairies panda et numpy sous python. De mon point de vue, la documentation des libraries panda et matplotlib sous python est assez naissante (et horriblement compliquée pour matplotlib) et si elle se complète chaque jour, l'abondance d'informations pour R et ggplot2 a généralement conduit à choisir cette option. On peut ajouter que R est en lui-même un programme d'analyse de données et qu'il intègre naturellement de nombreuses fonctions nécessaires à ces analyses. Python nécessite l'importation de nombreuses librairies, ce qui rend l'utilisation moins intuitive, chaque librairie ayant ses mécaniques différentes.

Le second concerne le traitement des fichiers textes, omniprésents en bioinformatique. Il peut s'agir de remplacer certains caractères, d'extraire certaines lignes ou certaines colonnes ou de reconnaître des expressions rationnelles. Les programmes intégrés aux distributions GNU/Linux réalisent ces opérations avec une efficacité inégalable. On peut citer sed, grep, awk, cat, head, tail, wc, ou paste. Transformer un fichier se fait alors en une ligne de commande et je pense que la bioinformatique passe par une maîtrise élémentaire de ces programmes.

Enfin, le dernier problème s'applique aux opérations plus compliquées, lorsqu'il faut écrire un programme car aucun existant ne réalise une tâche précise. Le langage python, très utilisé dans la communauté, est alors indiqué.

On pourrait ajouter un dernier obstacle, qui est la gestion de tous les programmes utilisés. La bonne connaissance d'un *shell* (comme le bash) et de l'informatique sous GNU/Linux est à mon avis indispensable. Par exemple, être capable de gérer des programmes lancés en parallèle permet des gains de temps considérables. Mais plus simplement, la gestion des chemins, des librairies et des variables d'environnement peut rendre l'installation de certains programmes difficile.

Chapitre 1

Syntaxe des facteurs de réponses à l'auxine

1.1 Introduction

Si l'auxine et son rôle important ont été mis en évidence dès le début du XX^{e} siècle, il faut attendre une centaine d'années avant que ses premiers gènes cibles soient identifiés. L'étude de *GRETCHEN HAGEN 3* (*GH3*), permet de mettre en évidence les premiers éléments de la voie de signalisation nucléaire par l'auxine. En altérant le promoteur du gène, Liu et al. (1994) isolent la partie qui rend son niveau d'expression contrôlé par l'auxine. Ces éléments sont les *Auxin Response Elements* (AuxRE) (évoqués dans le I.4.1.2), définis comme la séquence TGTCTC. Le premier ARF (ARF1) est identifié sur une construction contenant plusieurs AuxRE (Ulmasov et al., 1997a). D'autres membres de la famille sont rapidement identifiés, les recherches s'essayant alors à expliquer les différentes réponses à l'auxine en fonction des spécificités des ARF.

Ainsi, des études qui suivent peu après montrent que les ARF1-10 sont tous capables de lier la séquence TGTCNN avec de effets plus ou moins importants suivant les variations sur les nucléotides NN (Ulmasov et al., 1999). Pour permettre aux ARF de former des dimères, la configuration de deux AuxRE (figure 29) et leur espacement sur un même promoteur jouent également un rôle important : ARF1 lie les *Everted Repeat* (ER) avec un espacement de 7 ou 8 (ER7/8) et les *Direct Repeat* (DR) avec un espacement de 5 (DR5) de façon privilégiée. Ces deux types d'éléments peuvent conférer une inductibilié par l'auxine (Ulmasov et al., 1997a,b) lorsqu'elles sont présentes dans un promoteur. Dans des expérience effectuées sur des protoplastes [1], les ARF5-8 sont capables d'induire des promoteurs possédant des éléments de type DR5 et ER7/8 alors que ARF3,4,9 répriment les gènes dont le promoteur contient ER7 (Tiwari et al., 2003). Cela tend à montrer que des combinaisons d'ARF et de configurations d'AuxRE induisent des réponses différentes.

1. Cellules végétales débarrassées de leur paroi cellulosique externe

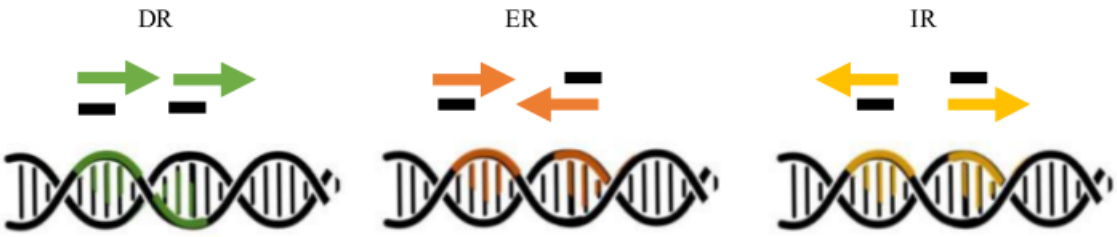


FIGURE 29: Deux AuxRE peuvent soit former un *Direct Repeat* (DR), un *Everted Repeat* (ER) ou un *Inverted Repeat* (IR). D'après Martin-Arevalillo (2017).

Pour expliquer les différences de spécificité, Boer et al. (2014) proposent le modèle du "pied à coulisse" (*caliper* en anglais). Les ARF auraient seraient tous capables de lier les mêmes configurations de dimère, mais alors que certaines protéines lieraient des espacement strictement définis, d'autres seraient plus permissives. Ainsi, ARF1 lie strictement ER7/8 alors que ARF5/MP montre une bonne affinité pour ER5-9. Cependant, ce modèle n'a été établi qu'en utilisant les domaines de liaison de ARF1 et ARF5 et non pas les protéines entières.

À ce jour, si certains éléments liés à la voie de signalisation nucléaire par l'auxine ont été capturés, la compréhension des règles de liaison et de régulation demeure incomplète. Ainsi, Franco-Zorrilla et al. (2014) utilisent le PBM (expliqué dans la revue en annexe) pour montrer que les ARF lient préférentiellement TGTCGG plutôt que TGTCTC et Zemlyanskaya et al. (2016) proposent un modèle où la régulation positive ou négative des gènes par l'auxine dépend des bases NN du motif TGTCNN mais des modèles de liaison plus évolués, comme les PWM et les TFFM, n'ont pas été appliqués pour une meilleure compréhension de ces règles.

Dans leur article, O'Malley et al. (2016) proposent une étude des spécificités dimériques de deux ARF (ARF2 et ARF5/MP) grâce au DAP-Seq. La recherche de sites de liaisons monomériques est alors basée sur la séquence consensus TGTC. Considérant que cette méthode est insuffisante pour capturer le comportement des protéines, nous avons fait le choix de reprendre et compléter une telle analyse en remplaçant le consensus par les matrices poids position. Nous nous intéresserons d'abord aux spécificités de liaison de ces deux TF et nous confronterons ensuite ces spécificités à l'activité des gènes régulés par l'auxine.

1.2 Article

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

Molecular Plant
Research Article

CellPress
PARTNER JOURNAL

Capturing Auxin Response Factors Syntax Using DNA Binding Models

Arnaud Stigliani[1], Raquel Martin-Arevalillo[1,2], Jérémy Lucas[1], Adrien Bessy[1], Thomas Vinos-Poyo[1], Victoria Mironova[3,4], Teva Vernoux[2], Renaud Dumas[1] and François Parcy[1,*]

[1]Univ. Grenoble Alpes, CNRS, CEA, INRA, BIG-LPCV, 38000 Grenoble, France

[2]Laboratoire de Reproduction et Développement des Plantes, Univ. Lyon, ENS de Lyon, UCB Lyon1, CNRS, INRA, 46 allée d'Italie, 69364, Lyon, France

[3]Novosibirsk State University, Pirogova Street 2, Novosibirsk, Russia

[4]Institute of Cytology and Genetics SB RAS, Lavrentyeva Avenue 10, Novosibirsk, Russia

***Correspondence: François Parcy (francois.parcy@cea.fr)**

https://doi.org/10.1016/j.molp.2018.09.010

ABSTRACT

Auxin is a key hormone performing a wealth of functions throughout the life cycle of plants. It acts largely by regulating genes at the transcriptional level through a family of transcription factors called auxin response factors (ARFs). Even though all ARF monomers analyzed so far bind a similar DNA sequence, there is evidence that ARFs differ in their target genomic regions and regulated genes. Here, we report the use of position weight matrices (PWMs) to model ARF DNA binding specificity based on published DNA affinity purification sequencing (DAP-seq) data. We found that the genome binding of two ARFs (ARF2 and ARF5/Monopteros [MP]) differ largely because these two factors have different preferred ARF binding site (ARFbs) arrangements (orientation and spacing). We illustrated why PWMs are more versatile to reliably identify ARFbs than the widely used consensus sequences and demonstrated their power with biochemical experiments in the identification of the regulatory regions of *IAA19*, an well-characterized auxin-responsive gene. Finally, we combined gene regulation by auxin with ARF-bound regions and identified specific ARFbs configurations that are over-represented in auxin-upregulated genes, thus deciphering the ARFbs syntax functional for regulation. Our study provides a general method to exploit the potential of genome-wide DNA binding assays and to decode gene regulation.

Key words: *Auxin*, Auxin Response Factor, DAP-seq, DNA binding model

Stigliani A., Martin-Arevalillo R., Lucas J., Bessy A., Vinos-Poyo T., Mironova V., Vernoux T., Dumas R., and Parcy F. (2019). Capturing Auxin Response Factors Syntax Using DNA Binding Models. Mol. Plant. ■ ■, 1–11.

INTRODUCTION

Auxin is a key plant hormone affecting multiple developmental processes throughout the lifecycle of the plant. Most long-term developmental auxin responses (such as embryo polarity establishment, tropisms, phyllotaxis, or secondary root emergence) involve modifications of gene expression by the nuclear auxin pathway (Lavy and Estelle, 2016; Weijers and Wagner, 2016). This pathway includes a family of transcription factors (TFs) called auxin response factors (ARFs) (Weijers and Wagner, 2016; Leyser, 2018). In the absence of auxin, the Aux/IAA repressors bind ARF TFs and form inactive multimers thereby preventing their activity (Han et al., 2014; Korasick et al., 2015). The presence of auxin leads to the degradation of Aux/IAA proteins and therefore allows ARFs to activate transcription.

ARF proteins exist in three classes (A, B, and C) with class A corresponding to ARF activators and B and C to ARF repressors (Finet et al., 2013). Understanding ARF biochemical properties (DNA binding specificity, capacity to activate or repress transcription, capacity to interact with partners) is important to decipher how different tissues could respond differently to the same auxin signal (Leyser, 2018). ARFs are modular proteins with several functional domains: most ARFs (except ARF3/ETTIN, ARF17, and ARF23) have a PB1 domain (previously called domain III/IV) responsible for interaction with the Aux/IAA repressors, TFs from other families, and possible homo-oligomerization through electrostatic head-to-tail assembly

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

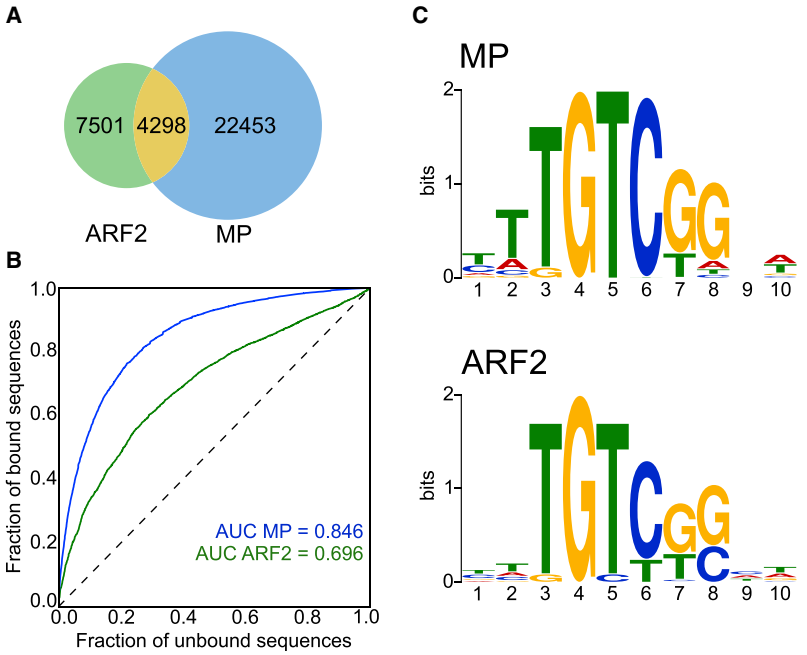


Figure 1. Global Analysis of DNA regions and DNA Motifs Bound by MP and ARF2 in DAP-seq

(A) Venn diagram of regions bound by ARF2 or MP in DAP-seq.

(B) ROC curves and AUC values for MP and ARF2 PWM models.

(C) Logo for MP and ARF2 PWMs.

(Korasick et al., 2014; Nanao et al., 2014; Parcy et al., 2016; Weijers and Wagner, 2016; Mironova et al., 2017). ARFs also possess a DNA binding domain (DBD) from the plant-specific B3 family. The structure of this DBD has been solved for ARF5 (also called Monopteros [MP]) and ARF1 revealing a B3 domain embedded within a flanking domain (FD) and a dimerization domain (DD) (Boer et al., 2014). The DD allows ARF proteins to interact as a face-to-face dimer with a DNA element called an everted repeat (ER) made of two ARF binding sites (ARFbs). ARFbs were originally defined as TGTCTC (Ulmasov et al., 1995; Guilfoyle et al., 1998), and this knowledge was used to construct a widely used auxin transcriptional reporter, DR5 (Ulmasov et al., 1997). More recently, protein binding microarray (PBM) experiments suggested that TGTCGG are preferred ARFbs, and a new version of DR5, DR5v2, was built based on this *cis* element (Boer et al., 2014; Franco-Zorrilla et al., 2014; Liao et al., 2015). ARFs are able to bind different ARFbs configurations in addition to ER such as direct repeats (DR) or, as recently suggested, inverted repeats (IR) (O'Malley et al., 2016). Whether ARF oligomerization through the PB1 domain contributes to binding of some ARFbs configurations such as IR or DR that are not compatible with DD dimerization has been proposed but not yet demonstrated (O'Malley et al., 2016; Parcy et al., 2016). Based on affinity measurements of interaction between ARF DBD (for ARF1 and MP) and a few ER *cis* elements, it was proposed that ARFs differ by the type of ER configuration they prefer: the ARF1 repressor has a much narrower range of preferences than the MP activator (this was called the molecular caliper model) (Boer et al., 2014). However, this model was established using isolated ARF DBD lacking the PB1 domain and did not include interaction with DR and IR ARFbs configurations.

Despite the central importance of ARF TFs, models reliably predicting their DNA binding specificity are still scarce (Keilwagen et al., 2011; Mironova et al., 2014), and simple consensus sequences are often used (Berendzen et al., 2012; O'Malley et al., 2016; Zemlyanskaya et al., 2016) that hardly capture possible sequence variation within the *cis* element. Recently, DNA affinity purification sequencing (DAP-seq) data have offered a genome-wide view for two full-length ARF proteins of *Arabidopsis* (the repressor ARF2 and the activator MP) (O'Malley et al., 2016). The DAP-seq assay is technically similar to ChIP-seq but with chromatin-free isolated genomic DNA and with a single recombinant protein added. Based on a TGTC consensus sequence as the ARFbs definition, the ARF5/MP activator and the ARF2 repressor appear to have different preferred DNA binding sites. They appear to share a novel inverted repeat (IR7-8) element, but also have specific binding sites with different spacing and orientation of ARFbs (O'Malley et al., 2016). Here, we undertook an extensive reanalysis of DAP-seq data using position weight matrix (PWM) as the DNA binding specificity model (Wasserman and Sandelin, 2004). PWM represents a simple but efficient tool that captures the base preference at each position of the motif. PWMs give a score to any DNA sequence with zero for the optimal sequence and more negative scores as the sequence diverges from the optimum. The PWM score is then a quantitative value directly related to the affinity of the DNA molecule for the protein (Berg and von Hippel, 1987). Using PWMs, we establish differences between ARF2 and MP and show that they reliably identify a binding site syntax explaining their specificity. We further illustrate the predictivity of PWM compared with consensus using binding assays and identify ARFbs configurations enriched in promoters of genes regulated by auxin.

RESULTS

ARF2 and MP DNA Binding Site Similarities and Genomic Locations

Using the published DAP-seq data (O'Malley et al., 2016), we first compared the sets of genomic regions bound by ARF2 and MP. Two DAP-seq regions were considered bound by both factors when they overlapped by at least 50% (see Methods). As expected for two TFs from the same family, there is a significant overlap and many regions are bound by both factors (Figure 1A). However, the large number of regions specifically bound by only one of them indicates a clear difference between ARF2 and MP DNA binding preferences (Figure 1A). This remains true even when focusing on regions bound with the highest confidence (top 10%, see Methods) by each of the factors (Supplemental Figure 1). We intended to

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

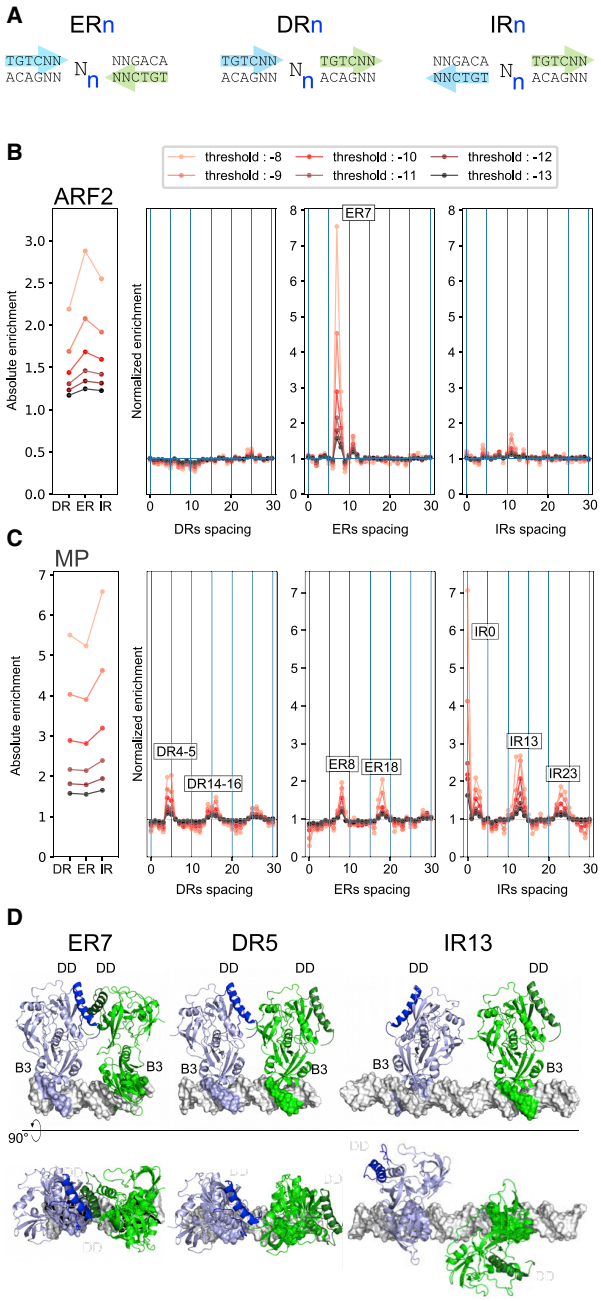


Figure 2. ARFbs Configuration Enrichment

(A) Definition of ERn, DRn, and IRn.

(B and C) Over-representation of dimeric ARFbs configurations in DAP-seq regions compared with an unbound set of sequences generated for ARF2 **(B)** and MP **(C)**. The left panels quantify the absolute enrichment for all ERn, DRn, and IRn ($0 \leq n \leq 30$) compared with the background set. Right panels present the normalized enrichment for each ERn, DRn, and IRn (see Methods).

(D) Structural modeling of DR5 and IR13 ARF complexes based on ER7 ARF1 structure (PDB: 4LDX) (Boer et al., 2014). Note that the dimerization interface present on ER7 is absent in the other two configurations.

explain these differences by characterizing ARF2 and MP DNA binding specificity. The examination of the DNA motif logo derived from regions recognized by ARF2 or MP monomers revealed only minor differences (Figure 1C). For both logos, the G[4] position corresponding to a direct protein-base contact in the ARF1 structure (Boer et al., 2014) is highly invariant. At positions [7,8] where the original ARFbs harbored TC (Guilfoyle et al., 1998), the preferred sequence is GG as recently proposed from PBM experiments (Boer et al., 2014; Franco-Zorrilla et al., 2014), but this preference is not as pronounced as in PBM-derived logos, and sequence variations at these positions are tolerated.

We built ARF2 and MP PWMs to model their DNA binding. We evaluated the prediction power of each PWMs using receiver operating characteristics area under the curve analysis (ROC-AUC or AUROC) (Hanley and McNeil, 1982) based on the best-score ARFbs present in each bound region (Figure 1B). Such analysis yields an AUROC value of 1 for a perfect model and 0.5 for a model with no predictive value. This analysis requires the generation of a negative set of regions for comparison. For this, we improved a previously designed tool, a negative set builder (Sayou et al., 2016), to extract from the *Arabidopsis* genome a set of non-bound regions with similar features as bound ones (size, GC content, genomic origin; see Methods). Based either on the full set of bound regions (Figure 1B) or only the 10% top ranked regions (Supplemental Figure 1), we found that the MP model is highly predictive (AUROC = 0.84) while the ARF2 model has a lower performance (AUROC = 0.69).

PWM models assume additive contributions of each nucleotide position, a hypothesis that is not always true (Bulyk, 2002; Moyroud et al., 2011; Zhao et al., 2012; Mathelier and Wasserman, 2013). We used Enologos (Workman et al., 2005) to test for the presence of dependencies between some of the positions, particularly for positions [7,8] (Figure 1C) where mostly GG and TC doublets have been proposed so far. Enologos did not detect any dependency (Supplemental Figure 1) indicating that standard PWM can be adequately used. We also wondered whether the small differences between ARF2 and MP PWMs (as visible on their logos from Figure 1C) could contribute to their binding specificity. We thus tested the MP PWM on ARF2 regions and, conversely, ARF2 PWM on MP regions. The performance is indeed slightly weaker showing there is some specificity in the monomer PWM (Supplemental Figure 1). However, the very small difference suggests there must be other parameters explaining ARF2 and MP different specificities.

Comparisons of ARF2 and MP Binding Site Configurations

Published analyses (O'Malley et al., 2016) suggested that MP and ARF2 might differ in their preferred ARFbs dimeric configurations (ER, DR, or IR; Figure 2A). We thus analyzed the distribution of spacings between ARFbs using PWM models. To do this, a score threshold needs to be chosen above which a transcription factor binding site (TFBS) is considered. As this threshold cannot be experimentally determined, we performed the analysis within a range of scores (from −8 to −13, where

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

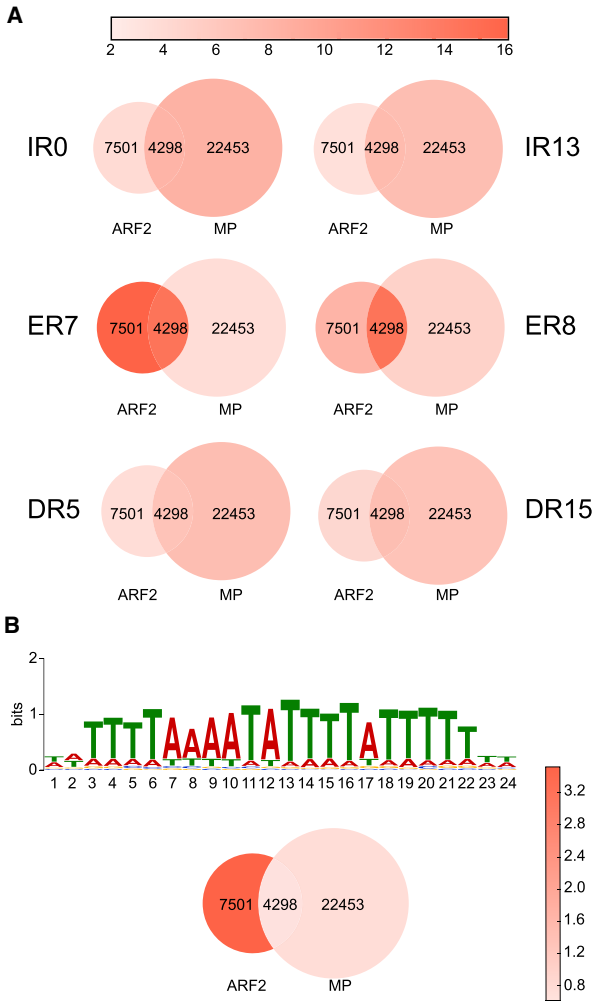


Figure 3. Example of ARFbs Configurations Frequency in ARF2- and MP-Bound Regions

(A) Venn diagrams colored according to the frequency (in %) of a few ARFbs conformations in MP-specific, ARF2-specific, and MP/ARF2 common regions.

(B) Fraction of regions containing at least one AT-rich motif in MP-specific, ARF2-specific, and MP/ARF2 common regions. AT-rich motifs are enriched in ARF2-specific regions as described in the Methods.

−8 indicates better affinity than −13). We studied the over-representations of all dimer configurations (DR, ER, and IR) compared with a negative set of regions. Overall, DR, ER, and IR are more frequent in the ARF-bound regions than in the negative set (Figure 2B, left panels), consistent with the higher density of ARFbs in these regions. We next estimated the over-representation of each particular configuration (ERn, DRn, or IRn with the spacing n varying between 0 and 30 bp) within the whole population of configurations and normalized it to the equivalent parameter in the negative set of regions (Figure 2). For example, if, for a given value of n, DRn represents 10% of all configurations (ER/DR/IR with $0 \leq n \leq 30$) in the positive set and only 2% in the negative one, DRn enrichment will be 5-fold.

This analysis revealed a striking difference between ARF2 and MP. For ARF2, ER7-8 are the only over-represented configurations, whereas MP showed a wider range of preferred distances and configurations, including DR4-5, DR14-16, DR25-26, ER7-8, ER17-18, IR0, IR3, IR12-13, IR23-24 (Figure 2B). Our results contrast with O'Malley's where IR8 was the most over-represented configuration for both factors (O'Malley et al., 2016). In this study, the authors chose NNGACA as the ARFbs definition (the reverse complement of TGTCNN) and the conventional IR and ER definitions (Nelson et al., 1996), opposite of what has been used until now in the auxin field. Because of these two differences, O'Malley's and our definition of ER are equivalent and should have yielded the same enrichments. Since their result was obtained using a TGTC consensus as ARFbs definition, we repeated our analysis with TGTC (Supplemental Figure 2A). We still validated our result, suggesting there might be a confusion at some point in O'Malley et al.'s analysis. The MP graph (Figure 2B) suggests a periodicity of over-represented distances every 10 bp, a hypothesis we confirmed by extending the distance window, revealing this trend for MP but not for ARF2 (Supplemental Figure 2B). Modeling of DR5 and IR13 protein/DNA complex structures based on ARF1 crystallographic data (PDB: 4LDX) clearly illustrates that these configurations are incompatible with the dimerization mode described for ER7 and could involve a different dimerization interface (Figure 2D).

ARF2 and MP DNA Binding Syntax

We re-examined the Venn diagram from Figure 1A in the light of the identified preferred configurations. We separated ARF2- and MP-bound regions in three sets: ARF2 specific, MP specific, ARF2/MP common regions. Because the two PWMs are very similar, we used the ARF2 matrix and performed the same analysis as in Figure 2 but on the three sets of regions (Supplemental Figure 3). DR4/5/15 and IR0/13 are over-represented only in MP-specific regions, ER7 in ARF2-specific regions, and ER7/8 mostly in the MP/ARF2 common regions. Remarkably, MP-specific regions are even depleted in ER7/8 compared with the negative set of sequences, because these elements are bound by both ARF2 and MP (Supplemental Figure 3). Plotting the frequency of a few selected configurations illustrates the group-specific characteristics (Figure 3). We also used RSAT (Medina-Rivera et al., 2015) to search for other sequence features that could distinguish the three groups of regions. For ARF2-bound regions only, we found an enrichment for nine long AT-rich motifs similar to the one shown in Figure 3B. These motifs are found all along the bound regions (not shown). One example of enrichment of such a motif is illustrated in Figure 3B.

Comparison between Improved PWM Models and Consensus

We incorporated the ARF2- and MP-specific features in new PWM-based models and tested their prediction power using AUROC. The improvement is marginal for MP but better for ARF2 (Figure 4C; AUROC for monomeric ARF2bs = 0.69, for ER7/ER8 model = 0.74). To illustrate the fundamental differences between PWM and consensus, we plotted the specificity (false-positive rate) and sensitivity (true-positive rate) parameters on the PWM ROC curve (Figure 4). For the monomeric ARFbs models, the TGTC consensus is poorly

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

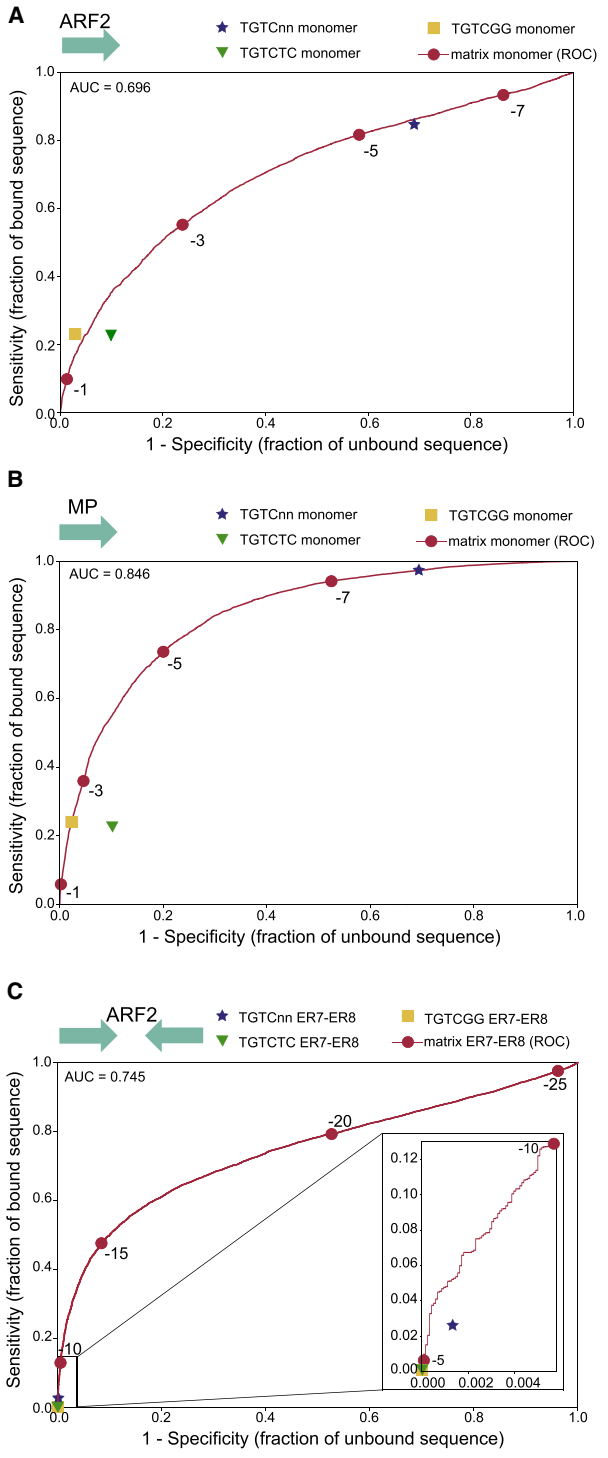


Figure 4. Comparison between PWM and Consensus Sensitivity and Specificity

For all graphs, red dots correspond to score thresholds used to plot the PWM ROC curves. For the consensus search, a sequence is considered positive for TGTC, TGTCGG, or TGTCTC if this sequence is present at least once in the DNA region. The ER7-8 models were built as described in the Methods.

(A) ARF2 PWM and consensus on ARF2-bound regions.

(B) MP PWM and consensus on MP-bound regions.

(C) ER7-8 PWM and consensus models on ARF2-bound regions.

specific with almost 70% false-positive rate. Conversely, TGTCGG or TGTCTC perform correctly but leave no freedom in terms of sensitivity and specificity; only the quantitative model allows these parameters to be chosen by adjusting the score threshold. For ARF2 ER7/8 dimeric models, using any of the three consensus is extremely stringent and detects very few sites in the positive set (at best 2.5% for TGTC), whereas the PWM model is again more versatile as it allows the desired specificity/sensitivity combination to be reached by adjusting the score threshold.

DNA binding models are extremely useful to detect TFBSs and challenge their role *in vitro* or *in vivo*. To scan individual sequences, PWMs are superior as they provide quantitative information linked to TFBS affinity (Berg and von Hippel, 1987) and allow the detection of possible non-consensus sites of high affinity. We used our models to identify binding sites in the well-studied promoter of the *IAA19* gene (Pierre-Jerome et al., 2016 and references therein). Scanning the *IAA19* promoter sequence with ARF2 and MP PWMs identified several ARFbs (Figure 5), including a high-scoring ER8 site bearing one non-canonical gGTCGG that lacks the TGTC consensus (Figure 5A). This site is located at the center of a DAP-seq peak for MP and ARF2. We tested ARF binding to this particular ER8 element and tested the impact of the consensus presence on binding to ARF. For this, we restored the TGTC consensus for this non-canonical ER8 element and also created an artificial ER8 that has both TGTC consensus but suboptimal bases in other positions according to the PWM (Figure 5B). Strikingly, the PWM score better predicts the binding than the presence of the consensus sequence; we observed intense binding on the non-canonical ER8, only a slight improvement when the consensus is restored, and no binding on a consensus-bearing ER8 of low score (Figure 5C).

PWM Models Reveal Preferred ARFbs Configurations in Auxin-Regulated Genes

We next tested the PWM models on *in vivo* data. ChIP-seq data are available for ARF6 and ARF3 (Oh et al., 2014; Simonini et al., 2017). However, no obvious ARFbs could be identified in any of these datasets. Testing ARFbs monomeric or dimeric models yielded a very poor AUROC value (0.61 for ARF6 and 0.58 for ARF3), suggesting that these data might not be adequate to evaluate our model. We also used the auxin-responsive gene datasets derived from a meta-analysis of 22 microarray data (see Methods, Supplemental Table 2). We defined four groups of regions of either auxin-induced or -repressed genes with high or very high confidence (very high confidence, 153 upregulated genes, 36 downregulated; high confidence, 741 upregulated, 515 downregulated, Supplemental Data 1). We first analyzed the 1500 bp promoters of the regulated genes compared with unregulated ones. This analysis revealed a mild but detectable over-representation of ER8 in upregulated promoters (Supplemental Figure 4) compared with unregulated ones and nothing in downregulated genes. We also analyzed a set of genes that responded quickly to auxin (within 30 min to 1 h) in microarray data (Supplemental Data 1) and are thus likely enriched in direct ARF targets. Indeed, the ER8 enrichment becomes slightly stronger in promoter sequences of these auxin-induced genes (Supplemental Figure 4).

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

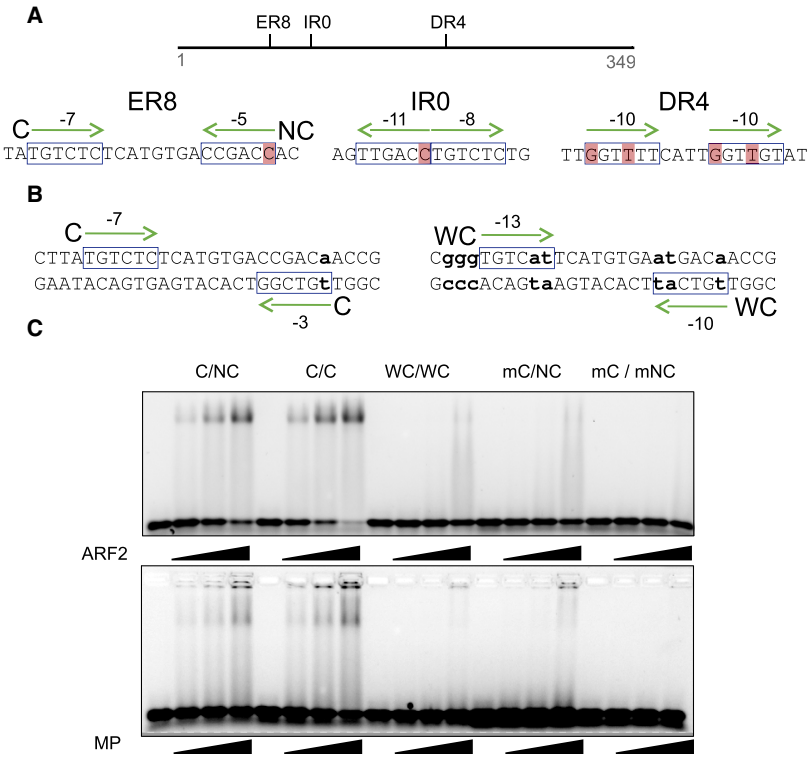


Figure 5. Analysis of *IAA19* promoter

(A) *Arabidopsis IAA19* promoter with position, sequence, and scores of ARFbs.

(B) ER8 and its variants used in EMSA.

(C) EMSA using ARF2 and MP proteins on probes described in **(B)** and two mutant probe controls with one (mC) or two mC/mNC sides of the ER8 mutated. ARF2 and MP are used at increasing concentrations: 0, 125, 250, and 500 nM. Color shading indicates difference from consensus.

C, consensus; mC, mutant consensus; NC, non-consensus; WC, worst consensus; mNC, mutant non-consensus.

Next, we tested whether more information could be extracted from these promoters if only the DNA segments bound by ARF in DAP-seq were analyzed. We focused on auxin-induced genes and regions bound by the MP activator, because the mechanism of gene induction by auxin is well understood, while repression by auxin and the role of repressor ARFs such as ARF2 is less clear. We therefore compared MP-bound regions present in regulated versus non-regulated promoters. We observed that the over-representation of ER8 and IR13 is higher in auxin upregulated genes than in non-regulated ones (Figure 6A and 6B). This is particularly striking for the high-confidence auxin-induced genes even if this list likely also contains indirect ARF targets (Figure 6A). When focusing on early auxin-induced genes intersected with MP-bound regions, we observe these same enriched elements plus others distant by one DNA helix turn (DR5/15, ER8/17, IR3/13/23) (Supplemental Figure 7). We experimentally tested MP binding to the IR13 probe and observed a strong and well-defined MP/IR13 complex (Figure 6C), similar to those obtained with ER7/8 probes. The IR0 element, also enriched in MP-bound DAP-seq regions but not in auxin-regulated promoters, gives a weaker smeary band. For auxin-repressed genes, two configurations (ER18 and IR3) are more over-represented in the MP-bound regions from promoters of downregulated genes than for non-regulated genes (Supplemental Figure 5). This might indicate that some ARFbs configurations could be involved in repression by auxin but this attractive hypothesis clearly needs to be tested with additional experiments. We used the same type of analysis (intersection between ARF-bound regions and auxin-regulated genes) for ARF6 and ARF3 regions obtained by ChIP-seq (Oh et al., 2014; Simonini et al., 2017). Whereas ARF3-bound regions or promoter of genes repressed by auxin yielded a number of novel configurations (Supplemental Figure 6B and 6D), the analysis of ARF6-bound regions present in promoter of auxin-induced genes recovered the same ER8 and IR13 over-represented configurations found for MP (Supplemental Figure 6A and 6C).

DISCUSSION

PWM versus Consensus for Auxin-Responsive Elements

A key question in auxin biology is how this structurally simple molecule evokes such diverse responses. TFs of the ARF family are likely the main contributors of auxin response diversity. Predictive tools to infer the presence of ARF binding sites in regulatory regions are essential both for functional and evolutionary analyses. Most studies so far have used TGTC-containing consensus sequences as a tool to detect ARFbs (Berendzen et al., 2012; O'Malley et al., 2016; Zemlyanskaya et al., 2016). Here, we built PWM-based models and showed that they provide a greater versatility than consensus sequences as they allow the sensitivity and specificity to be adjusted. Even if a TGTC consensus is perfectly suitable to detect the over-representation of some configurations (such as ER7-8 for ARF2 and MP) (O'Malley et al., 2016) (Supplemental Figure 2), it cannot be used to search regulatory regions because of its lack of specificity when used as monomer and its extremely low sensitivity when used as ER7/8 dimer (Figure 4). We illustrated on a chosen example (the *IAA19* promoter) that a site can be bound without a TGTC consensus and not necessarily bound even when the consensus is present (Figure 5). The non-canonical ER8 site we detected in *IAA19* promoter was challenged and functionally validated by studies in yeast (Pierre-Jerome et al., 2016). Even if more elaborate models exist (Mathelier and Wasserman, 2013), PWMs have emerged as the simplest and most performant models. Still, we were surprised that, in a DAP-seq context where no other parameters (such as cofactors, histones, or chromatin accessibility) should influence TF/DNA binding, the PWM models could not reach better AUROC values especially for ARF2. We have tried models that integrate the DNA shape feature (Mathelier and Wasserman, 2013), but they did not significantly improve the prediction

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

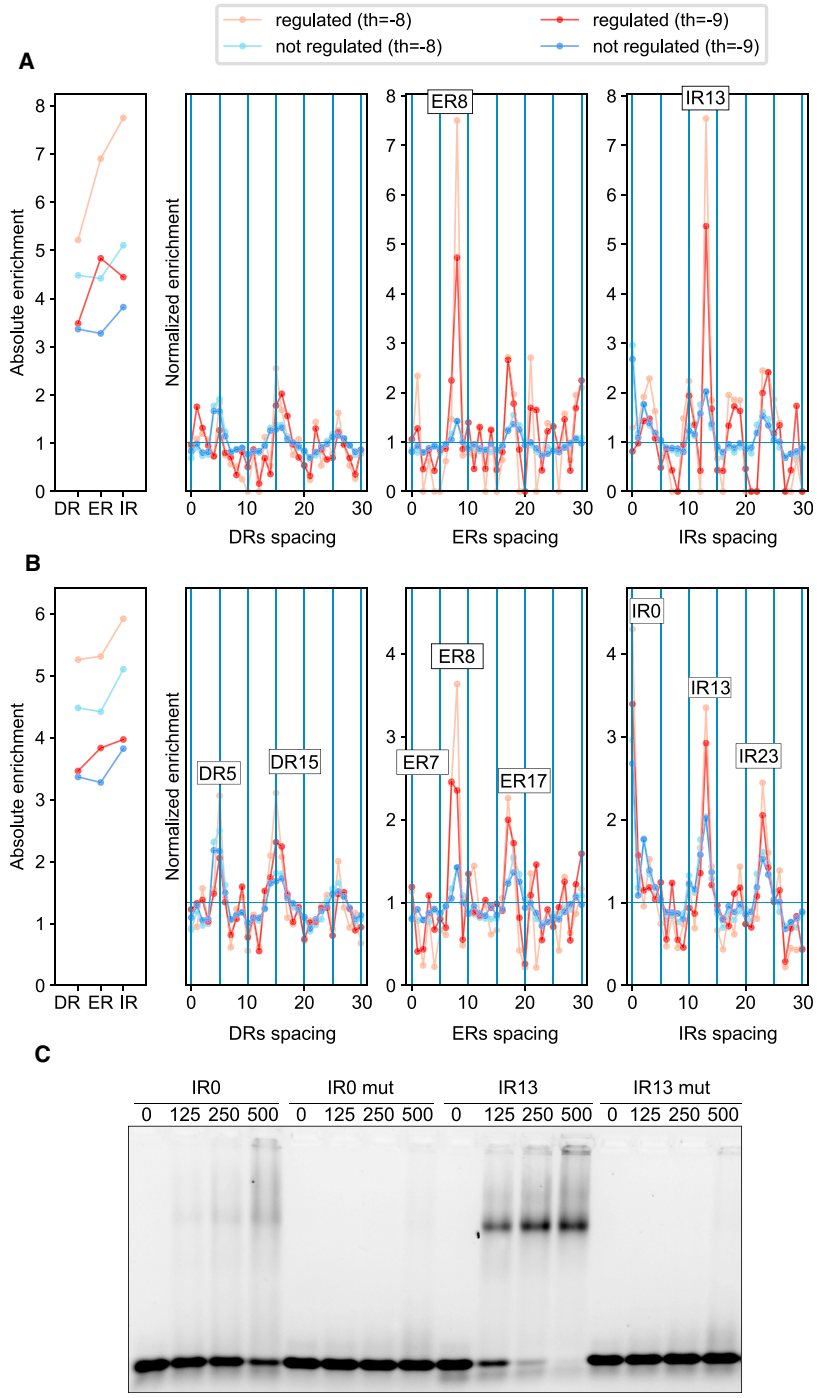


Figure 6. ARFbs Configuration Enriched in Auxin Upregulated Genes

(A and B) Spacing enrichment in promoter regions bound by MP were analyzed in auxin upregulated very high-confidence **(A)** or high-confidence genes **(B)** (red colors) and non-auxin-regulated genes **(B)** (blue colors) at two different score thresholds. The enrichment of ER7/8 and IR13 is increased for the very high confidence auxin upregulated genes.

(C) EMSA showing the binding of MP to IR0 and IR13 motifs and the corresponding control mutant probes. MP is used at increasing concentrations: 0, 125, 250, and 500 nM.

power (data not shown). The newly identified sequences with stretches of As and Ts (Figure 3B) were not easy to integrate in improved models but might affect the overall context of ARF2-binding sites and contribute to ARF2-specific regions. This finding is reminiscent of the family of AT-rich motifs found as over-represented in promoters of auxin-responsive genes (Cherenkov et al., 2018). These elements were mostly found in ARF2-binding regions, and they were more associated with downregulation than with upregulation. More studies are needed to elucidate their role.

ARF2 versus MP

ARFs exist as activators and repressors (Dinesh et al., 2016). Affinity measurements on a few DNA sequences *in vitro* (the molecular caliper model) and consensus-based searches in genome-wide binding data both indicate that activator ARF MP and repressor ARF (ARF1 and ARF2) might have different preferences for ARFbs configurations (Boer et al., 2014; O'Malley et al., 2016). But one study examined only a few ER elements (Boer et al., 2014), whereas the other did not recover the long known ER7/8 elements and instead proposed IR7/8 (O'Malley et al., 2016). Using PWM-based models and re-analyzing DAP-seq data, we confirmed that MP and ARF2 have a similar monomeric binding site but differ in the syntax of binding sites (combinations of binding sites of ARF monomers) they recognize: ARF2 prefers ER7/8 while MP has a much wider range of preferences. For ER motifs (face-to-face DBDs), our results extend the molecular caliper model (Boer et al., 2014) at the genome-wide level with some larger spacings. Moreover, MP has wider syntax than ARF2 as it also includes enriched DR and IR motifs. Such findings cannot be accommodated with the molecular caliper model as they involve different orientations of the two DBDs than in ER (head to tail for DR and tail to tail for IR). As previously reported (O'Malley et al., 2016), MP shows an increased binding frequency every 10 bp for all DR-, ER-, and IR-enriched configurations. Because this spacing corresponds to a DNA helix turn, we can imagine that this configuration allows interaction between ARF proteins on the same side of the DNA. 3D modeling using the published ARF1 structure indicates that these interactions are unlikely to involve the same dimerization surface as for ARF1 (Figure 2D). The proximity of some ARF DD domains in 3D,

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

combined with possible flexibility of ARF DBD, suggest that these proteins might have evolved different dimerization modes with the same protein domain. Confirming this hypothesis will await their structural characterization. An alternative hypothesis is that the PB1 oligomerization domain contributes to stabilize the MP binding to preferred motifs, but this also remains to be tested. However, it should be also noted that a preference for 10-bp spaced binding sites does not necessarily imply the presence of protein–protein contacts. Indeed, it has been shown that the binding of a first protein in the DNA major groove favors the binding of a second one at a 10 bp distance through allosteric changes in DNA conformation (Kim et al., 2013). This mechanism could also be at work for ARF DNA binding.

It is interesting to note that ER7-8 is bound by both ARF2 and MP, whereas some configurations such as DR5 or IR13 are more specific to MP. If repressor ARFs act by competing with activator ARFs for ARFbs (as proposed in Vernoux et al., 2011), this competition will therefore depend on the nature of ARFbs (shared between activators and repressors or specific to only one class of ARFs). Examining the DR5 auxin response reporter that corresponds to a succession of seven to nine CCTTT**TGTCTC** elements expectedly identifies DR5, DR16, and DR27 as highest scoring sites for ARF2 and MP PWMs. These elements are enriched in MP DAP-seq data (Figure 2) but not in ARF2 data, suggesting that the DR5 reporter construct is highly auxin inducible because it is better recognized by activator ARFs (such as MP) than by repressors (such as ARF2) (Ulmasov et al., 1997; Liao et al., 2015). Specificity differences between ARF2 and MP help explain that *arf2* and *mp* mutations have different developmental impacts (*mp* predominantly affecting identity and polarity versus *arf2* mainly affecting growth and senescence) even in tissues such as embryo, leaf, or flowers where both genes are active as attested by the presence of a mutant phenotypes (Berleth and Jürgens, 1993; Hardtke and Berleth, 1998; Okushima et al., 2005; Lim et al., 2010). Extending the analysis performed on MP and ARF2 to all members of the ARF family should indicate whether ARFs from a given class (A, B, or C) (Finet et al., 2013) have a stereotypic behavior or whether there is also a diversity of properties within the class A ARFs, for example. Such differences would help explain how a single auxin signal can trigger different responses depending on the cell type where it is perceived (provided different cell types express different sets of ARF proteins). *In vivo*, other parameters will also play an important role for the response to auxin, such as the ARF interaction partners (Mironova et al., 2017) and chromatin accessibility.

ARF Binding versus Auxin Regulation

The analysis of auxin-induced genes using PWM models identified only a small over-representation of ER8 (Supplemental Figure 4), a motif shared by ARF2 and MP. As we anticipated that ARFbs might be diluted in whole promoter sequences, we collected the set of DNA regions present in promoters from auxin-induced genes that are also bound by MP in DAP-seq and compared it with MP-bound promoter regions from non-auxin-regulated genes. This analysis confirmed the over-representation of ER8 in auxin-induced genes but also identified IR13 as enriched motifs (Figure 6). IR13 is a novel element, well bound by MP *in vitro* that now requires *in planta* characterization. It is not enriched in ARF2-bound regions, suggesting it will likely be insensitive to competition by repressor ARF2. We also characterized auxin-repressed gene and revealed a few novel motifs (ER18 and IR3 [Supplemental Figure 5] and IR7 [Supplemental Figure 4]). Again, functional analysis of such motifs *in planta* will be important in the future. We anticipate that the strategy we designed here (combining DAP-seq data with expression studies) is a very general method to increase the signal/noise ratio in regulatory regions and better detect binding sites involved in regulation. DAP-seq is a powerful technique but it suffers from giving access to DNA that might never be accessible in the cell. The combination with differential expression studies (+/− stimulation or +/− TF activity) will be a powerful way to narrow down the number of regions examined and extract functional regulatory information.

METHODS

Bio-informatic Analyses

The TAIR10 version of the *Arabidopsis* genome was used throughout the analyses. The DAP-seq peaks were downloaded from http://neomorph.salk.edu/PlantCistromeDB. We sorted the peaks (200 bp) extracted from the narrowPeaks file accordingly to their Q value. An ARF2-bound region was considered to overlap with an MP-bound region if the overlap exceeded 100 bp. We used the Bedtools suite to assess the overlaps and retrieve genome sequences. The PWMs were generated using MEME Suite 4.12.0 (Bailey et al., 2009) on the 600 top peaks of ARF2 and MP according to the Q value.

For the ROC-AUC analysis, performing an ROC analysis requires a background set of unbound genomic regions. This set was built with a Python script that takes a *bed* file of bound genomic regions as input and randomly selects in the *Arabidopsis* genome regions of the same size, similar GC content, and with similar origin (intron, exon, or intergenic).

To search for dependencies between positions of the ARF PWM, we used the sequence alignment inferred by the MEME Suite (Bailey et al., 2009) to build a PWM and used it as input for Enologos, selecting the option "mutual info" (Workman et al., 2005).

Analysis of ARFbs Configurations

The absolute enrichment (A) for each type of configuration (DR, ER, IR) was calculated as the ratio between the total number of sites in each configuration C in the bound set of regions divided by the same number in the background set. Such calculations were done for different score thresholds and normalized by the ratio between the total number of monomeric sites (BS, with no threshold applied) in the foreground and in the background to account for the different sequence sizes of the two sets. S_{max} stands for the maximum spacing.

$$A_{C=\mathrm{DR,ER,IR}} = \frac{\sum_{i=0}^{S_{max}} C_{i_{pos}}}{\sum_{i=0}^{S_{max}} C_{i_{neg}}} \cdot \frac{\sum BS_{neg}}{\sum BS_{pos}}.$$

For the normalized enrichment, we inventoried all the dimer configurations made of two monomeric ARFbs with scores above the chosen threshold. We then calculated the frequency (*f*) of each particular conformation (DRn, ERn or IRn with $0 \leq n \leq S_{max}$) among all dimeric sites in the positive set of bound regions and in the background set.

$$f_{i,C=\mathrm{DR,ER,IR}} = \frac{C_i}{\sum_{C=\mathrm{DR,ER,IR},k=0}^{S_{max}} C_k}$$

The normalized enrichment (*N*) shown in Figure 2 corresponds to the ratio between frequencies in the positive set and in the negative set for a given spacing.

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

$$N_{i,C=\mathrm{DR,ER,IR}} = \frac{f_{i,c_{\mathrm{pos}}}}{f_{i,c_{\mathrm{neg}}}}$$

To illustrate the enrichment of a few chosen motifs (DR4-15, ER7-8, IR0-13), we identified all sequences displaying a potential binding site with a score higher than a −8 threshold. Next, we plotted the % of regions displaying a given motif in the Venn diagram regions. The same was done for AT-rich motifs with a score threshold for each AT-rich PWM of −10.

The ER7/ER8 PWM for ARF2 was built from the ARF2 monomer PWM. Both ARF2 bound and unbound sets of regions were scanned with these two PWMs, and the best score given to each region by either ER7 or ER8 was used to plot the ROC curve. For the analysis of specificity and sensitivity of TGTC-containing consensus sequences, we analyzed each region for the presence or absence of ER7 or ER8 consensus (TGTCNN-7/8N-NNGACA, TGTCGG-7/8N-CCGACA, TGTCTC-7/8N-GAGACA). A region was scored positive when containing at least one site.

For the analysis of auxin-regulated promoters, we used 1500 bp upstream of the first exon of each gene. All DAP-seq regions overlapping with the promoters were then selected for analyses.

The major scripts used are available on github: https://github.com/Bioinfo-LPCV-RDF. The frequency matrices used to infer the PWM can be downloaded on https://github.com/Bioinfo-LPCV-RDF/Scores.

Selection of Auxin-Regulated Genes

We selected auxin-regulated genes over 22 publicly available gene expression profiling datasets from the GEO database (Supplemental Table 2). The datasets were generated on seedlings or roots of *Arabidopsis thaliana* with different auxin concentrations and times of exposure to auxin (explored in Zemlyanskaya et al., 2016). Differentially expressed genes were defined as those expressed at least 1.5 times higher (lower) after auxin treatment comparing with control, with a false discovery rate (FDR) adjusted p value <0.05 (Welch t test with Benjamini-Yekutieli correction). We compiled four groups of auxin-regulated genes: induced or repressed genes with high or very high confidence (Supplemental Data 1); high-confidence genes, 741 auxin upregulated and 515 downregulated genes significantly (more than 1.5-fold, FDR adjusted $p < 0.05$) changed their expression after auxin treatment in two or more datasets; very high confidence genes, 153 auxin upregulated and 36 downregulated genes significantly changed their expression in four or more datasets.

In addition, we compiled two lists of early responsive genes: 235 auxin upregulated and 87 downregulated genes (more than 1.5-fold, FDR adjusted $p < 0.05$) changed their expression in at least one of 11 datasets within 30 min to 1 h of auxin treatment (Supplemental Data 1).

Expression and Purification of Recombinant Proteins

ARF2 and ARF5 coding sequences were cloned into pHMGWA vectors (Addgene) containing N-terminal His-MBP-His tags. His-MBP-His-tagged ARF proteins were expressed in *Escherichia coli* Rosetta2 strain. Bacteria cultures were grown in liquid LB medium to an $OD_{600\ nm}$ of 0.6–0.9. Protein expression was induced with isopropyl-β-D-1-thyogalactopiranoside (IPTG) at a final concentration of 400 μM. Protein production was done overnight at 18°C. Bacteria cultures were centrifuged, and the resulting pellets were resuspended in lysis buffer (Tris-HCl 20 mM [pH 8]; NaCl 500 mM; Tris(2-carboxyethyl) phosphine (TCEP) 1 mM for ARF2 and Tris-HCl 20 mM [pH 8]; NaCl 500 mM; EDTA 0.5 mM; PMSF 0.5 mM; TCEP 1 mM; Triton 0.2% (w/v) for ARF5) with EDTA-free antiprotease (Roche) for sonication. Proteins were separated from the soluble fraction on Ni-Sepharose columns (GE Healthcare) previously equilibrated with the corresponding lysis buffer. Elutions were done with Imidazole 300 mM diluted in the corresponding lysis buffer.

Electrophoretic Mobility Shift Assays

DNA–protein interactions were characterized by electrophoretic mobility shift assays (EMSAs). The ER8 binding site was isolated from *Arabidopsis IAA19* promoter and ER8 variant sequences are given in Supplemental Table 1. IR0 and IR13 sequences were artificially designed with TGTCGG consensus sites (Supplemental Table 1). EMSA DNA probes were prepared from lyophilized oligonucleotides corresponding to the sense and antisense strands (Eurofins). Oligonucleotides for the sense strand presented an overhanging G in 5′ for DNA labeling. Sense and antisense oligonucleotides were annealed in Tris-HCl 50 mM, NaCl 150 mM. The annealing step was performed at 98°C for 5 min, followed by progressive cooling overnight. Annealed oligonucleotides, at a final concentration of 200 nM, were incubated at 37°C for 1 h with Cy5-dCTP (0.4 μM) and Klenow enzyme in NEB2 buffer (New England Biolabs). The enzyme was inactivated by a 10-min incubation at 65°C. Oligonucleotides were conserved at 4°C in darkness. EMSAs were performed on agarose 2% (w/v) native gels prepared with Tris-borate, EDTA (TBE) buffer 0.5×. Gels were pre-run in TBE buffer 0.5× at 90 V for 90 min at 4°C. Protein-DNA mixes nonspecific unlabeled DNA competitor (salmon and herring genomic DNA, Roche Applied Science; final concentration 0.045 mg/ml) and labeled DNA (final concentration 20 nM) in the interaction buffer: 25 mM HEPES (pH 7.4); 1 mM EDTA; 2 mM $MgCl_2$; 100 mM KCl; 10% glycerol (v/v); 1 mM DTT; 0.5 mM PMSF; 0.1% (w/v) Triton. Mixes were incubated in darkness for 1 h at 4°C and next loaded in the gels. Gels were run for 1 h at 90 V at 4°C in TBE 0.5× DNA–protein, and bindings were visualized on the gels with Cy5-exposition filter (Bio-Rad ChemiDoc MP Imaging System).

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the Agence Nationale de la Recherche (ANR-12-BSV6- 0005 Auxiflo to R.D., T.V., F.P.), PhD fellowships from the University Grenoble Alpes (R.M.-A.), the Grenoble Alliance for Cell and Structural Biology (ANR-10-LABX-49-01) to F.P., R.D., A.S., Russian State Budget (0324-2019-0040) to V.M., and Russian Foundation for Basic Research (18-04-01130) to V.M.

AUTHOR CONTRIBUTIONS

F.P. and R.D. designed and supervised the project, A.S., J.L., A.B., and V.M. performed the bioinformatic analyses, R.M.-A. and T.V.-P. performed the biochemical experiments, all authors discussed the results, F.P. wrote the manuscript with the help of A.S., R.D., T.V.-P., R.M.-A., and V.M.

ACKNOWLEDGMENTS

We thank Anthony Mathelier for discussions, Line Andresen and Chloe Zubieta for critical reading of the manuscript, David Mast and Laura Grégoire for input at the early stage of this work. No conflict of interest declared.

Received: June 18, 2018
Revised: August 31, 2018
Accepted: September 28, 2018
Published: October 15, 2018

REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. **37**:W202–W208.

Berendzen, K.W., Weiste, C., Wanke, D., Kilian, J., Harter, K., and Dröge-Laser, W. (2012). Bioinformatic cis-element analyses performed in *Arabidopsis* and rice disclose bZIP- and MYB-related binding sites as potential AuxRE-coupling elements in auxin-mediated transcription. BMC Plant Biol. **12**:125.

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

Berg, O.G., and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. **193**:723–750.

Berleth, T., and Jürgens, G. (1993). The role of the *monopteros* gene in organising the basal body region of the *Arabidopsis* embryo. Development **118**:575–587.

Boer, D.R., Freire-Rios, A., van den Berg, W.A., Saaki, T., Manfield, I.W., Kepinski, S., López-Vidrieo, I., Franco-Zorrilla, J.M., de Vries, S.C., Solano, R., et al. (2014). Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. Cell **156**:577–589.

Bulyk, M.L. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. **30**:1255–1261.

Cherenkov, P., Novikova, D., Omelyanchuk, N., Levitsky, V., Grosse, I., Weijers, D., and Mironova, V. (2018). Diversity of cis-regulatory elements associated with auxin response in *Arabidopsis thaliana*. J. Exp. Bot. **69**:329–339.

Dinesh, D.C., Villalobos, L.I.A.C., and Abel, S. (2016). Structural biology of nuclear auxin action. Trends Plant Sci. **21**:302–316.

Finet, C., Berne-Dedieu, A., Scutt, C.P., and Marlétaz, F. (2013). Evolution of the ARF gene family in land plants: old domains, new tricks. Mol. Biol. Evol. **30**:45–56.

Franco-Zorrilla, J.M., López-Vidriero, I., Carrasco, J.L., Godoy, M., Vera, P., and Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc. Natl. Acad. Sci. U S A **111**:2367–2372.

Guilfoyle, T., Hagen, G., Ulmasov, T., and Murfett, J. (1998). How does auxin turn on genes? Plant Physiol. **118**:341–347.

Han, M., Park, Y., Kim, I., Kim, E.H., Yu, T.K., Rhee, S., and Suh, J.Y. (2014). Structural basis for the auxin-induced transcriptional regulation by Aux/IAA17. Proc. Natl. Acad. Sci. U S A **111**:18613–18618.

Hanley, J.A., and McNeil, B.J. (1982). Maximum attainable discrimination and the utilization of radiologic examinations. J. Chronic Dis. **35**:601–611.

Hardtke, C.S., and Berleth, T. (1998). The *Arabidopsis* gene MONOPTEROS encodes a transcription factor mediating embryo axis formation and vascular development. EMBO J. **17**:1405–1411.

Keilwagen, J., Grau, J., Paponov, I.A., Posch, S., Strickert, M., and Grosse, I. (2011). De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. PLoS Comput. Biol. **7**:e1001070.

Kim, S., Brostromer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y.Q., et al. (2013). Probing allostery through DNA. Science **339**:816–819.

Korasick, D.A., Westfall, C.S., Lee, S.G., Nanao, M.H., Dumas, R., Hagen, G., Guilfoyle, T.J., Jez, J.M., and Strader, L.C. (2014). Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. Proc. Natl. Acad. Sci. USA **111**:5427–5432.

Korasick, D.A., Chatterjee, S., Tonelli, M., Dashti, H., Lee, S.G., Westfall, C.S., Fulton, D.B., Andreotti, A.H., Amarasinghe, G.K., Strader, L.C., et al. (2015). Defining a two-pronged structural model for PB1 (Phox/Bem1p) domain interaction in plant auxin responses. J. Biol. Chem. **290**:12868–12878.

Lavy, M., and Estelle, M. (2016). Mechanisms of auxin signaling. Development **143**:3226–3229.

Leyser, O. (2018). Auxin signaling. Plant Physiol. **176**:465–479.

Liao, C.-Y., Smet, W., Brunoud, G., Yoshida, S., Vernoux, T., and Weijers, D. (2015). Reporters for sensitive and quantitative measurement of auxin response. Nat. Methods **12**:207–210.

Lim, P.O., Lee, I.C., Kim, J., Kim, H.J., Ryu, J.S., Woo, H.R., and Nam, H.G. (2010). Auxin response factor 2 (ARF2) plays a major role in regulating auxin-mediated leaf longevity. J. Exp. Bot. **61**:1419–1430.

Mathelier, A., and Wasserman, W.W. (2013). The next generation of transcription factor binding site prediction. PLoS Comput. Biol. **9**:e1003214.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., et al. (2015). RSAT 2015: regulatory sequence analysis tools. Nucleic Acids Res. **43**:W50–W56.

Mironova, V.V., Omelyanchuk, N.A., Wiebe, D.S., and Levitsky, V.G. (2014). Computational analysis of auxin responsive elements in the *Arabidopsis thaliana* L. genome. BMC Genomics **15**:S4.

Mironova, V., Teale, W., Shahriari, M., Dawson, J., and Palme, K. (2017). The systems biology of Auxin in developing embryos. Trends Plant Sci. **22**:225–235.

Moyroud, E., Minguet, E.G., Ott, F., Yant, L., Posé, D., Monniaux, M., Blanchet, S., Bastien, O., Thévenon, E., Weigel, D., et al. (2011). Prediction of regulatory interactions from genome sequences using a biophysical model for the *Arabidopsis* LEAFY transcription factor. Plant Cell **23**:1293–1306.

Nanao, M.H., Vinos-Poyo, T., Brunoud, G., Thévenon, E., Mazzoleni, M., Mast, D., Lainé, S., Wang, S., Hagen, G., Li, H., et al. (2014). Structural basis for oligomerization of auxin transcriptional regulators. Nat. Commun. **5**:3617.

Nelson, C.C., Hendy, S.C., Faris, J.S., and Romaniuk, P.J. (1996). Retinoid X receptor alters the determination of DNA binding specificity by the P-box amino acids of the thyroid hormone receptor. J. Biol. Chem. **271**:19464–19474.

O'Malley, R.C.O., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. Cell **166**:1598.

Oh, E., Zhu, J.Y., Bai, M.Y., Arenhart, R.A., Sun, Y., and Wang, Z.Y. (2014). Cell elongation is regulated through a central circuit of interacting transcription factors in the *Arabidopsis* hypocotyl. Elife https://doi.org/10.7554/eLife.03031.

Okushima, Y., Mitina, I., Quach, H.L., and Theologis, A. (2005). AUXIN RESPONSE FACTOR 2 (ARF2): a pleiotropic developmental regulator. Plant J. **43**:29–46.

Parcy, F., Vernoux, T., and Dumas, R. (2016). A glimpse beyond structures in auxin-dependent transcription. Trends Plant Sci. **21**:574–583.

Pierre-Jerome, E., Moss, B.L., Lanctot, A., Hageman, A., and Nemhauser, J.L. (2016). Functional analysis of molecular interactions in synthetic auxin response circuits. Proc. Natl. Acad. Sci. USA **113**:11354–11359.

Sayou, C., Nanao, M.H., Jamin, M., Posé, D., Thévenon, E., Grégoire, L., Tichtinsky, G., Denay, G., Ott, F., Peirats Llobet, M., et al. (2016). A SAM oligomerization domain shapes the genomic binding landscape of the LEAFY transcription factor. Nat. Commun. **7**:11222.

Simonini, S., Bencivenga, S., Trick, M., and Østergaard, L. (2017). Auxin-induced modulation of ETTIN activity orchestrates gene expression in *Arabidopsis*. Plant Cell **29**:1864–1882.

Ulmasov, T., Liu, Z.B., Hagen, G., and Guilfoyle, T.J. (1995). Composite structure of auxin response elements. Plant Cell **7**:1611–1623.

Please cite this article in press as: Stigliani et al., Capturing Auxin Response Factors Syntax Using DNA Binding Models, Molecular Plant (2018), https://doi.org/10.1016/j.molp.2018.09.010

Ulmasov, T., Murfett, J., Hagen, G., and Guilfoyle, T.J. (1997). Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. Plant Cell **9**:1963–1971.

Vernoux, T., Brunoud, G., Farcot, E., Morin, V., Van den Daele, H., Legrand, J., Oliva, M., Das, P., Larrieu, A., Wells, D., et al. (2011). The auxin signalling network translates dynamic input into robust patterning at the shoot apex. Mol. Syst. Biol. **7**:1–15.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. **5**:276–287.

Weijers, D., and Wagner, D. (2016). Transcriptional responses to the auxin hormone. Annu. Rev. Plant Biol. **67**:539–574.

Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D., and Benos, P.V. (2005). enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. **33**:W389–W392.

Zemlyanskaya, E.V., Wiebe, D.S., Omelyanchuk, N.A., Levitsky, V.G., and Mironova, V.V. (2016). Meta-analysis of transcriptome data identified TGTCNN motif variants associated with the response to plant hormone auxin in *Arabidopsis thaliana* L. J. Bioinform. Comput. Biol. **14**:1641009.

Zhao, Y., Ruan, S., Pandey, M., and Stormo, G.D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. Genetics **191**:781–790.

Supplemental data

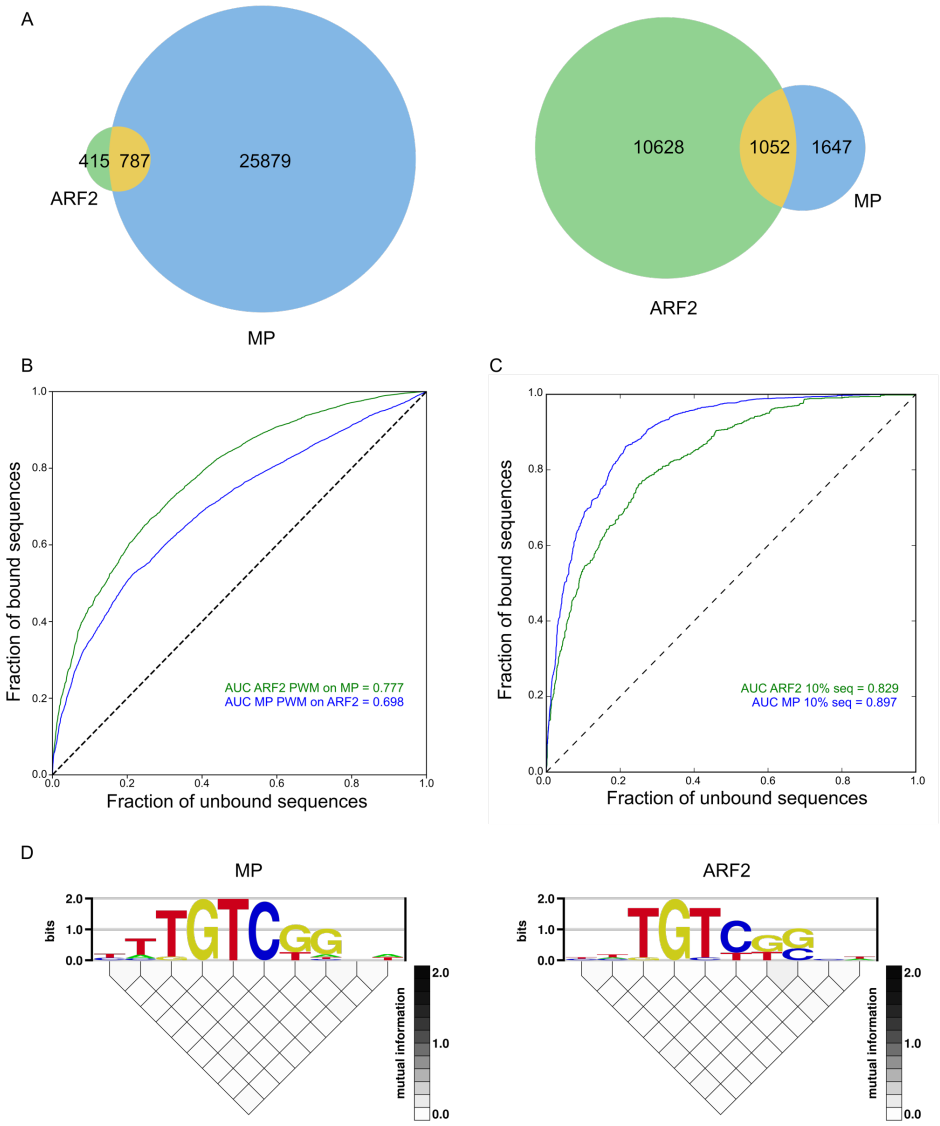
Capturing auxin response factors syntax using DNA binding models

Arnaud Stigliani[1], Raquel Martin-Arevalillo[1,2], Jérémy Lucas[1], Adrien Bessy[1], Thomas Vinos-Poyo[1], Victoria Mironova[3,4], Teva Vernoux[2], Renaud Dumas[1] and François Parcy[1,*]

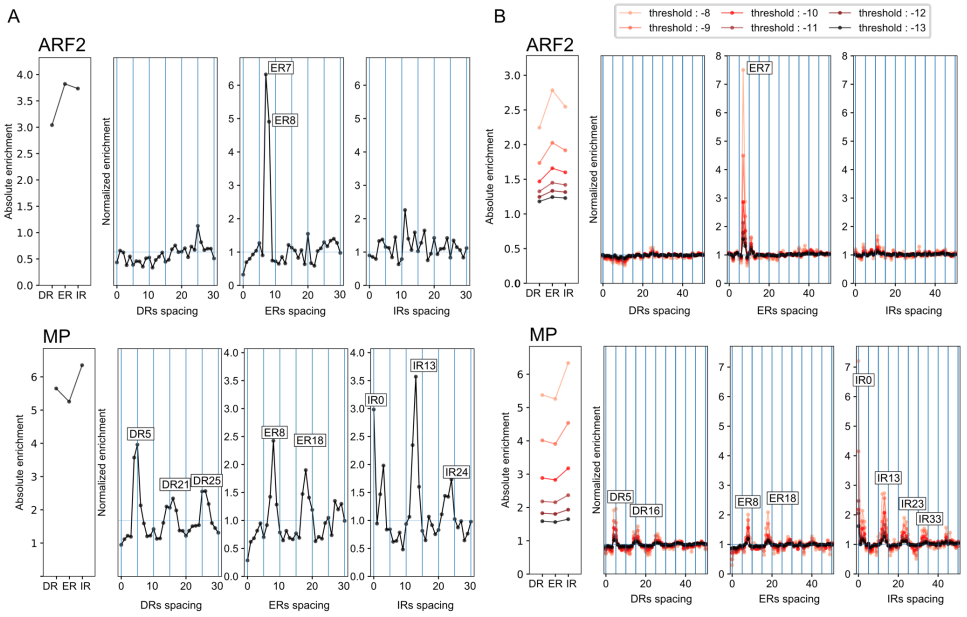
This file contains

7 Supplemental figures

2 Supplemental tables



Supplemental Figure 1: **(A)** 2 Venn diagrams with the 10% top bound regions for ARF2 against all MP regions and the 10% top bound regions for MP against all ARF2 regions. This shows that there are regions specifically bound by a single factor even in the highest confidence regions **(B-C)** ROC curves with ARF2 PWM on MP bound regions and MP PWM on ARF2 regions. AUROC value decrease slightly as compared to Figure 1 **(D)** Enologos analysis of MP and ARF2 motifs (Workman et al., 2005). No dependency between nucleotide position is detected.

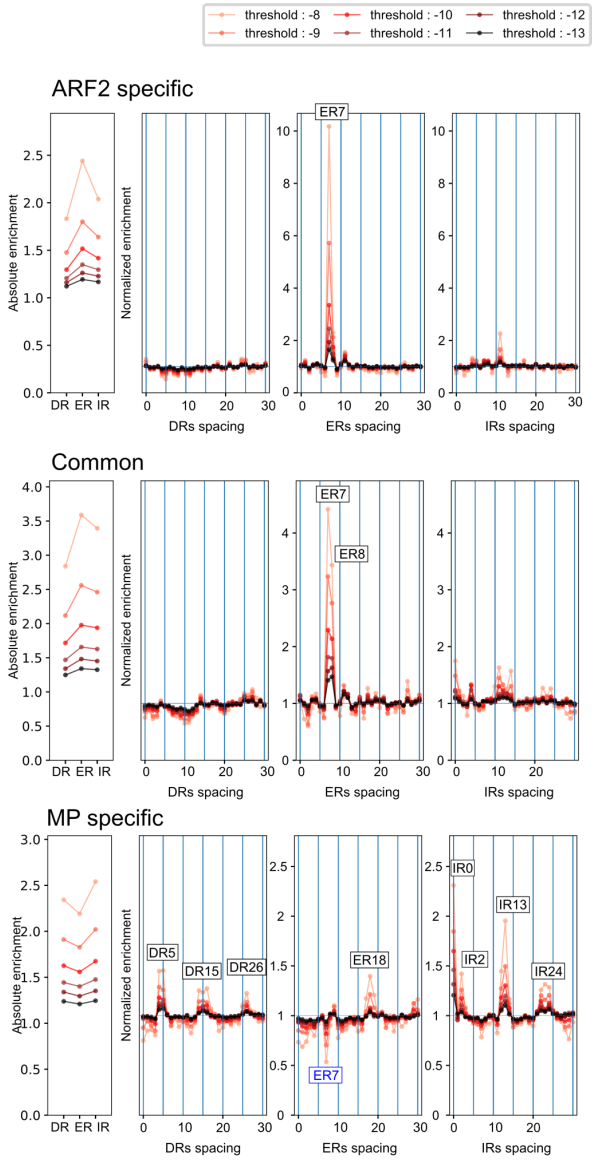


22

23 **Supplemental Figure 2**: **(A)** Enrichment of spacings between TGTC **(B)** Spacing
24 enrichment for DRn, ERn and IRn for 0≤n≤50. Threshold indicates the PWM score
25 threshold value used for ARFbs detection.

26

27



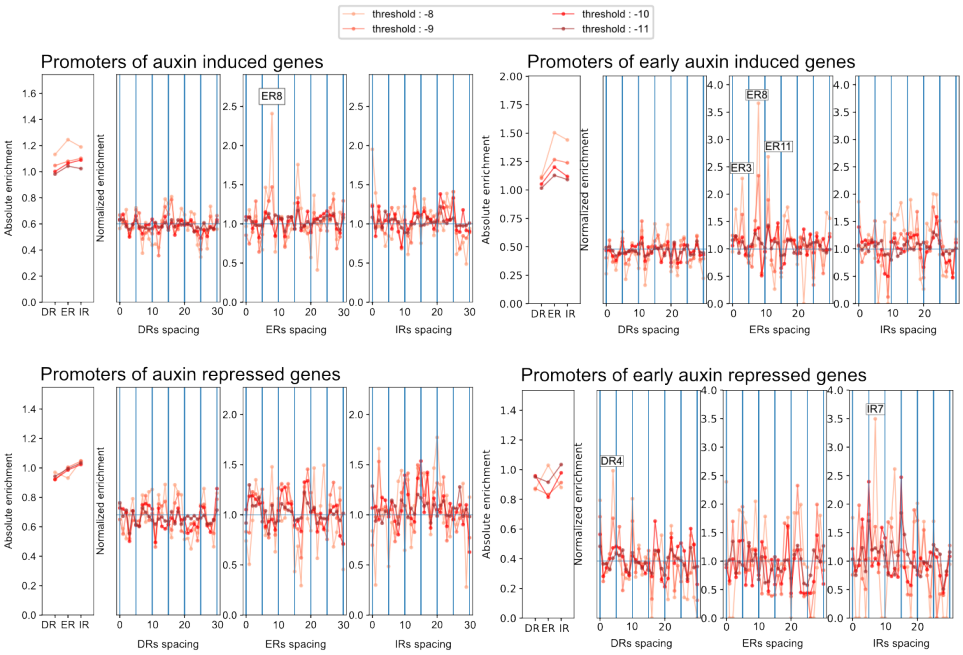
28

29

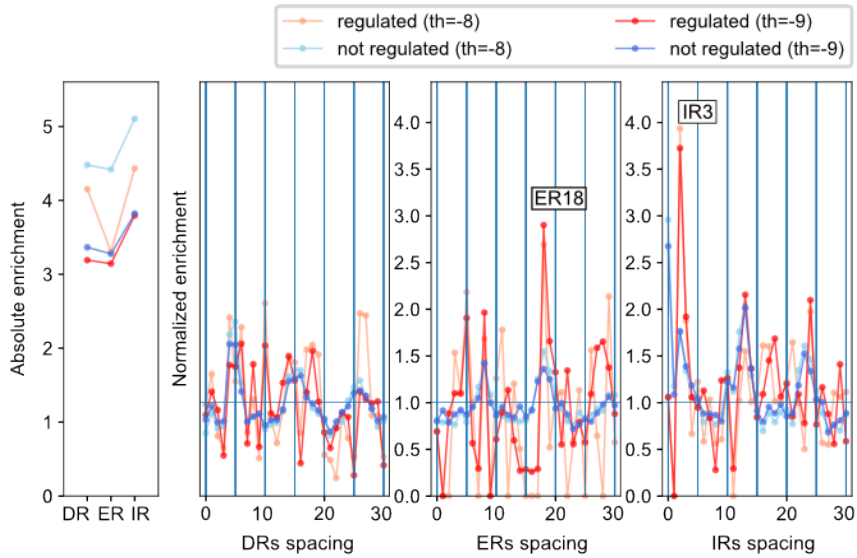
30 **Supplemental Figure 3:** Spacing enrichment in MP-specific, ARF2-specific and
31 MP/ARF2 common regions, compared to unbound sets of sequences. Threshold
32 indicates the PWM score threshold value used for ARFbs detection. Note ER7 is
33 depleted in MP-specific bound regions.

34

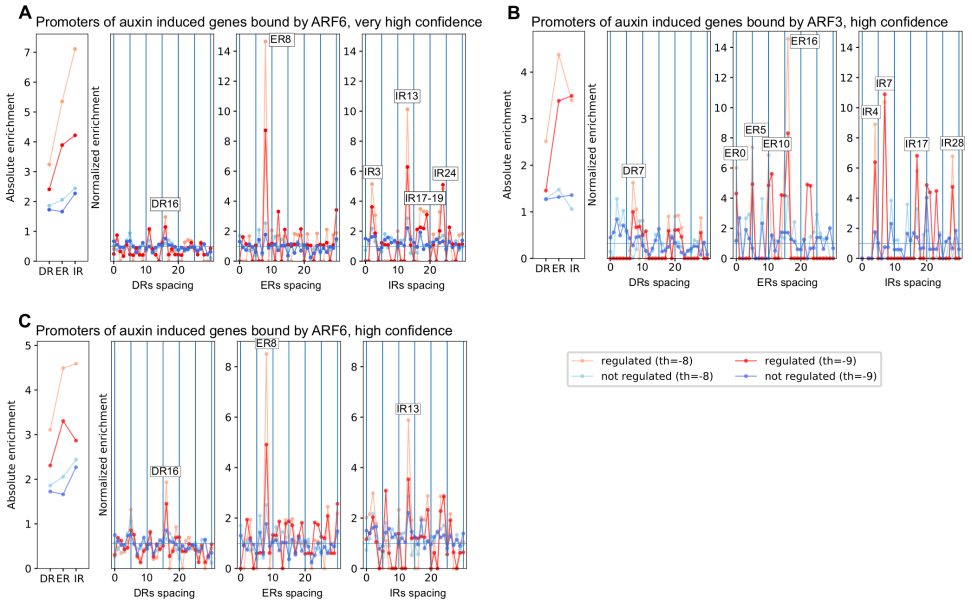
35



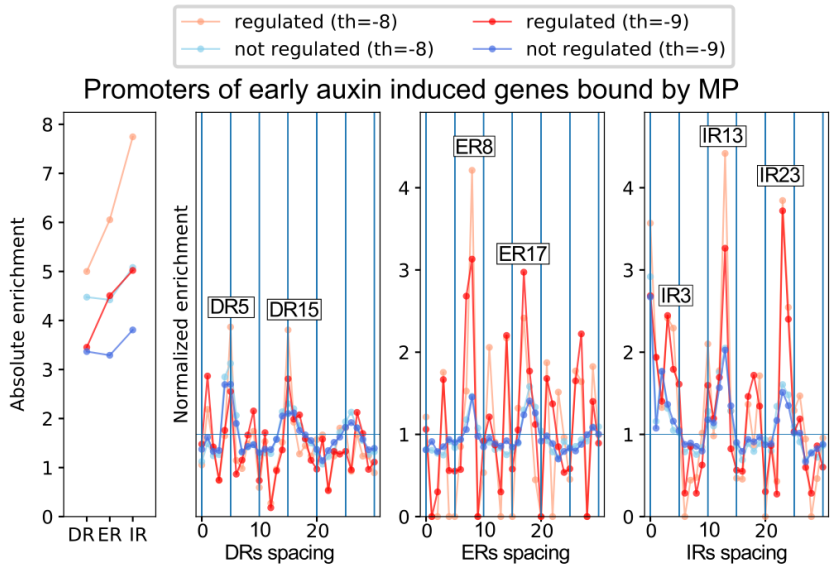
Supplemental Figure 4: ARFbs over-represented conformations in the promoters of the auxin up-regulated genes (upper panel) or the down-regulated genes (lower panel) (Supplemental File). We used combined very high and high confidence genes (left panels) or early auxin regulated genes (right panels) compared to auxin insensitive gene promoters. Threshold indicates the PWM score threshold value used for ARFbs detection.



Supplemental Figure 5: Promoter regions bound by MP were analysed in down-regulated (red colours) and non-regulated genes (blue colours) in high confidence gene lists (Supplemental File). The regions not bound by MP from auxin insensitive promoters were used as background. Threshold indicates the PWM score threshold value used for ARFbs detection.



Supplemental Figure 6: Spacing enrichment in promoters regions bound by ARF6 (A, C) or ARF3 (B) were analysed in auxin up-regulated high (B-C) or very high (A) confidence genes and non-auxin regulated genes (blue colours) at two different score thresholds.



Supplemental Figure 7: ARFbs over-represented conformations in the promoters of early auxin up-regulated genes (Supplemental File). We compared MP-bound regions from promoters of genes induced by auxin (after 0.5 to 1h treatment) to MP-bound regions of auxin insensitive gene promoters. Threshold indicates the PWM score threshold value used for ARFbs detection.

Supplemental Table 1. Sequences of DNA probes for EMSAs. Bold letters show ARF binding sequence. Lower case letters indicate the nucleotides variation introduced.

Oligonucleotide	DNA sequence (5'->3')
ER8 C/NC	GCAAACTTA**TGTCTC**TCATGTGA**CCGACC**ACCGCATC
ER8 C/C	GCAAACTTA**TGTCTC**TCATGTGA**CCGACa**ACCGCATC
ER8 WC/WC	GCAAACggg**TGTCat**TCATGTGA**atGACa**ACCGCATC
ER8 mC/NC	GCAAACTTA**TGTCTC**TCATGTGA**CCGttC**ACCGCATC
ER8 mC/mNC	GCAAACTTA**TaaCTC**TCATGTGA**CCGttC**ACCGCATC
IR0	GATGCAGTCATGTG**CCGACATGTCGG**CATGTGCTCACAT
IR0 mut	GATGCAGTCATGTG**CCGttATaaCGG**CATGTGCTCACAT
IR13	GATGCAG**CCGACA**AAACACATGATTT**TGTCGG**CTCACAT
IR13 mut	GATGCAG**CCGttA**AAACACATGATTT**TaaCGG**CTCACAT

Supplemental Table 2. Gene expression profiling datasets on auxin treatments used to create the lists of auxin repressed and auxin activated genes.

No.	GEO accession number	Tissue, Developmental Stage	Treatment (concentration, time)	Number of replicates
1-5	GSE35580[1]	Roots of 7 dag seedlings, epidermis, pericycle, stele, columella	5 µM IAA, 2 h	3
6	GSE627[2-3]	7 dag seedling	5 µM IAA, 2 h	3
7	GDS3505[4]	Roots of 3 dag seedlings	1 µM IAA, 4 h	2
8-9	GDS1515[5]	Root segments of 3 dag seedlings	10 µM NAA, 2 h; 10 µM NAA, 6 h	2
10	GDS1044[6]	7 dag seedlings	10 µM IAA, 1 h	4
11	GDS744[7]	5 dag seedlings	10 µM IAA, 2 h	2
12-15	GDS672[8]	10 dag seedlings	0.1 µM IAA, 1 h; 0.1 µM IAA, 3 h; 1 µM IAA, 1 h; 1 µM IAA, 3 h	2
16-22	GSE42007[9]	Roots of 6 dag seedlings	1 µM IAA, 0.5 h; 1 µM IAA, 1 h; 1 µM IAA, 2 h; 1 µM IAA, 4 h; 1 µM IAA, 8 h; 1 µM IAA, 12 h; 1 µM IAA, 24 h	3

1. Bargmann BO, Vanneste S, Krouk G, Nawy T, Efroni I, Shani E, Choe G, Friml J, Bergmann DC, Estelle M, Birnbaum KD, A map of cell type-specific auxin responses, *Mol Syst Biol* **9**:688, 2013.
2. Okushima Y, Mitina I, Quach HL, Theologis A, AUXIN RESPONSE FACTOR 2 (ARF2): a pleiotropic developmental regulator, *Plant J* **43**:29–46, 2005.
3. Overvoorde PJ, Okushima Y, Alonso JM, Chan A, Chang C, Ecker JR, Hughes B, Liu A, Onodera C, Quach H, Smith A, Yu G, Theologis A, Functional genomic analysis of the *AUXIN/INDOLE-3-ACETIC ACID* gene family members in *Arabidopsis thaliana*, *Plant Cell* **17**:3282–3300, 2005.
4. Stepanova AN, Yun J, Likhacheva AV, Alonso JM, Multilevel interactions between ethylene and auxin in *Arabidopsis* roots, *Plant Cell* **19**:2169–2185, 2007.
5. Vanneste S, De Rybel B, Beemster GT, Ljung K, De Smet I, Van Isterdael G, Naudts M, Iida R, Gruissem W, Tasaka M, Inze D, Fukaki H, Beeckman T, Cell cycle progression in the pericycle is not sufficient for SOLITARY ROOT/IAA14-mediated lateral root initiation in *Arabidopsis thaliana*, *Plant Cell* **17**:3035–3050, 2005.
6. Armstrong JI, Yuan S, Dale JM, Tanner VN, Theologis A, Identification of inhibitors of auxin transcriptional activation by means of chemical genetics in *Arabidopsis*, *Proc Natl Acad Sci USA* **101**:14978–14983, 2004.
7. Nemhauser JL, Mockler TC, Chory J, Interdependency of brassinosteroid and auxin signaling in *Arabidopsis*, *PLoS Biol* **2**:E258, 2004.
8. Redman JC, Haas BJ, Tanimoto G, Town CD, Development and evaluation of an *Arabidopsis* whole genome Affimetrix probe array, *Plant J* **38**:545–561, 2004.

Lewis DR, Olex AL, Lundy SR, Turkett WH, Fetrow JS, Muday GK, A kinetic analysis of the auxin transcriptome reveals cell wall remodeling proteins that modulate lateral root development in *Arabidopsis*, *Plant Cell* **25**:3329–3346, 2013.

1.3 Bilan

Cet article montre que malgré un motif de liaison presque similaire, ARF2 et ARF5 ne lient pas les mêmes régions du génome. Nous avons pu voir que des préférences de configurations de liaison en dimères différentes pouvaient expliquer que ARF2 et ARF5 ne lient pas les mêmes régions. Ainsi, ARF2 a une préférence unique pour ER7 alors que ARF5 lie des régions contenant DR4-5, DR14-16, ER8, ER18, IR0, IR13, IR23. Nous avons ainsi pu voir que les régions liées par MP contiennent un enrichissement en sites de liaison tous les 10 paires de base – 10.5 paires de bases correspondant à un tour d'hélice d'ADN – qui suggère peut-être que MP peut former des oligomères. Enfin, en étudiant les promoteurs des gènes liés par MP et régulés par l'auxine, nous nous sommes aperçus que ER8, ER17, IR23 et particulièrement IR13 sont sur-représentées en comparaison des régions liées mais pas régulées. Ceci suggère que si un TF a diverses préférences de liaison, certaines sont plus propices à la régulation.

1.4 Discussion

1.4.1 Chez *A. thaliana*

Dans cet article, au moment d'étudier les gènes régulés par l'auxine, nous avons fait le choix arbitraire de prendre des promoteurs de 1500 paires de bases avant le $5'UTR$. Depuis, nous nous sommes rendus compte que dans les gènes sur-exprimés, les pics de de DAP-Seq de ARF5/MP se situent en majorité dans les 400 paires de bases précédant le $5'UTR$ (figure 30). Si ce résultat peut aider à choisir la taille des promoteurs, il constitue également une signature pour identifier les gènes régulés.

Nous avons également analysé la nature des gènes régulés possédant un élément de type IR13 avec un bon score. En effet, dans l'article, nous avions montré que la configuration IR13 était nettement sur-représentée dans les gènes sur-exprimés en présence d'auxine. Ainsi, nous avons étudié les promoteurs (de 500 paires de bases) des gènes activés par l'auxine et liés en DAP-Seq par ARF5, nous permettant d'établir une liste de 95 promoteurs. Sur ces 95 promoteurs, 16 contiennent un IR13 dont le score de chaque monomère est bon (compris entre 0 et -12) Un de ces gènes est remarquable par le très bon score du site IR13 : chaque score de monomère est égal à -3. Nous avons été agréablement surpris de constater qu'il s'agissait de l'emblématique gène *GH3* chez lequel la présence d'un tel élément dimérique n'avait jamais été décrite comme tel. Il serait intéressant dans le futur d'étudier l'activité de ces éléments chez la plante et de tester si leur expression dans les différents tissus d'Arabidopsis est la même ou si au contraire, ils confèrent des profils d'expression distincts en répondant à des ARF aux spécificités et profils d'expression différents. Suite à nos résultats, des telles expériences ont été initiées chez notre collaborateur à Lyon (T. Vernoux, RDP, ENS de Lyon) pour les éléments IR13, ER8, DR5 et DR15.

1.4.2 Dans le maïs

Depuis la parution de notre article, une étude similaire (Galli et al., 2018) à celle de O'Malley et al. (2016) a été conduite sur de nombreux ARF de maïs : des activateurs et une sous-famille des ARF répresseurs. Il y est suggéré que tous les ZmARF appartenant respectivement aux activateurs ou aux répresseurs lient les mêmes configurations de dimères, avec parfois quelques variations. Ainsi, les ZmARF activateurs ont tendance à préférer $\mathrm{DR}(4/5+10k), k \in \boldsymbol{N}$, ER8 et IR13. Les ZmARF répresseurs semblent lier ER7/8 et ER29. Le signal ER29 pose question dans la mesure où il est extrêmement marqué et où ne nous l'avons pas observé pour AtARF2, une protéine répresseur d'*A. thaliana*.

FIGURE 30: Position des pics d'ARF5/MP par rapport aux gènes dans *A. thaliana*. L'axe y donne la densité de pics par gène et l'axe x donne la position des pics sur les gènes. Les positions négatives correspondent aux positions avant le $5'UTR$, le sens des gènes ayant été pris en compte.

Les motif des sites monomériques des ZmARF activateurs montrent que ces TF ont une préférence partagée pour TGTCGG, où les nucléotides GG semblent avoir une moindre influence sur la liaison que le reste du motif. De manière plus surprenante, les ZmARF répresseurs semblent lier TGTCCCCC. En réalisant le DAP-Seq des ZmARF sur de l'ADN génomique d'*A. thaliana*, l'équipe menant l'étude a pu montrer que les ZmARF répresseurs retrouvent la préférence connue pour TGTCGG. Ce résultat a montré que la nature de l'ADN génomique – et notamment le pourcentage en GC du génome – utilisé influencent fortement les résultats du DAP-Seq.

Cette étude possède cependant des limites. Les configurations de dimères observées n'ont pas été obtenues en regard d'un set de régions non liées. Compte tenu de la forte influence de la nature de l'ADN sur les résultats observés, ceci pourrait sensiblement altérer les résultats obtenus.

Chapitre 2

LEAFY, un exemple pour mieux comprendre la liaison des TF

Note : Dans cette partie, je présenterai des résultats obtenus en collaboration avec Jérémy Lucas.

LEAFY est un facteur de transcription clé dans le développent floral et le sujet historique de mon équipe d'accueil. Ce facteur a été très étudié et nous disposons aujourd'hui de données de SELEX, de DAP-Seq, et deux jeux de données de ChIP-Seq. Un premier jeu permet d'étudier les sites de liaison de LFY constitutivement exprimé sous le contrôle du promoteur *35S* (p35S :LFY) dans des plantules de 2 semaines (Sayou et al., 2016). L'autre jeu étudie le comportement de LFY dans le méristème d'inflorescence. La construction p35S :LFY-GR permet de contrôler la présence conditionnelle de LFY dans le noyau. Les tissus ont été récoltés 4h après le traitement à la dexaméthasone, qui permet de « simuler » la transition florale (Goslin et al., 2017). LFY offre donc l'opportunité de répondre à des questions fondamentales à la fois à propos des modèles de liaison, dont on peut comparer l'efficacité sur les différents jeux de données et sur l'influence des facteurs cellulaires, en comparant DAP-Seq et ChIP-Seq.

Dans ce chapitre, je vais tout d'abord utiliser les données DAP-Seq pour évaluer nos modèles en incluant les TFFM, encore jamais utilisés pour LFY. Dans un deuxième temps, je comparerai les résultats de ChIP-Seq et de DAP-Seq pour voir s'ils coïncident parfaitement ou s'il existe des régions qui sont préférentiellement liées *in vivo* ou *in vitro*. Si ces régions existent, j'essaierai d'expliquer leur raison d'être.

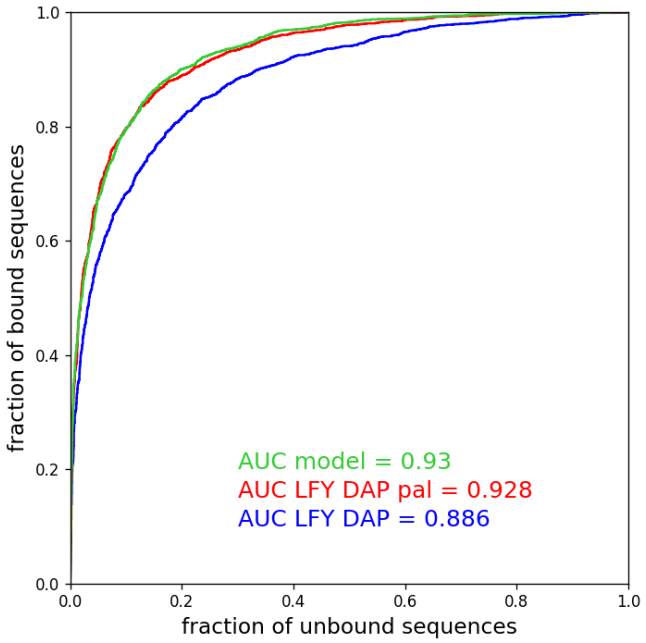
2.1 Construction d'un modèle de liaison

2.1.1 Construire un modèle basé sur les PWM

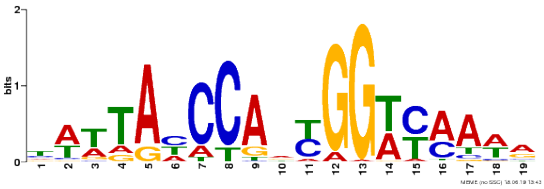
Comme nous venons de l'expliquer, un DAP-Seq de LFY nous a permis de construire et d'évaluer différents modèles de liaison. Le DAP-Seq, réalisé au sein de l'équipe par Xuelei Lai, utilise de l'ADN génomique nu au préalable amplifié par PCR (cette variation de la technique est appelée ampDAP-Seq). L'amplification par PCR fait disparaître les marques de méthylations sur les cytosines, ce paramètre ne sera donc pas pris en compte dans le modèle.

Les données ont été filtrées selon la méthode du paragraphe II.1.1. Ne disposant pas de réplicat pour s'assurer de la reproductibilité des expérience de DAP-Seq, nous avons choisi de ne considérer que les 3000 meilleurs pics d'après le critère de la couverture normalisée (paragraphe II.1.2.1), sur les 9165 obtenus au départ.

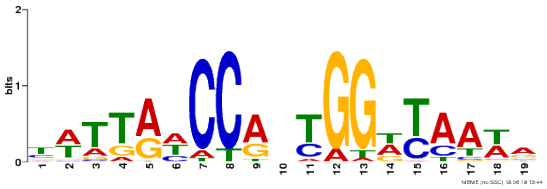
LFY ayant la propriété de lier l'ADN comme un dimère symétrique (Hamès et al., 2008), nous avons cherché à produire des modèles qui prennent en compte cette propriété. Aidés du programme meme-suite (paragraphe II.3.1.2), les 300 meilleurs pics nous ont permis de déterminer deux PWM, l'une palindromique qui incorpore cette symétrie (appelée PWM "LFY DAP PAL", figure 31.c) et l'autre non (PWM "LFY DAP", figure 31.b). Le pouvoir prédictif de ces deux PWM a été comparé à celui de la PWM symétrique avec dépendances (appelée PWM "model", figure 31.d), construite par Moyroud et al. (2011b). À cette fin, nous avons évalué la capacité de nos modèles à discriminer les 3000 régions liées par rapport à un set de régions non liées d'après le critère de l'AUROC (voir paragraphe II.3.2). Les résultats sont présentés dans la figure 31.a.



(a)



(b) PWM "LFY DAP"



(c) PWM "LFY DAP pal"

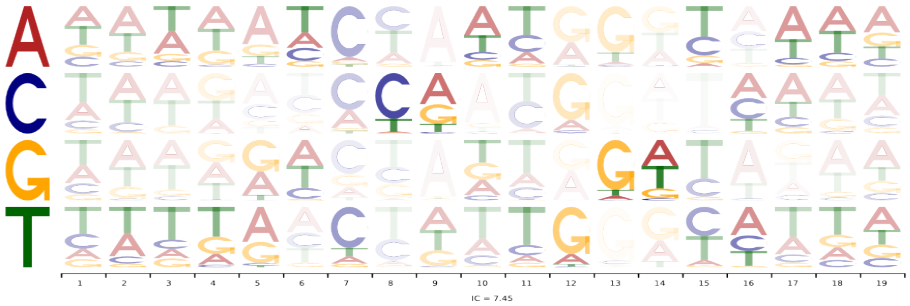
(d) PWM "model"

FIGURE 31: (a) Courbes ROC obtenues sur les 3000 régions liées. (b) PWM obtenue à partir des régions liées en DAP-Seq. (c) PWM obtenue à partir des régions liées en DAP-Seq, en forçant la symétrie. (d) Modèle tenant compte des dépendances, obtenu en SELEX par Moyroud et al. (2011b)

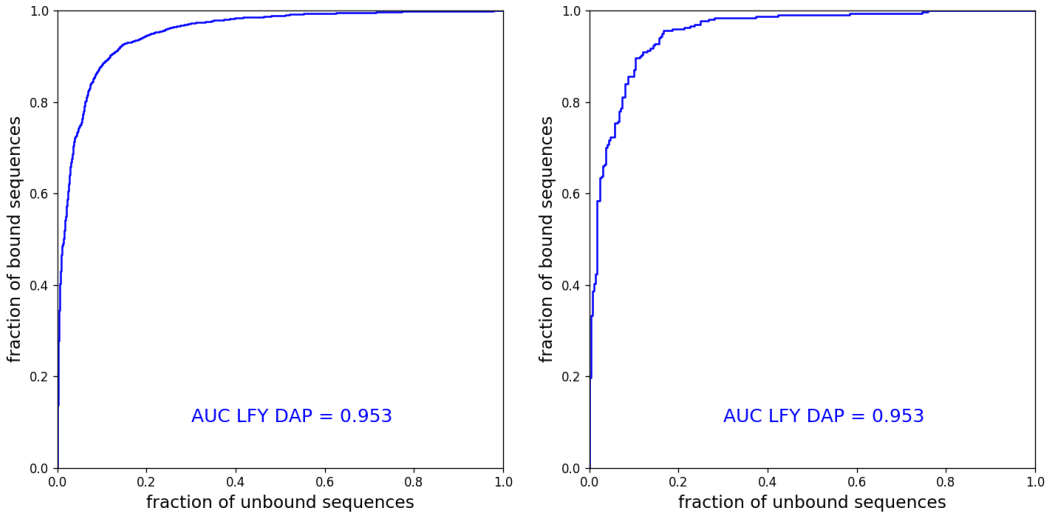
La figure 31 montre que forcer la palindromie de la PWM améliore la prédiction des régions liées (figure 31.(a,b,c)). Le pouvoir prédictif de la "PWM DAP PAL" est sensiblement le même que celui de la PWM "model" (figure 31.(a,c,d)).

2.1.2 Construire un modèle basé sur les TFFM et la structure de l'ADN

Dans la suite, nous avons cherché à savoir si ces modèles pouvaient être améliorés par des éléments qui interviennent dans la liaison sans être capturés par les PWM. D'abord, nous avons produit une TFFM à partir de la PWM "LFY DAP pal" (figure 31.a) et des 300 meilleurs pics de DAP-Seq. Les performances de cette TFFM ont été évaluées par AUROC (figure 32.b).



(a)



(b)

(c)

FIGURE 32: (a) TFFM obtenue sur les 300 meilleurs pics du DAP-Seq de LFY. (b) Test de la TFFM sur les 3000 régions liées en DAP-Seq par LFY. (c) Test du modèle intégrant TFFM et structure de l'ADN sur les 300 régions liées en DAP-Seq du set de test

Dans un deuxième temps, nous avons implémenté l'influence de la structure de l'ADN dans notre modèle. Ainsi, nous avons pu prendre en compte l'*Helix Twist*, la largeur du grand sillon, le *Proppeller Twist* et le *Roll* (figure 14). Le paquet *DNAshapedTFBS* (Mathelier et al., 2016) permet de produire un modèle qui intègre à la fois le score calculé par la TFFM et ces éléments de structure de l'ADN au niveau des sites de liaison. Si nous avons pu constater que peu de régions suffisent pour entraîner une TFFM

correctement, l'algorithme implémenté dans le paquet *DNAshapedTFBS* fait appel à des méthodes de *machine learning* qui nécessitent un plus grand set d'apprentissage et un set de régions non liées. En effet, il apparaît que les performances du modèle restaient assez faibles lorsque celui-ci était entraîné sur les 300 meilleurs pics du DAP-Seq de LFY. À l'inverse, l'apprentissage du modèle sur l'ensemble des régions liées contre les régions non liées le conduit à s'adapter aux sets utilisés. Par conséquent, on ne peut pas tester le modèle sur les régions utilisées pour le produire, sous peine de surestimer son pouvoir prédictif. Ainsi, nous avons scindé les pics de DAP-Seq en un set d'apprentissage (2700 pics pris au hasard) et un set de test (les 300 pics restants). Le modèle a été testé par AUROC grâce aux 300 pics du set de test (figure 32.c).

Il apparaît que si le pouvoir prédictif de la TFFM est supérieur à celui des PWM (figures 31.a et 32.b), la prise en compte de la structure de l'ADN ne contribue pas à améliorer le modèle. Pour cette raison, nous n'avons pas cherché à examiner les éléments de structure caractéristiques du site de liaison de LFY. L'apport de la TFFM sur la PWM "model" (figure 31.d) est assez surprenant compte tenu qu'elle implémente déjà des dépendances entre nucléotides. Il semble néanmoins que les dépendances données par la TFFM et celles de la PWM "model" sont différentes. Ainsi, la TFFM semble indiquer que dans les 4 premières et dans les 3 dernières positions du site de liaison, LFY préfère les successions de T, les successions de A ou les transitions A-T. Ceci n'apparaît pas dans la PWM, qui donne les principales dépendances aux positions (4-5-6), (9-10-11) et (14-15-16) visibles également dans la TFFM. Ces dépendances semblent cependant différentes d'un modèle à l'autre : alors que la PWM donne la préférence pour TAC en positions (4-5-6), la TFFM privilégie TAT, TGT ou TGA (le dernier triplé n'apparaissant pas dans la PWM). Comme pour la PWM, on observe une symétrie de ces dépendances aux positions (14-15-16). Enfin les dépendances aux positions (9-10-11) sont également différentes de la PWM à la TFFM, qui donne principalement AAT ATT ou GTT (AAT n'apparaît pas dans la PWM).

2.1.3 Discussion

Lors de la dernière mise à jour de la base de données JASPAR (Khan et al., 2017) (voir article en annexe V.2), nous avons contribué à ajouter 250 PWM obtenus en DAP-Seq publiés par O'Malley et al. (2016). Au préalable, nous avions étudié les performances des différentes PWM sur les régions liées (figure 33). En comparaison avec les AUROC qui sont généralement obtenues avec des PWM sur des données de DAP-Seq, 0.93 (figure 31.a) représente une valeur très haute.

L'AUROC basée sur la TFFM (0.95, figure 32.(a,b)) correspond donc à une amélioration notable sur un modèle déjà bon. Compte tenu du fait que nous ne disposons d'aucun réplicat et que la variabilité observée dans la reproduction des expériences (voir paragraphe II.1.2.1), il semble difficile d'obtenir un modèle ayant un plus grand pouvoir prédictif. Néanmoins, plusieurs pistes restent à explorer. Parmi celles-ci, les séquences « flanquantes » du site de liaison peuvent jouer un rôle dans le recrutement du TF (Raccaud et al., 2019; Marklund et al., 2013). Ainsi, nous avons étudié la (sur/sous)-représentation de nucléotides ou de di-nucléotides dans les 100 nucléotides adjacents au meilleur site de liaison de chaque séquence, dans un set de régions liées et dans un set de régions non liées, sans que nous puissions mettre en évidence une quelconque différence. Comme LFY se lie en dimère, le modèle utilisé est celui du double site de liaison, insensible aux sites monomériques. Une piste à explorer est donc la sensibilité de LFY aux sites monomériques en évaluant les performances d'un modèle monomérique sur les régions où le pouvoir prédictif du modèle actuel est faible. Remarquons enfin que l'amplification de l'ADN par PCR conduit à la suppression des cytosines méthylées, qui sont alors remplacées par des cytosines normales. Ce n'est donc pas un paramètre à prendre en compte pour augmenter le pouvoir prédictif du modèle en DAP-Seq.

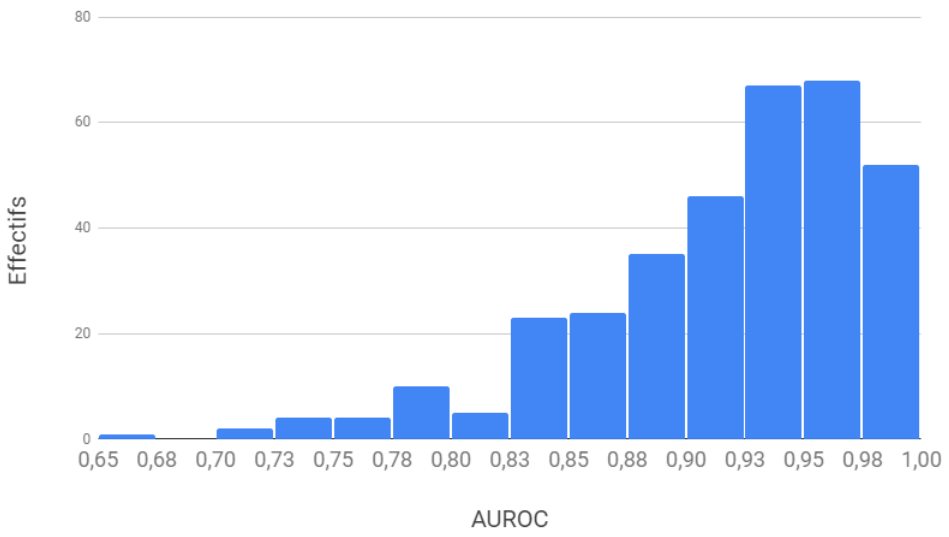


FIGURE 33: AUROC calculées à partir des DAP-Seq réalisés par O'Malley et al. (2016) et de leurs PWM respectives

2.2 Comparaison entre ChIP-Seq et DAP-Seq

Dans la partie précédente, nous avons optimisé un modèle de liaison à l'ADN de LFY en l'absence de facteurs cellulaires. Dans cette partie, nous souhaitons comprendre quels sont les facteurs exogènes pouvant influer sur la liaison. Dans ce but nous nous sommes aidés des ChIP-Seq de LFY que nous avons brièvement présentés en début de ce chapitre.

Après quelques précisions sur le traitement des données, nous allons consacrer un paragraphe aux différences observées entre le ChIP-Seq et le DAP-Seq. Aidés par une modélisation de l'interaction ADN/TF dont nous avons montré qu'elle était très efficace, nous avons ensuite souhaité identifier les éléments présents dans la cellule qui peuvent favoriser la liaison du TF. Nous avons donc fait l'étude des régions liées uniquement par ChIP-Seq. Enfin, nous avons pris le problème par l'autre bout en nous demandant quels paramètres peuvent empêcher la liaison de LFY dans la cellule. Pour cette question, nous avons fait l'étude des régions liées uniquement en DAP-Seq.

2.2.1 Traitement des données

Comme nous l'avons mentionné dans l'introduction de ce chapitre, nous avons analysé deux de données de ChIP-Seq : un premier qui donne accès aux régions liées par LFY dans le méristème d'inflorescence et un second qui donne accès aux régions liées par LFY, constitutivement exprimé dans des plantules. Afin d'obtenir la couverture normalisée du nombre de *reads* dans ces différentes expériences, les données ont été traitées d'après la procédure décrite dans le paragraphe II.2. L'obtention des pics n'a pas suivi de protocole unifié. En raison de contraintes temporelles, et de la confiance que nous accordions au ChIP-Seq de LFY en plantule (réalisé au sein de l'équipe), nous avons choisi d'utiliser les mêmes pics que dans l'étude de Sayou et al. (2016). Quant au ChIP-Seq réalisé dans l'inflorescence, les pics ont été obtenus d'après la commande macs2 détaillée dans le paragraphe II.1.2.1. Cependant, comme le ChIP-Seq n'a pas été répliqué, nous n'avons pas pu filtrer les pics et en raison de leur faible nombre, nous avons décidé de les

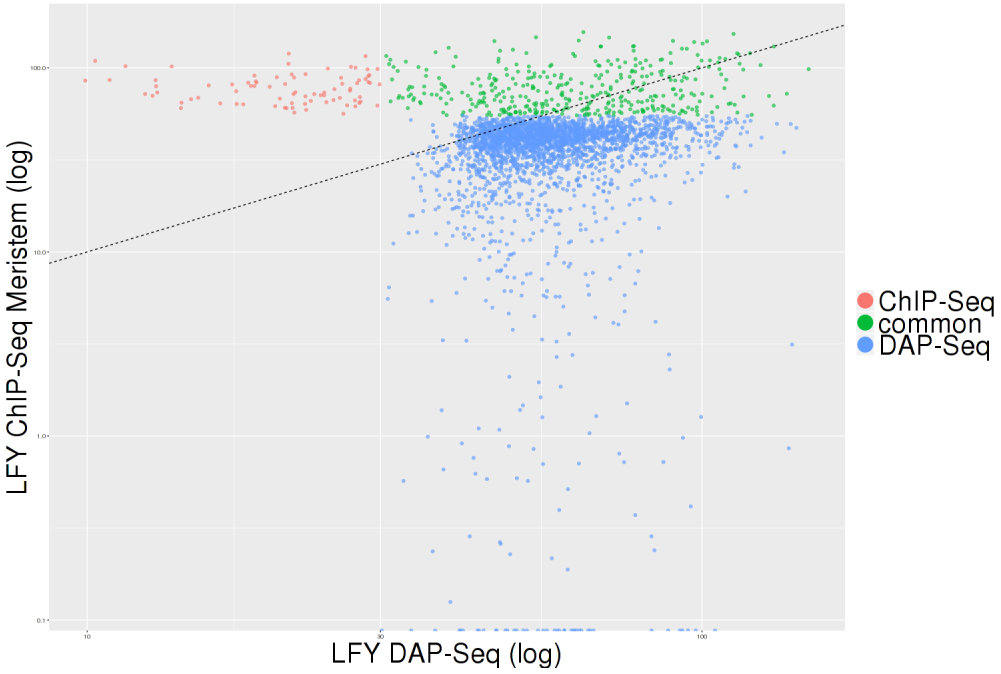
conserver tous. Les couvertures normalisées des signaux de ChIP-Seq et de DAP-Seq sont ensuite calculées sous les pics de ChIP-Seq et sous les 3000 meilleurs pics de DAP-Seq (figure 34). Afin de supprimer les duplicats des pics qui pourraient exister en ChIP-Seq comme en DAP-Seq, les pics communs sont fusionnés d'après l'algorithme de la figure 24. Nous avons fait 3 groupes (couleur des points sur la figure 34) :

— Les pics communs, car ayant une couverture normalisée forte à la fois en ChIP-Seq et en DAP-Seq

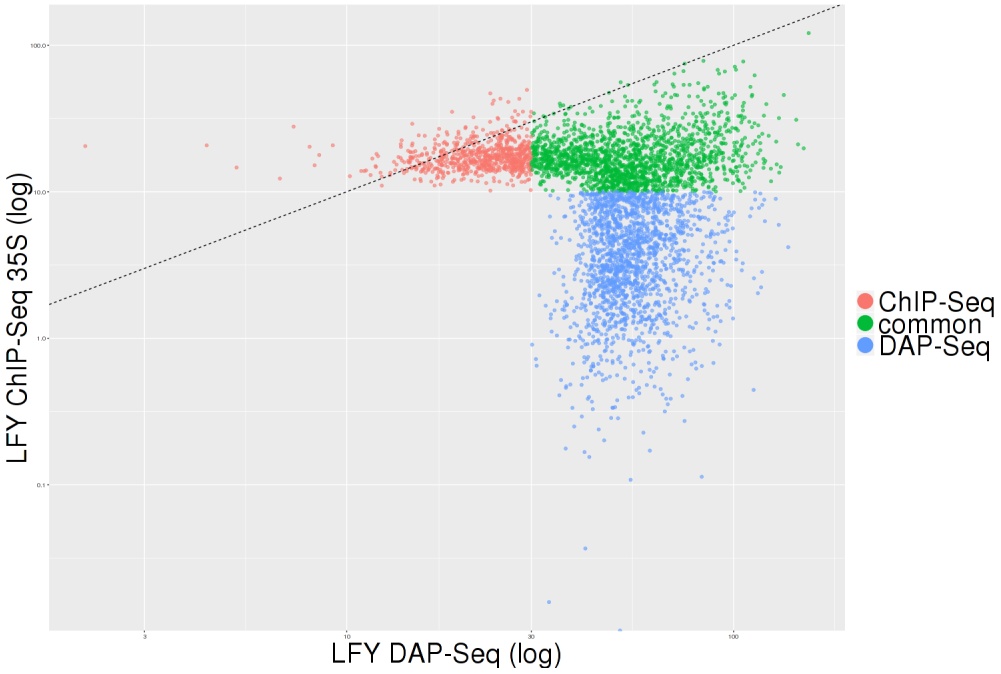
— Les pics uniques au DAP-Seq, ayant une couverture normalisée forte en DAP-Seq mais pas en ChIP-Seq

— Les pics uniques au ChIP-Seq, suivant la même méthodologie

Les seuils pour construire les 3 groupes ont été déterminés par l'observation de la figure 34.



(a) Pics de DAP-Seq et de ChIP-Seq en méristème



(b) Pics de DAP-Seq et de ChIP-Seq en plantule

FIGURE 34: Les couleurs bleu et rouge donnent respectivement les pics uniques de LFY en DAP-Seq et en ChIP-Seq. Les pics communs sont marqués en vert. Les axes x et y indiquent la couverture normalisée sous les pics.

2.2.2 Observations

Les tables suivantes détaillent les effectifs dans chaque expérience.

Pics ChIP-Seq méristème	Pics communs	Pics DAP-Seq
81	376	2683

Pics ChIP-Seq plantule	Pics communs	Pics DAP-Seq
702	1814	1767

Ces chiffres montrent de grandes différences entre le DAP-Seq et le ChIP-Seq. Parmi les régions différentiellement liées dans ces deux types d'expériences, on compte le gène *AP3*, impliqué dans l'activité B qui détermine l'identité des pétales et des étamines. Alors que des régions du gène sont liées uniquement en DAP-Seq, le promoteur du gène contient à la fois un pic propre au ChIP-Seq, et un pic commun au DAP-Seq et au ChIP-Seq (figure 35). Ces deux pics tombent dans deux régions régulatrices du gène *AP3* (Lamb et al., 2002), chacune étant strictement nécessaire pour que le gène soit activé. Ce gène illustre donc la complexité et la variété des phénomènes à expliquer : une région est liée seulement *in vivo*, une autre seulement *in vitro* et une dernière est liée dans les deux conditions.

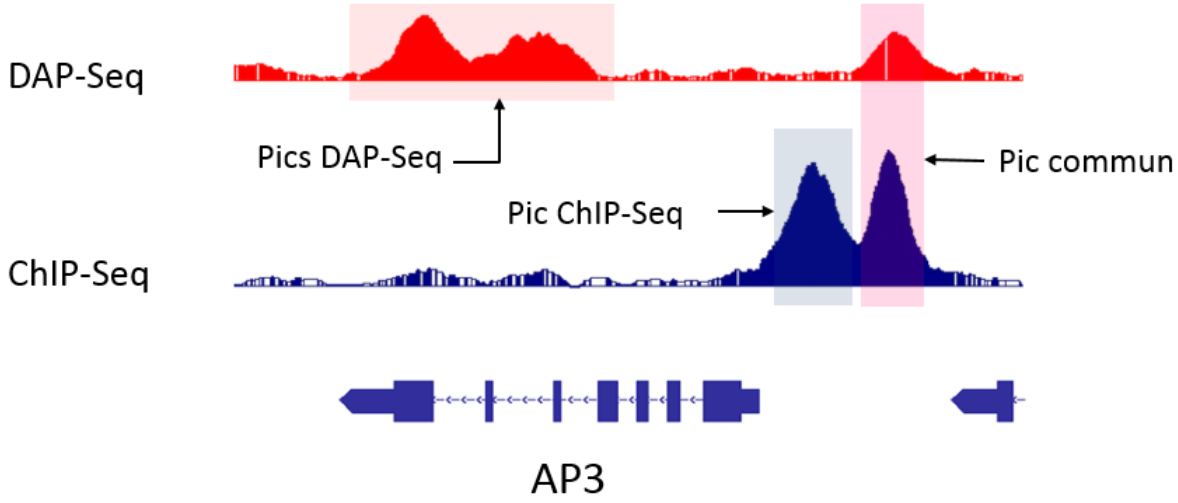
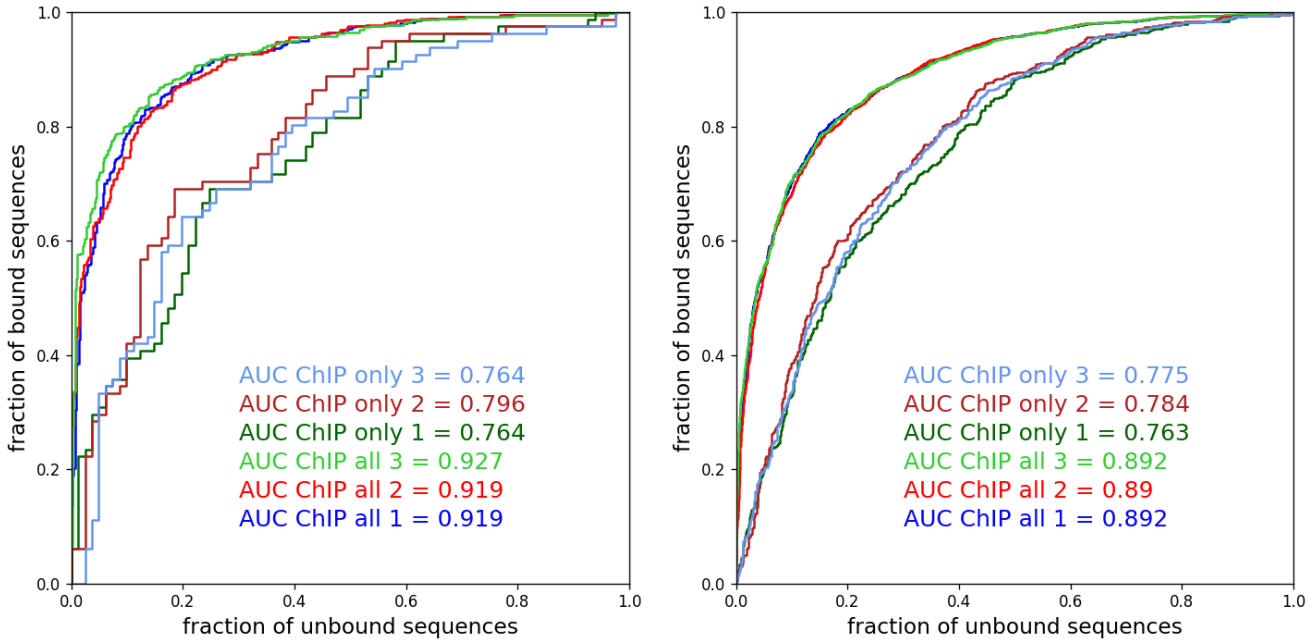


FIGURE 35: Représentation du gène AP3. Le signal en rouge montre la couverture du DAP-Seq de LFY alors que le signal en bleu représente la couverture du ChIP-Seq de LFY dans un plantule

Nous terminerons ce paragraphe en comparant les deux ChIP-Seq. Sur les 457 pics en ChIP-Seq de méristème, 374 sont partagés dans le celui en plantule. On peut donner plusieurs pistes expliquant ces différences. Il se peut que le contexte chromatinien et l'interactome de LFY exprimé diffèrent dans ces deux types cellulaires. Il est aussi possible que sa présence dans le noyau pendant 2 semaines ait activé la production de ces interacteurs. Ceux-ci peuvent ainsi l'aider à lier de nouvelles régions. Ajoutons que LFY est potentiellement pionnier, et qu'il aura pu permettre l'ouverture progressive de la chromatine en plusieurs endroits, la rendant ainsi plus accessible.

2.2.3 Déterminer les paramètres qui favorisent la liaison dans la cellule

Dans un premier temps, nous nous sommes interrogés sur la présence de site de liaison de LFY dans les régions uniquement liées en ChIP-Seq. En effet, la liaison de LFY y est potentiellement indirecte, puisqu'ils ne sont pas liés en DAP-Seq. Pour chaque ChIP-Seq, nous avons donc fait 2 groupes de régions : celles liées uniquement en ChIP-Seq, et celles liées communément en ChIP-Seq et en DAP-Seq. Nous y avons étudié la présence de site de liaison grâce à la TFFM (figure 32.a) à l'aide de courbes ROC suivant la méthode du II.3.2. Les jeu de données étant de petites tailles, nous les avons respectivement confrontés à trois sets de régions non liées, nous assurant ainsi que le choix du jeu de données non liées n'affecte pas les résultats (figure 36).



(a) Régions liées par LFY en ChIP-Seq de méristème

(b) Régions liées par LFY en ChIP-Seq de plantule

FIGURE 36: Performances de la TFFM sur toutes les régions liées en ChIP-Seq (all) ou sur les régions liées uniquement en ChIP-Seq (only). Les régions liées ont été confrontées à 3 set de régions non liées.

Les résultats de la figure 36 sont résumés dans le tableau suivant :

	ChIP-Seq méristème	ChIP-Seq plantule
AUROC Pics communs + pics ChIP-Seq	0.92	0.89
AUROC Pics ChIP-Seq	0.77	0.77

Si les performances du modèle sont à chaque fois moins bonnes sur la sous partie des régions liées uniquement en ChIP-Seq, les AUROC de 0.77 montrent l'existence de sites de liaison, et laissent donc supposer d'une interaction directe. Enfin, remarquons que les performances de la TFFM sur les régions liées en ChIP-Seq sont moins bonnes que sur les régions liées en DAP-Seq.

2.2.3.1 Déterminer des éventuels co-facteurs de LFY

La figure 36 nous apprend que des sites existent dans les régions liées uniquement en ChIP-Seq et que ceux-ci ont des affinités plus faibles que les sites de liaison trouvés en DAP-Seq. En effet, les courbes ROC qui se rapportent aux régions liées uniquement en ChIP-Seq ont une pente plus faible à l'origine que les courbes qui se rapportent à la totalité des régions liées en ChIP-Seq, montrant qu'elles contiennent moins de bons sites de liaison (la seule aire de la courbe ROC ne donne pas cette information, il se pourrait que les sites de liaison des régions liées uniquement en ChIP-Seq soient un mélange de très bons sites et de très mauvais sites). Ceci suggère que des éléments *in vivo* – des co-facteurs – pourraient aider le recrutement de LFY sur ces sites de faible affinité. De plus, l'existence de sites de liaison pour LFY, même s'ils ont des affinités faibles, suggère que le contact entre LFY et l'ADN est probablement direct. Afin de mettre en évidence des co-facteurs éventuels, nous avons utilisé le programme meme-suite et nous avons confronté les résultats à la base de données de profils de liaison JASPAR Plants (Khan et al., 2017). La procédure est décrite dans la rubrique "Confronter les motifs à des bases de données" du paragraphe II.3.1.2.

Sans différence notable dans les 2 ChIP-Seq, la sortie du programme donne un enrichissement important pour les motifs des TF à doigts de zinc, dont de nombreux facteurs de transcription de la famille DOF. On retrouve aussi des motifs de BCP, des facteurs de transcription impliqués dans le développement de la plante. Une CArGbox, le site de liaison des gènes à boîte MADS, dont beaucoup sont impliqués dans le développement floral, a également été mise en évidence. Enfin, des motifs de TCP sont présents dans les régions liées par LFY.

Ici, nous avons pu voir que la présence de certains co-facteurs pourraient expliquer comment LFY est capable de lier des sites de liaison de faible affinité. Nous nous sommes naturellement intéressés aux gènes cibles d'un potentiel complexe formé par LFY et un autre de ces TF. À cette fin, chaque gène du génome a été étendu de 3000 paires de bases en 5' et de 2000 paires de bases en 3'. Nous avons ensuite sélectionné les gènes où figure à la fois un pic propre au ChIP-Seq et un site de liaison d'un des éventuels co-facteurs de LFY. Parmi les gènes contenant un motif de TCP, on peut citer *AUX1* (qui contient aussi une CArGbox), impliqué dans le transport de l'auxine, *GI*, impliqué dans la floraison et le gène *AT3G07760*, qui code pour un domaine SAM, un interacteur potentiel de l'extrémité N-terminale de LFY. Il est également intéressant de noter qu'un site de liaison pour un TCP est trouvé dans le pic unique au ChIP-Seq du promoteur d'*AP3*.

Le croisement de ces nouvelles observations avec des données d'expression pourrait permettre une meilleure compréhension des phénomènes de régulation orchestrés par *LFY*

2.2.3.2 Discussion

Ainsi, nous avons mis en évidence des motifs de TF qui sont des interacteurs potentiels de LFY et des gènes intéressants compte tenu qu'ils sont impliqués dans le programme floral.

Il convient de préciser que d'autres paramètres pourraient expliquer que des régions sont uniquement liées en ChIP-Seq. En effet, La présence de cytosines méthylées *in vivo* peut favoriser la liaison de certains TF. Étant donné que l'ADN utilisé dans notre cas (ampDAP-Seq) n'est pas méthylé, nous n'avons pas pu mesurer l'importance de ce paramètre dans la liaison. Pour pallier cette limite, de nouvelles expériences de DAP-Seq sont en cours dans le laboratoire ; au lieu de mettre en présence l'ADN amplifié par PCR et le TF d'intérêt, la technique consiste à utiliser de l'ADN génomique non amplifié. Seuls les fragments liés sont amplifiés avant le séquençage.

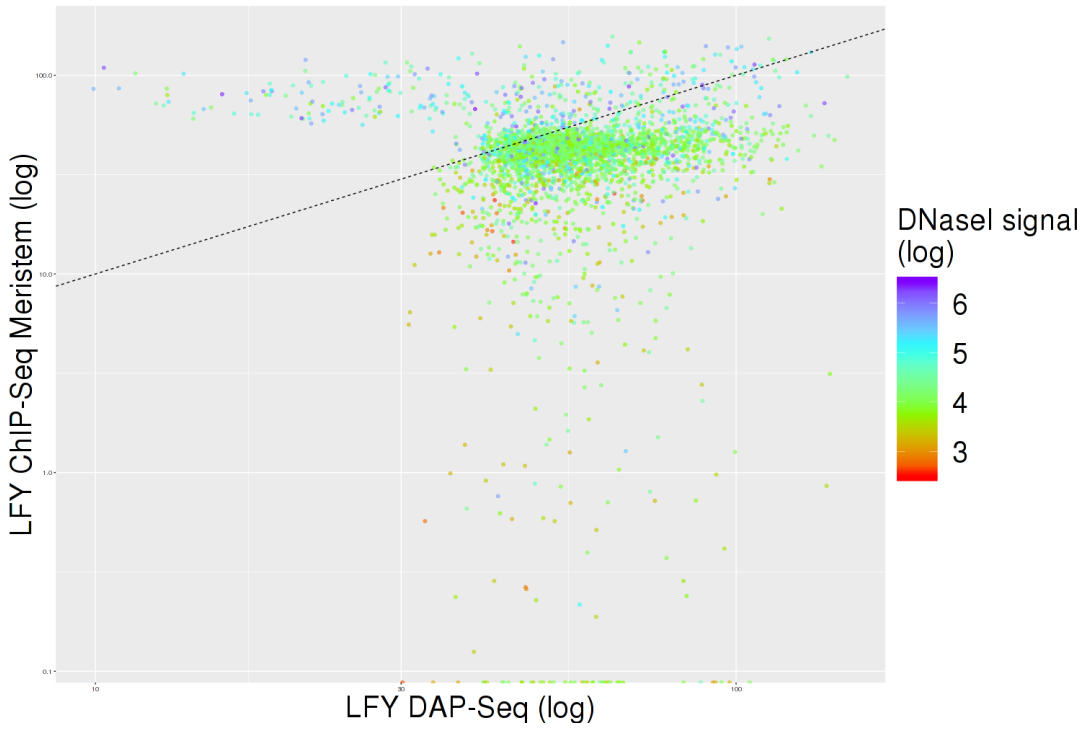
2.2.4 Déterminer les paramètres qui empêchent la liaison dans la cellule

2.2.4.1 Observation du signal DNaseI dans les régions liées

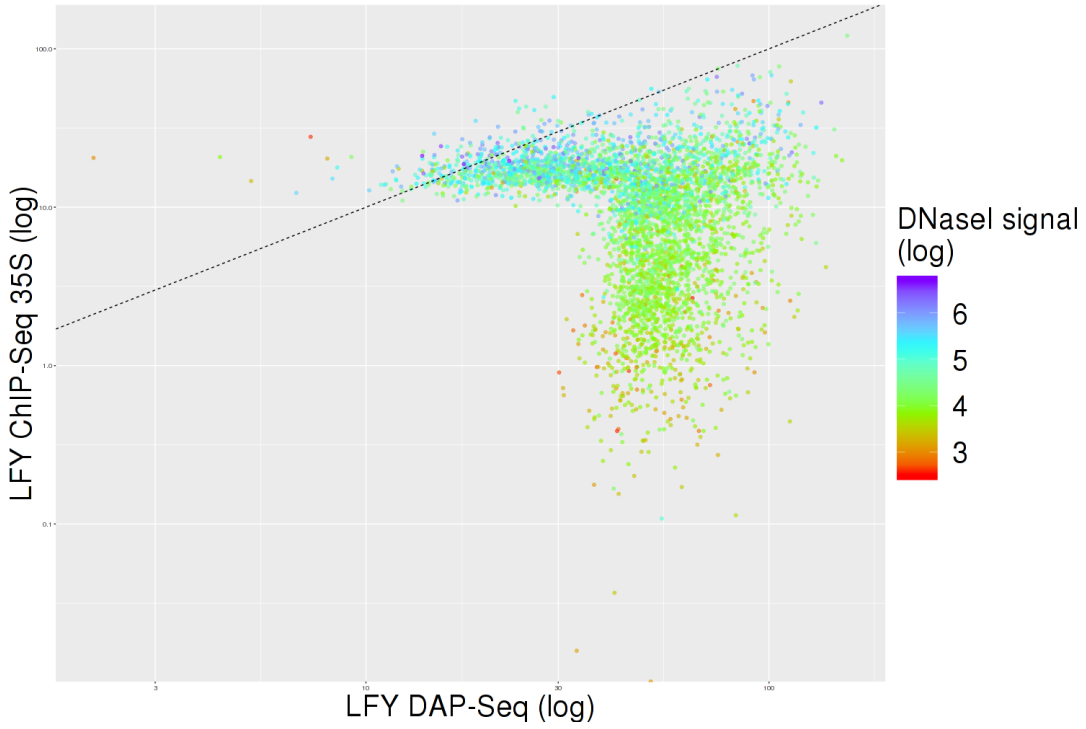
Alors que nous venons de voir que certains éléments présents dans la cellule peuvent favoriser la liaison de LFY, le grand nombre de régions liées exclusivement en DAP-Seq nous apprend qu'il existe *in vivo*, à l'inverse, des processus qui empêchent cette liaison. Dans cette partie, nous allons étudier ces éléments. À cette fin, nous nous focaliserons sur la sous partie des régions liées uniquement en DAP-Seq pour comparer leurs caractéristiques avec celles des régions liées en ChIP-Seq.

Dans les paragraphes I.2.1.2 et I.2.2, nous avons introduit les mécanismes épigénétiques qui peuvent altérer la liaison d'un TF sur l'ADN. Ces mécanismes conduisent à des degrés d'accessibilité plus ou moins importants de l'ADN et nous avons fait le choix d'en prendre la mesure. Dans ce but, nous disposions d'expériences de DNase-Seq réalisées dans des plantules.

La figure 37 montre que les régions uniquement liées en DAP-Seq ont une coloration de régions fermées. Cela suggère que les régions peu accessibles sont moins facilement liées. En même temps, cela montre qu'une grande partie des régions liées en ChIP-Seq peut être prédite à partir du DAP-Seq et du DNaseI-Seq.



(a) Pics de DAP-Seq et de ChIP-Seq en méristème



(b) Pics de DAP-Seq et de ChIP-Seq en plantule

FIGURE 37: La couverture normalisée sous les pics communs, uniques au DAP-Seq et uniques au ChIP-Seq est représentée sur axes x et y. La coloration donne la couverture normalisée du signal DNase-Seq sous les pics, dans des cellules de feuille.

2.2.4.2 Utiliser la DNaseI pour améliorer le modèle de liaison sur les régions liées en ChIP-Seq

Dans le début de cette partie (2.2.3), nous avons évalué la capacité de nos modèles à distinguer les régions non liées des régions liées en ChIP-Seq. Ici, nous avons souhaité voir si implémenter le signal DNaseI pouvait améliorer les prédictions.

Nous avons donc créé un modèle simple de la forme $score = score_{TFFM} + C * cov(DNaseI)$ où l'ouverture de la chromatine donne un bonus au score retourné par la TFFM. En ajustant le coefficient C sur $\frac{2}{3}$ des pics liés en ChIP-Seq d'après le critère de l'AUROC, puis en testant sur le tiers des pics restants, nous sommes parvenus à améliorer le modèle (voir table ci-dessous)

	ChIP-Seq méristème	ChIP-Seq plantule
AUROC TFFM	0.91	0.89
AUROC TFFM + DNaseI	0.94	0.92

2.2.4.3 Discussion

Dans cette partie, nous avons pu voir que l'état de la chromatine empêche l'accès de LFY à de nombreuses régions. En tenant compte de l'ouverture de la chromatine, le pouvoir prédictif de notre modèle (score TFFM et signal DNaseI) dans le ChIP-Seq du méristème atteint presque les performances du modèle basé sur les TFFM, en DAP-Seq.

Si la liaison est plutôt bien prédite par notre modèle, l'observation de la figure 37 nous montre qu'une partie des pics liés en ChIP-Seq ont un niveau de fermeture assez élevé (en vert sur le graphique), niveau sur-représenté dans les régions uniquement liées en DAP-Seq.

Ceci est en accord avec les observations faites par O'Malley et al. (2016) où les auteurs comparent les profils DAP-Seq et ChIP-Seq pour 5 TF différents. On peut y voir que suivant le TF, entre 15% et 70% des pics communs au ChIP-Seq et au Dap-Seq tombent dans sites sensibles à la DNaseI. Ainsi, combiner DAP-Seq et DNase-Seq ne suffit que partiellement à prédire les régions liées en ChIP-Seq dans le génome. On soulignera cependant le fait que le signal DNaseI ne fait que résumer l'état de la chromatine, il ne capture pas la complexité des très nombreuses marques chromatiniennes qui peuvent altérer le génome. Une approche plus juste serait sinon d'attribuer un état chromatinien à chaque région (Roudier et al., 2011) (voir paragraphe I.2.2.3) et de réaliser une étude similaire.

Étant donné la possible propriété pionnière de LFY (Sayou et al., 2016), l'étude de ces régions (fermées mais liées) et de leurs caractéristiques présente un intérêt certain. L'article de Sayou et al. (2016) montre que LFY peut former des oligomères *in vitro* grâce à son domaine *SAM*. Le ChIP-Seq de LFY_{TERE} (dont le domaine *SAM* est rendu inopérant) a montré que les régions liées les plus affectées par la perte de fonction du domaine *SAM* sont les régions liées les plus fermées. Il serait donc intéressant de comparer la densité de sites de liaison de LFY dans les régions liées fermées en ChIP-Seq et dans les régions d'un même niveau de fermeture liées uniquement en DAP-Seq. Enfin, puisque les TF pionniers sont capables de se lier sur les nucléosomes (Lai et al., 2018), l'étude de la position des nucléosomes sur les régions fermées et liées est une piste à suivre dans le futur.

2.3 Discussion du chapitre

Dans un premier temps, nous avons souhaité étudier les performances du modèle de liaison existant (PWM "model", figure 31.d) sur les régions liées par LFY en DAP-Seq. Ce modèle a d'abord été confronté à des PWM classiques, obtenues à partir des régions liées (figure 31.(b,c)), puis à une TFFM et enfin à un modèle prenant en compte la structure de la liaison et le score retourné par la TFFM (figure 32). Les performances de la PWM "model" ont été égalées par la PWM "LFY DAP pal" (retournée par meme-suite) et devancées par la TFFM. Ce résultat peut paraître surprenant : l'étude de Moyroud et al. (2011b) montre une grande corrélation entre les scores retournés par la PWM "model" et l'affinité de liaison de LFY. L'affinité d'un TF dépendant de l'ADN génomique utilisé (Galli et al., 2018), il est probable que le SELEX (utilisé pour produire la PWM "model") ne permette pas de capturer parfaitement la liaison de LFY dans le génome d'*A. thaliana*. Si les nouvelles dépendances observées entre les positions de la TFFM peuvent également provenir de ce biais, il serait possible de vérifier que les méthodes différentes pour obtenir les dépendances ne soit pas à l'origine de ces disparités ; les sites de liaison de DAP-Seq peuvent être mis en entrée du programme ENOLOGOS (Workman et al., 2005), utilisé pour produire la PWM "model". Enfin, nos analyses ont permis de montrer que des dépendances entre nucléotides adjacents suffisent à définir le modèle de liaison. En effet, le calcul de la structure de l'ADN est basé sur des pentamères et il n'ajoute rien au pouvoir prédictif du modèle.

Notre modèle de liaison nous a ensuite permis de confronter efficacement le DAP-Seq et le ChIP-Seq. D'abord, nous avons pu voir que beaucoup de régions sont exclusives du DAP-Seq et que quelques régions sont liées uniquement en ChIP-Seq. Nous avons d'abord choisi d'analyser cette dernière sous-partie, en supposant que des co-facteurs pourraient aider LFY à lier ces régions. Nous avons d'abord observé qu'elles avaient des sites de liaison de LFY d'affinités inférieures aux régions communément liées en DAP-Seq et en ChIP-Seq, suggérant que le contact entre l'ADN et LFY est direct même s'il devait être favorisé par des co-facteurs. Certains TF, comme les DOF, les TCP ou les protéines à boîte MADS se sont révélés être des co-facteurs potentiels de LFY.

Enfin, le grand nombre de régions liées uniquement par DAP-Seq nous a poussé à étudier l'ouverture de la chromatine de ces régions, celui-ci expliquant qu'une grande partie ne soit pas liée *in vivo* : il est ressorti que ces régions possèdent une chromatine très compacte, alors que son accessibilité est accrue dans les régions liées en ChIP-Seq.

Chapitre 3

Tétramérisation des facteurs de transcription à boîte MADS

Comme nous l'avons mentionné dans l'introduction (paragraphe I.4.4.2), des données génétiques (figure 23) suggèrent l'importance de la tétramérisation des TF à boîte MADS pour l'identité des organes floraux. Ces évidences indiquent que $SEP3_{del}$ se comporte différemment de SEP3 à l'échelle génomique. Dans cette section, nous chercherons à comprendre l'impact de la tétramérisation sur la spécificité des facteurs à boîte MADS en étudiant le quartet formé des deux dimères SEP3-AG. La description des CArGbox fait généralement appel au consensus et implémente parfois des séquences appelées A-tract (successions de A et de T entre CC et GG dans $CC(A_nT_m)_{n+m=6}GG$) (Muino et al., 2013) pour prendre en compte la sensibilité des TF à la structure de l'ADN. Nous nous proposons ici d'utiliser les PWM et les TFFM, pour décrire le site de liaison.

Pour réaliser cette étude, le DAP-Seq est une méthode particulièrement adaptée. En effet, la taille suffisante des fragments d'ADN permet d'étudier des configurations impossibles à observer dans le cadre du PBM ou du SELEX. De plus, utiliser l'ADN génomique d'*A. thaliana* est avisé dès lors que l'on souhaite établir un lien entre les gènes liés et ceux régulés.

Ici, l'idée est de comparer les DAP-Seq des complexes SEP3-AG et $SEP3_{del}$-AG, dont l'altération de l'interface de tétramérisation rend la formation du complexe $(SEP3_{del}\text{-AG})_2$ impossible. La réalisation de ces expériences présente une difficulté étant donné que la mise en présence des TF SEP3 et AG ne garantit pas nécessairement la formation du complexe SEP3-AG. En effet, l'homodimère SEP3 est capable de lier l'ADN et de former un homotétramère (Hugouvieux et al., 2018). Pour s'affranchir de cette situation, le DAP-Seq classique (figure 9) a été optimisé par un membre de l'équipe, Xuelei Lai. Si on exprime les TF SEP3 et AG avec des étiquettes différentes, une fois mis en présence de l'ADN d'intérêt, on peut d'abord récupérer les fragments liés par le premier TF. Au sein de ces fragments, l'étiquette différente sur le deuxième TF permet de n'extraire que la sous partie de ceux également liés par le celui-ci.

3.1 Pré-traitement des réplicats

3.1.1 Qualité des réplicats

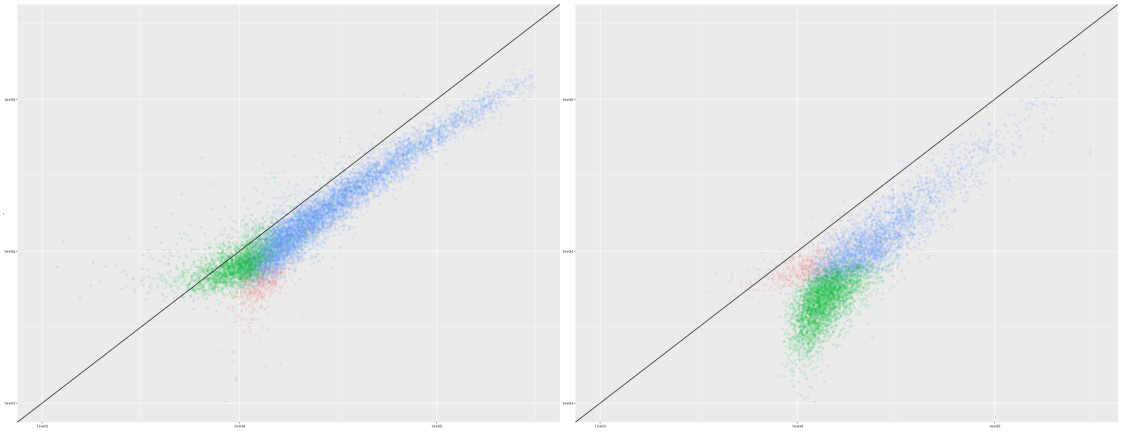
Les *reads* ont étés alignés comme détaillé dans le II.1.1. Trois réplicats ont été générés pour SEP3-AG et $SEP3_{del}$-AG. Ces réplicats ont été traités selon la méthode décrite dans le paragraphe II.1.2.1 et sont

comparés dans la figure 38. Les trois réplicats de SEP3-AG montrent une grande similarité et nous avons fait le choix de les garder tous (figure 38, a, c, d). Par contre, le réplicat 1 de SEP3$_{del}$ semble s'éloigner des réplicats 2 et 3 (figure 38, b et d), nous l'avons donc éliminé. On explique ce choix en considérant que la proximité des réplicats 2 et 3 ne peut pas être due au hasard. Par conséquent, les différences visibles avec le réplicat 1 laissent penser que celui-ci n'a pas bien fonctionné.

3.1.2 Choix des pics

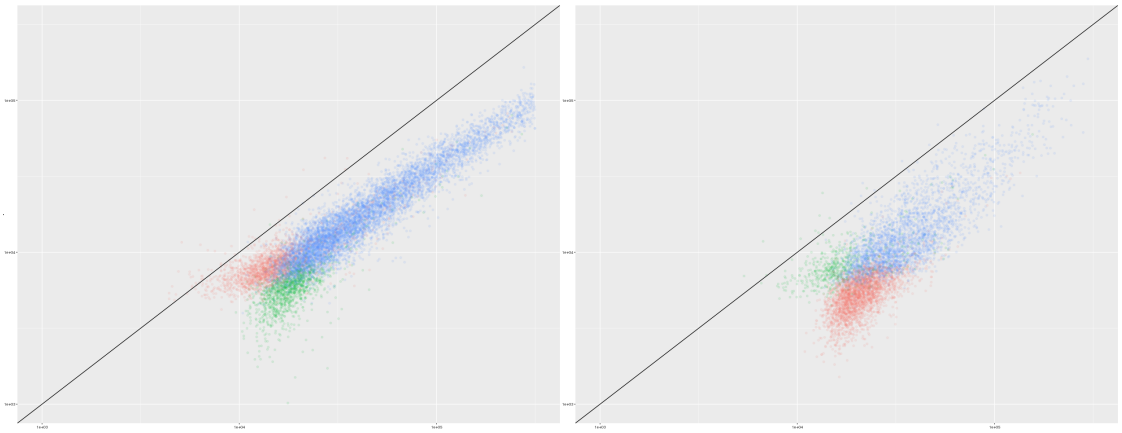
Nous avons choisi les pics d'après la méthode décrite dans le paragraphe II.1.2.2. Les réplicats 2 et 3 (figure 38, e, f) ont été analysés pour SEP3-AG et SEP3$_{del}$-AG. La figure 39 montre les pics avant et après l'application du filtrage IDR. La table ci-dessous donne le nombre de pics restants après le filtrage.

Pics SEP3-AG	Pics SEP3$_{del}$-AG
4072	2285



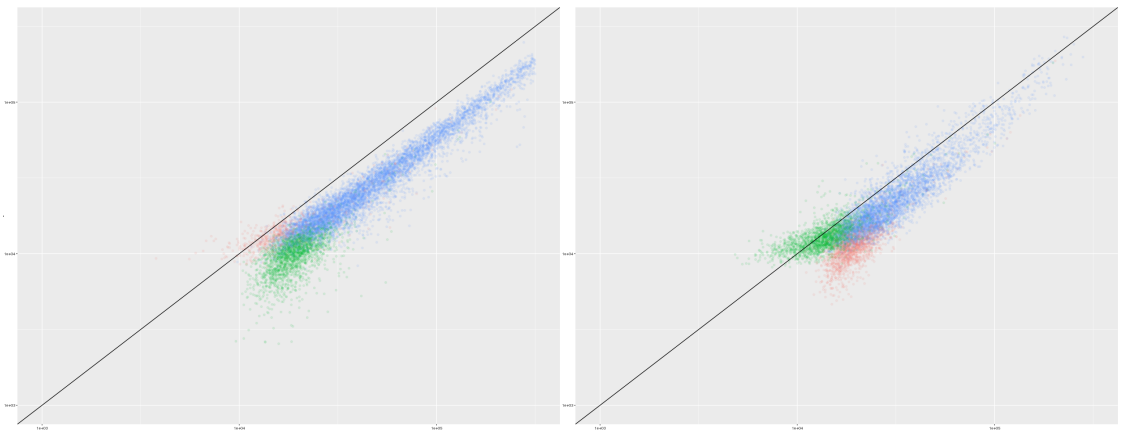
(a) SEP3-AG, réplicats 1 et 2

(b) SEP3$_{del}$-AG, réplicats 1 et 2



(c) SEP3-AG, réplicats 1 et 3

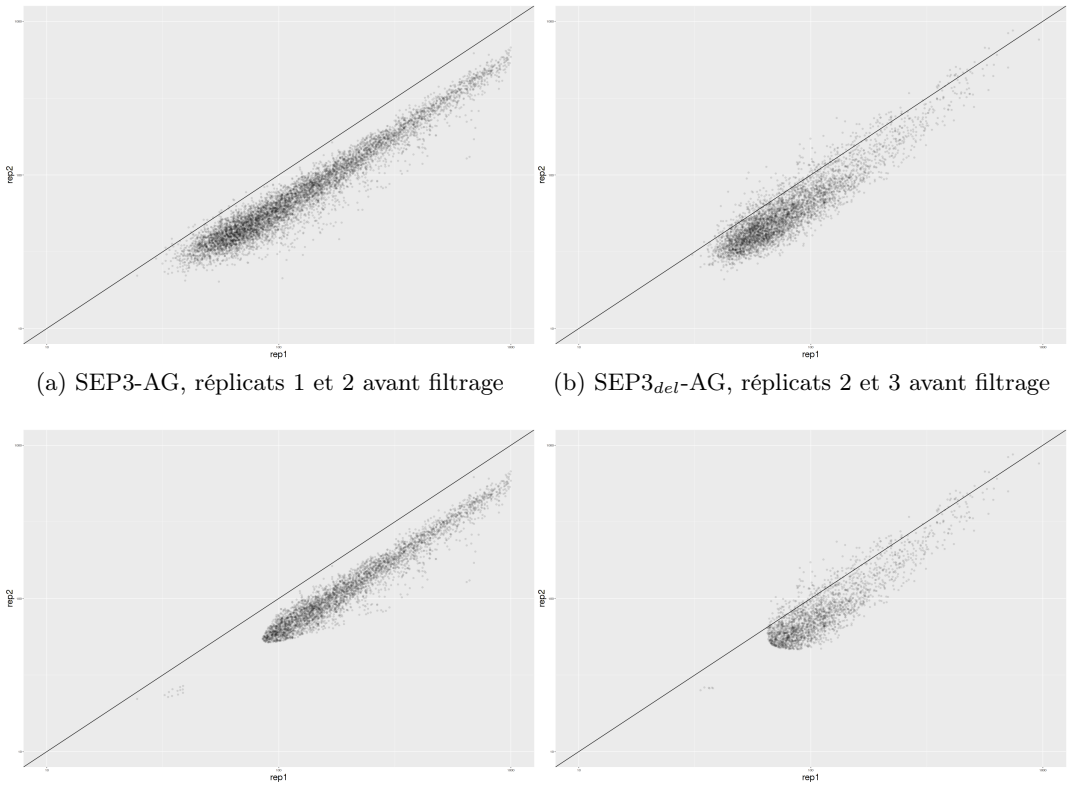
(d) SEP3$_{del}$-AG, réplicats 1 et 3



(e) SEP3-AG, réplicats 2 et 3

(f) SEP3$_{del}$-AG, réplicats 2 et 3

FIGURE 38: Comparaison des réplicats de SEP3AG (a, c, d) et de SEP3$_{del}$-AG (b, d, f). Chaque point représente un pic (commun à deux réplicats en bleu ou appartenant à l'un ou à l'autre en rouge ou en vert). Les axes x et y donnent la couverture normalisée sous sous chaque pic dans les différents réplicats.



(a) SEP3-AG, réplicats 1 et 2 avant filtrage

(b) SEP3$_{del}$-AG, réplicats 2 et 3 avant filtrage

(c) SEP3-AG, réplicats 1 et 2 après filtrage

(d) SEP3$_{del}$-AG, réplicats 2 et 3 après filtrage

FIGURE 39: Comparaison des réplicats de SEP3AG (a, c) et de SEP3$_{del}$-AG (b, d) avant filtrage (a, b) et après filtrage (c, d). Chaque point représente un pic. Les axes x et y donnent la couverture normalisée sous sous chaque pic dans les différents réplicats.

3.2 Expliquer la spécificité de liaison de SEP3-AG et SEP3$_{del}$-AG

3.2.1 SEP3-AG et SEP3$_{del}$-AG ne lient pas les mêmes régions

Comme décrit dans le paragraphe II.1.2.3, nous avons concaténé les *reads* des réplicats (1, 2, 3) et (2, 3) des DAP-Seq respectifs SEP3-AG et SEP3$_{del}$-AG.

En premier lieu, nous avons constaté que les deux complexes ne lient pas les mêmes régions (figure 40). Les pics uniques et communs sont déterminés selon l'algorithme de la figure 24. La table suivante récapitule la sortie du programme.

Pics SEP3-AG	Pics communs	Pics SEP3$_{del}$-AG
2524	1548	748

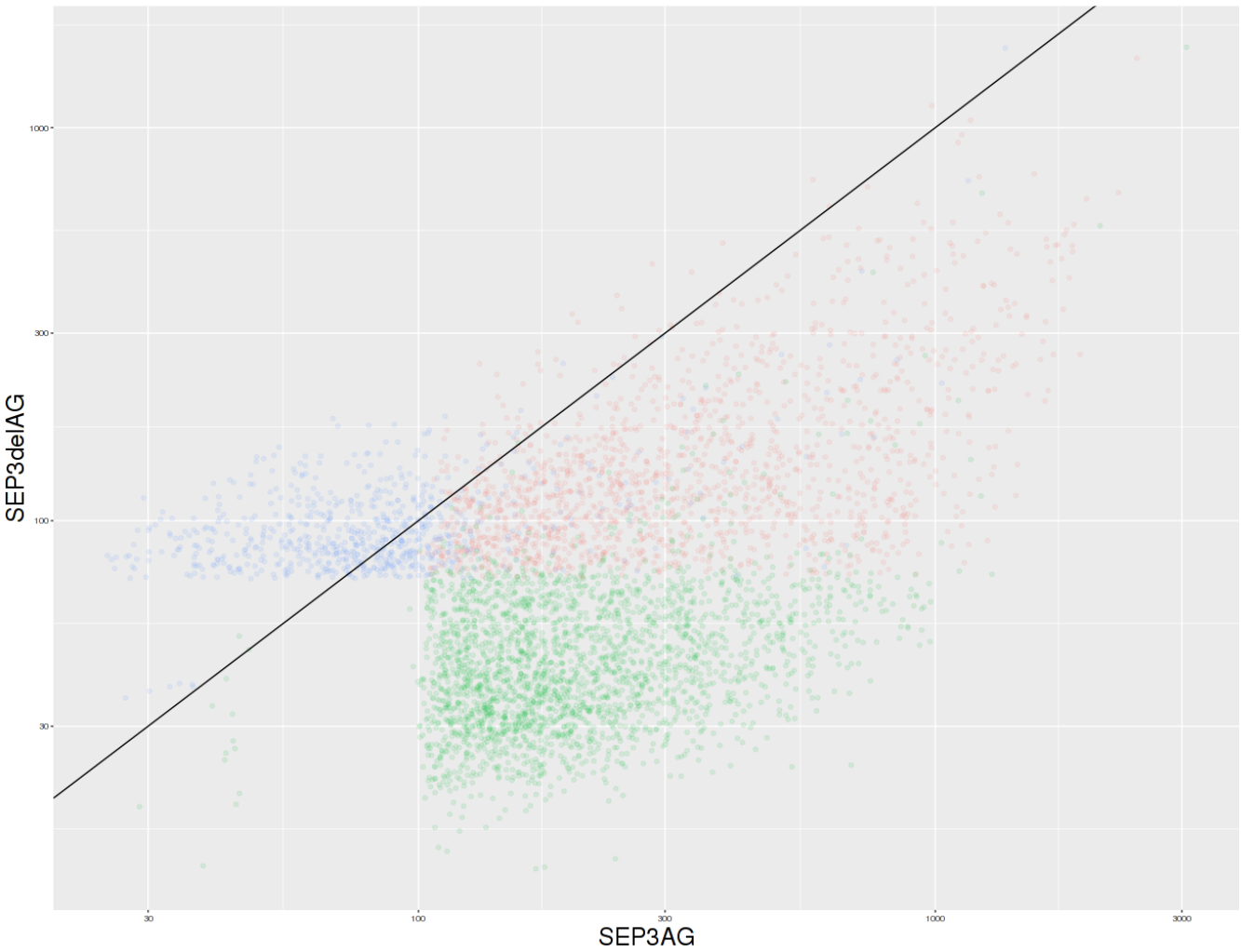


FIGURE 40: Les couleurs vert et bleu donnent respectivement les pics uniques à SEP3$_{del}$-AG et à SEP3-AG. Les pics communs sont marqués en rouge. Les axes x et y indiquent la couverture normalisée sous les pics.

3.2.2 Prédire la liaison des dimères

3.2.2.1 Prédire la liaison à l'aide des PWM

Comme décrit dans le paragraphe II.3.1.3, nous avons déterminé des PWM pour SEP3-AG et SEP3$_{del}$-AG. Ces PWM sont bien des CArGbox, les sites de liaison des TF à boîte MADS, ce qui donne une indication sur la réussite du DAP-Seq. Celles-ci ont été testées par AUROC suivant la méthode du paragraphe II.3.2 (figure 41). Deux résultats semblent ressortir. D'abord, les couples de PWM (SEP3AG, SEP3$_{del}$AG) et (SEP3AG pal, SEP3$_{del}$AG pal) donnent des AUROC très similaires. Cela tend à suggérer que la spécificité des deux complexes (SEP3-AG et SEP3$_{del}$-AG) est la même. Ensuite, en forçant la palindromie du motif (conformément à la structure du domaine de liaison à l'ADN), les performances du modèle augmentent considérablement.

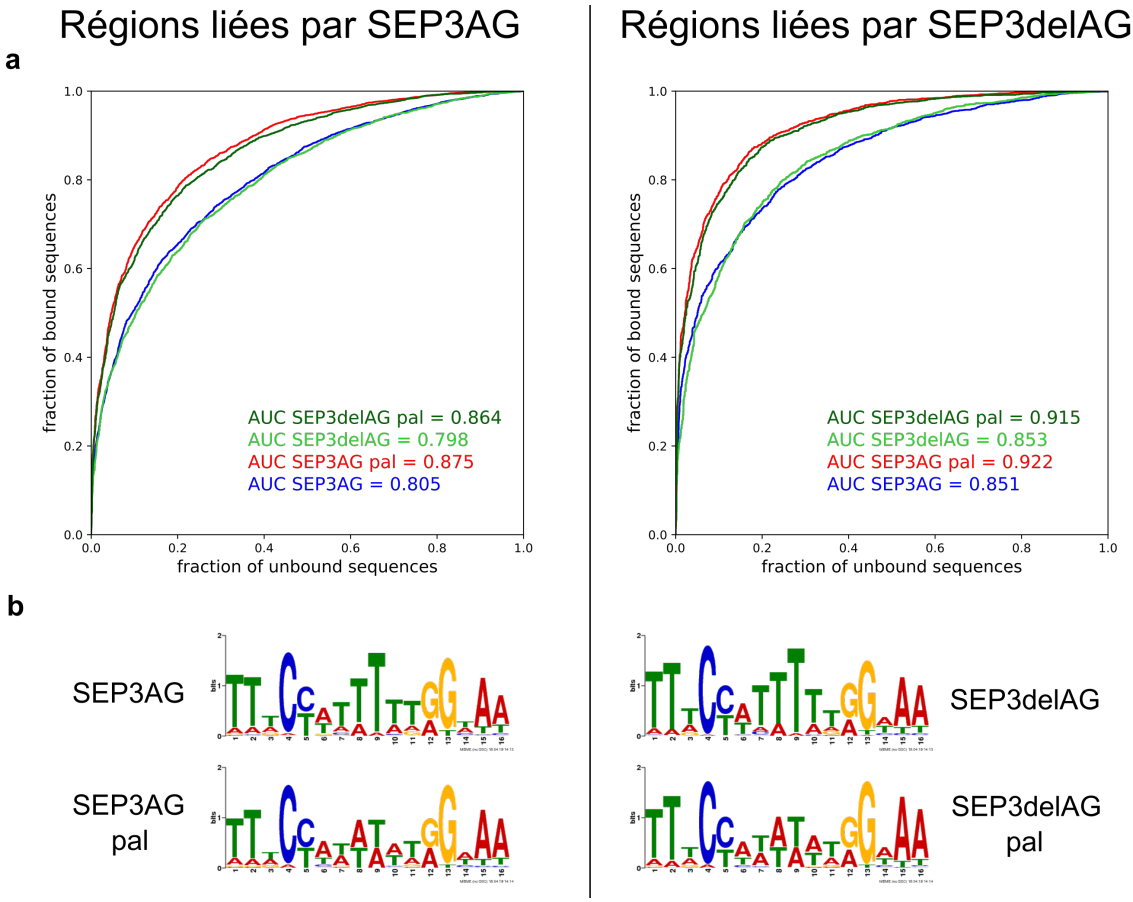


FIGURE 41: (a) Test des quatre motifs (b) sur les régions liées par SEP3-AG à gauche et sur les régions liées par $SEP3_{del} - AG$ à droite. (b) Les motifs de gauche et de droite ont respectivement été déterminés sur les régions liées par SEP3-AG et par $SEP3_{del}$-AG. Les motifs SEP3AG pal et $SEP3_{del}$AG pal sont palindromiques

3.2.2.2 Prédire la liaison à l'aide des TFFM

Comme nous l'avons mentionné dans l'introduction (figure 13, Mathelier et al. (2016)), les sites de liaison des facteurs de transcription à boîte MADS peuvent être mieux modélisés par des outils qui prennent en compte les dépendances entre nucléotides, la structure de l'ADN entre les bases CC et GG du motif étant plus ou moins favorable à la liaison des TF. Pour ces raisons, nous avons fait le choix de continuer notre étude en remplaçant les PWM par des TFFM. Celles-ci ont été déterminées selon la méthode du paragraphe II.3.1.2 à partir de la PWM palindromique SEP3-AG, qui obtient les meilleures performances sur les DAP-Seq de $SEP3_{del}$-AG et de SEP3-AG. Les TFFM respectivement déterminées sur les régions liées par SEP3-AG et sur celles liées par $SEP3_{del}$-AG sont quasiment identiques (figure 42,b) et ont des performances similaires (figure 42,a), aboutissant aux mêmes conclusions que le paragraphe précédent ; altérer le domaine de tétramérisation ne modifie pas la spécificité du dimère SEP3-AG. Cependant, on note le pouvoir prédictif très supérieur des TFFM par rapport à celui des PWM, avec d'excellentes performances sur les régions liées par $SEP3_{del}$-AG). Comme dans la figure 13, les logos (figure 42,b) montrent des dépendances entre les positions 4 et 11 des TFFM. En effet, les successions de T ou de A sont respectivement préférées aux transitions T-A ou A-T.

À la fois pour les PWM comme pour les TFFM, les modèles prédictifs de la liaison sont plus efficaces sur les régions liées par SEP3$_{del}$-AG, pour lesquelles on peut supposer que toute l'information de la liaison se situe dans les logos. À l'inverse, expliquer la liaison du tétramère (SEP3-AG)$_2$ nécessite peut-être l'existence d'un deuxième site situé non loin du premier. À partir d'ici, nous avons fait le choix de n'utiliser plus que la TFFM obtenue sur les régions liées par SEP3-AG.

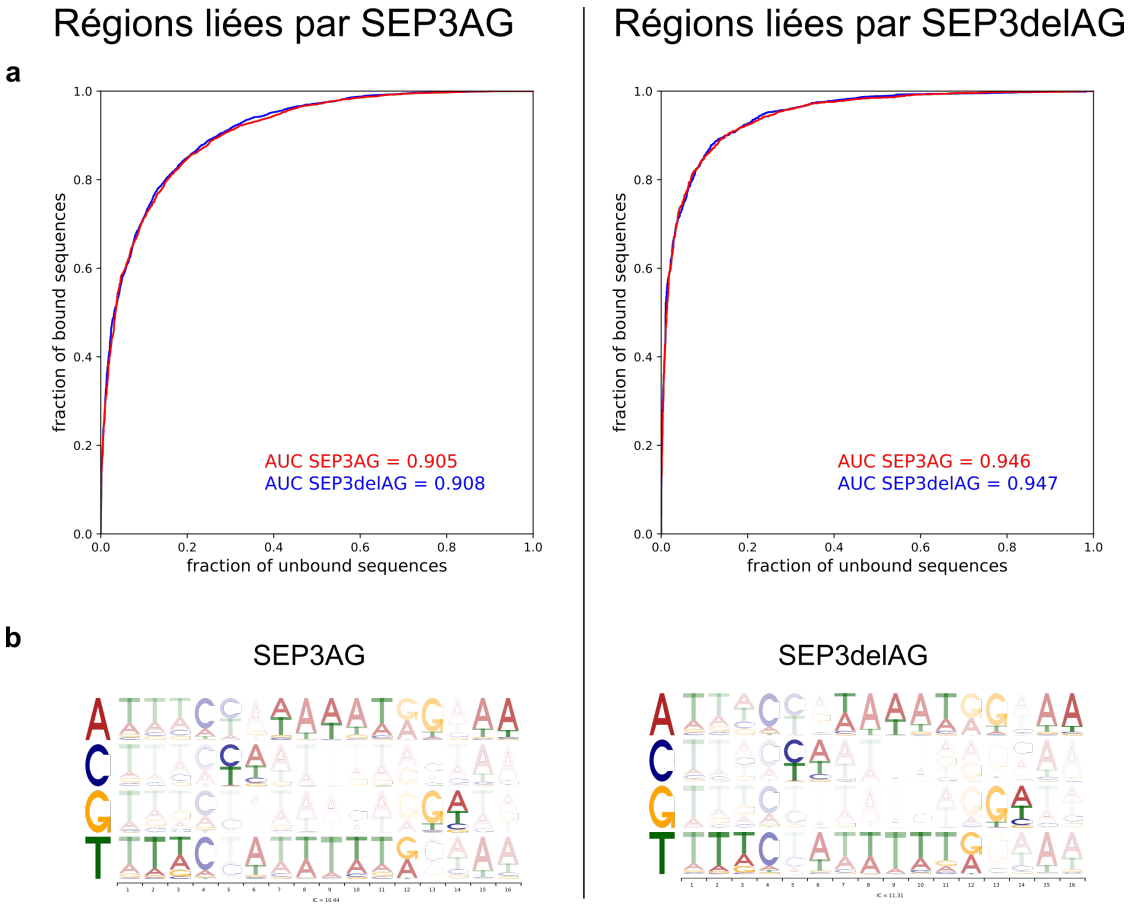


FIGURE 42: (a) Test des 2 TFFM sur les régions liées par SEP3-AG à gauche et sur les régions liées par SEP3$_{del}$-AG à droite. (b) Les motifs de gauche et de droite ont respectivement été déterminés sur les régions liées par SEP3-AG et par SEP3$_{del}$-AG.

3.2.3 Expliquer les spécificités différentes de SEP3-AG et SEP3$_{del}$-AG

Alors que les modèles prédictifs semblent indiquer l'identité des sites de liaison dimériques de SEP3-AG et de SEP3$_{del}$-AG (figures 42, 41), les régions liées par ces deux duos de TF sont très différentes (figure 40). Il faut donc que cette spécificité soit expliquée par un autre paramètre que l'affinité des dimères à l'ADN. Ce paramètre peut être par exemple une densité de sites de liaison supérieure dans les régions liées par SEP3-AG, qui permettrait aux TF de former des quartets (figure 22). Supposant que le modèle des quartets impose des contraintes sur de la courbure de l'ADN, nous avons poussé ce raisonnement en cherchant si la liaison du tétramère (SEP3-AG)$_2$ était favorisée par des espacements précis entre les deux sites de liaison des dimères. À cette fin, nous avons utilisé le script et les méthodes du chapitre 1 (`https://github.com/Bioinfo-LPCV-RDF/get_interdistances`).

Le script a été mis à jour pour accepter les TFFM. Il convient ici de noter que contrairement aux PWM, les TFFM ne peuvent pas être symétriques : parce ce qu'elles sont basées sur un modèle de Markov, l'ordre des nucléotides observé au niveau du site influe sur le score donné. Par conséquent, elles ne peuvent pas donner des scores identiques sur les deux brins de l'ADN. En théorie, il faudrait donc définir un sens à la matrice et donc aux configurations que l'on peut obtenir (comme dans la figure 29). Cependant, il apparaît que sans être identiques, les scores donnés par la TFFM de SEP3AG sur chaque brin sont très similaires. Pour cette raison, les configurations *Everted repeat*, *Direct repeat* et *Inverted repeat* observées donnent des résultats très proches. Nous avons donc décidé de ne pas donner un sens à la matrice et avons donc choisi de ne représenter qu'un type de configuration.

Les espacements entre sites de liaison sont définis comme le nombre de bases (**N** en gras) entre deux CArGbox (en italique) :

CCNNNNNNGG **NNNNNNNNNN** *CCNNNNNNGG*

3.2.3.1 Observe-t-on les mêmes préférences d'espacement dans les deux sets de régions liées ?

Trois sets de régions non liées (Cf. paragraphe II.3.2) ont été respectivement produits pour SEP3$_{del}$-AG et SEP3-AG. L'enrichissement pour chaque configuration [1] dans les sets de régions liées par rapport aux sets de régions non liées a ensuite été calculé pour différents seuils sur les scores renvoyés par la TFFM.

La figure 43 montre que les préférences des dimères SEP3-AG et SEP3$_{del}$-AG sont différentes, ce qui est en accord avec les différences de spécificités observées pour les deux complexes (figure 40). Ainsi, On constate que les distances autour de 26, 36 et 46 sont exclusivement préférées par SEP3-AG. De manière surprenante, on constate que le dimère SEP3$_{del}$-AG a une préférence également exclusive pour deux sites séparés de 10 nucléotides (ce résultat est résumé en figure 44). Il semble que la déplétion notable entre les distances 0 et 6, présente dans les deux panneaux, soit due au chevauchement entre les deux sites (la TFFM donne un score sur 16 nucléotide, pas seulement sur la CArGbox $CCA_nT_mGG_{n+m=6}$).

Les préférences des deux dimères sont difficiles à interpréter. En effet, pour SEP3-AG, si les sites situés au niveau des espacements 26, 36 et 46 sont à peu près sur la même face de la double hélice d'ADN, ils ne correspondent pas à des tours d'hélice complets, ce qui empêche à priori la formation du quartet (SEP3-AG)$_2$. La préférence notable de SEP3$_{del}$-AG pour l'espacement de 10 nucléotides ne trouve pas d'explication convenable (figure 44).

1. Tous les sites dont le score est supérieur à un seuil sont retenus. On regarde ensuite quels sont les espacements entre les sites 2 à 2

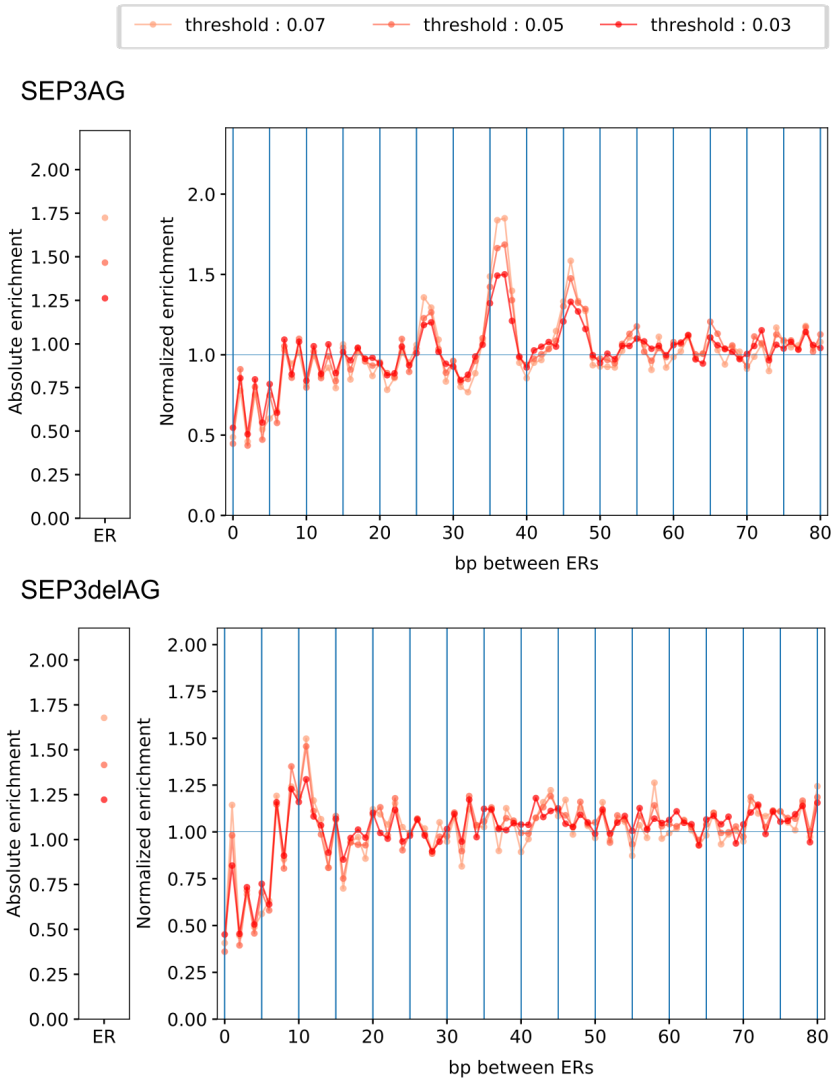


FIGURE 43: Enrichissement en configurations dans les régions liées par SEP3-AG et par SEP3$_{del}$-AG. L'*Absolute enrichment* donne le niveau de la sur/sous-représentation des configurations dans les régions liées par rapport aux régions non liées. Le *Normalized Enrichment* donne la même information pour chaque espacement mais a été normalisé tel que la moyenne de tous les *Normalized Enrichment* soit égale à 1. Des informations supplémentaires sur le calcul de ces deux grandeurs sont données dans les méthodes de l'article présenté dans le chapitre 1. Ces grandeurs sont calculées pour 3 différents seuils sur les scores renvoyés par la TFFM.

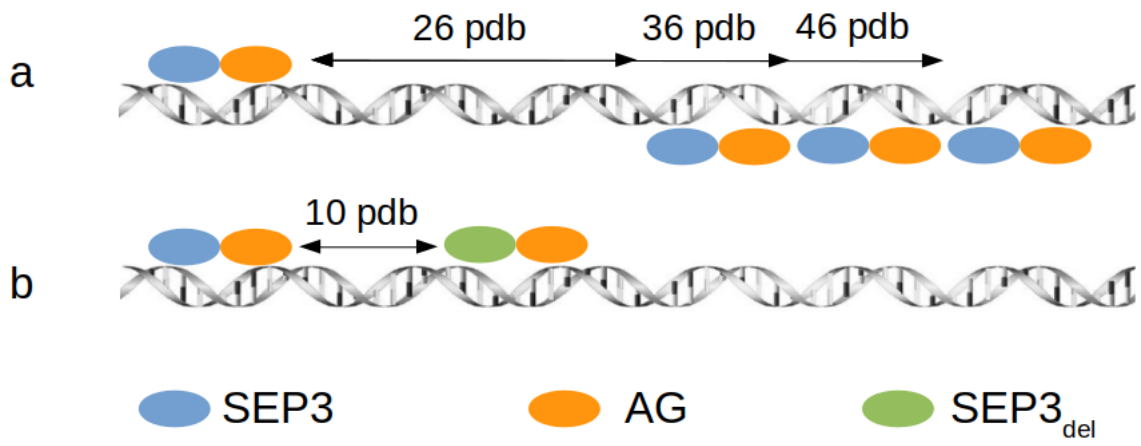


FIGURE 44: Préférences d'espacement de deux dimères SEP3-AG (a) et de deux dimères SEP3$_{del}$-AG (b) en paires de bases (pdb). Alors que les sites de deux dimères SEP3-AG sont situés de part et d'autre de l'hélice d'ADN, les deux dimères SEP3$_{del}$-AG se placent du même côté de l'hélice.

3.2.3.2 Ces préférences d'espacements expliquent-elles l'éclatement de la figure 40 ?

Afin de confirmer nos résultats, nous avons calculé la réduction de couverture (CFR) induite par la perte du domaine de tétramérisation. On définit donc le CFR comme

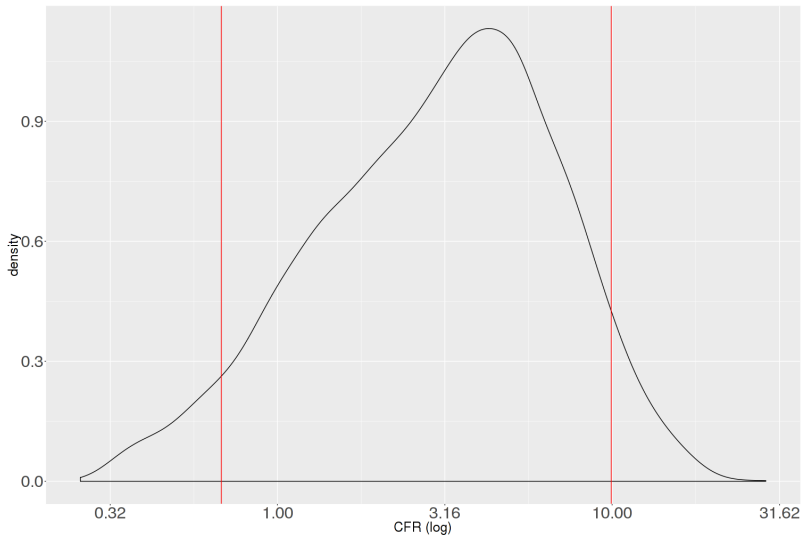
$$CFR = \frac{cov(SEP3 - AG)}{cov(SEP3_{del} - AG)}$$

Le CFR est calculé au niveau des pics montrés en figure 40, *ie.* sous les régions liées par SEP3-AG et/ou SEP3$_{del}$-AG. Ainsi, les régions telles que $CFR > 1$ sont mieux liées par SEP3-AG alors que $CFR < 1$ donne les régions mieux liées par SEP3$_{del}$-AG. À partir du CFR, on classe nos pics en 20 fractiles, les premiers et derniers (ayant respectivement un fort et un faible CFR) serviront comme base de comparaison (figure 45,a). La figure (figure 45,b) montre l'enrichissement en configurations dans le premier fractile par rapport au dernier.

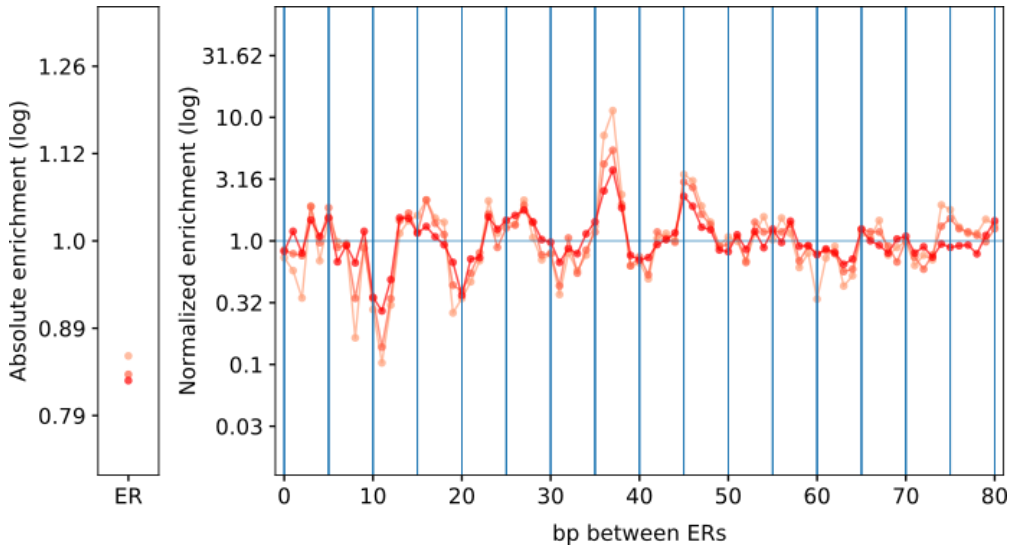
La figure 45.b accentue les résultats de la figure 43. En effet, on constate que pour le seuil 0.07, les enrichissements 36 et 11 sont respectivement 10 fois plus et 10 fois moins présents dans les régions les plus exclusivement liées par par SEP3-AG et SEP3$_{del}$-AG.

3.2.4 Bilan et discussion

Ici, nous avons en partie répondu à la question posée dans les objectifs de la thèse : nous avons constaté que les TF à boîte MADS capables de former des tétramères ont une spécificité différente de ceux qui en sont incapables. Nous avons également pu expliquer ces différences en constatant des préférences d'espacements entre 2 CArGbox distinctes pour les TF capables et incapables de former des tétramères. Ainsi, la capacité à former des tétramères confère une spécificité au quartet (SEP3-AG)$_2$. Il convient cependant de présenter les limites de notre analyse. D'abord, si elle suggère que les spécificités différentes des quartets sont dues à des préférences d'espacement distinctes, la simple étude du quartet (SEP3-AG)$_2$ ne permet pas de l'affirmer. Ensuite, les analyses *in vitro* ne permettent pas de conclure sur le comportement des TF dans la cellule.



(a) Distribution des CFR. Les deux barres rouges montrent le premier et le dernier des 20 fractiles.



(b) Enrichissement en configuration du premier fractile par rapport au dernier fractile. Les seuils utilisés sont les mêmes que dans la figure 43.

FIGURE 45: Comparaison des régions les plus différentiellement liées par SEP3-AG et SEP3$_{del}$-AG

3.3 Spécificité des gènes à boîte MADS *in vivo*

Étant donné ces observations, nous nous sommes questionnés sur l'importance de ces espacements dans la cellule. En effet, la liaison d'un TF *in vitro* donne simplement une indication sur les préférences de liaison du TF, sans permettre d'affirmer que des régions liées *in vivo* disposent de ces caractéristiques. Et ces préférences de liaison ne permettent pas de conclure sur la capacité d'un TF à réguler ou non un gène lié. Ainsi, comme nous l'avons observé dans le chapitre 1, les configurations observées sur les gènes régulés ne sont pas les mêmes que celles sur les gènes liés.

Dans la mesure du possible avec les moyens à notre disposition, nous avons donc cherché à répondre

à ces trois questions :

1. Les spécificités d'espacement de SEP3-AG *in vivo* sont elles observables *in vitro* ?
2. Cette spécificité d'espacement explique-t-elle la spécificité des différents quartets ?
3. Observons-nous des spécificités d'espacement particulières sur les gènes régulés par SEP3-AG ?

3.3.1 Spécificité de liaison de SEP3-AG

Pour réaliser cette étude, nous nous sommes aidés de 3 jeux de donnés :

— Un ChIPSeq de SEP3 dans des inflorescences sauvages ou dans le mutant *ag* (Kaufmann et al., 2009).
— Un ChIPSeq de SEP3 dans des inflorescences (Pajoro et al., 2014).
— Un ChIPSeq de AG dans des inflorescences (Ó'Maoiléidigh et al., 2013)

Pour chacun des ChIP-Seq les données brutes ont été analysées selon la méthode du II.2.

Dans le premier ChIPSeq, la protéine SEP3 étant capable de former des complexes avec de nombreux TF, la sous partie des régions exclusivement liées dans le sauvage (et pas dans le mutant *ag*) devrait nous donner accès aux régions liées par SEP3-AG, et ainsi répondre à la première question. En comparant avec les régions exclusivement liées dans le mutant *ag*, nous devrions être en mesure de répondre à la deuxième question (la méthodologie est donnée dans la figure 46)

FIGURE 46: Méthode pour déterminer les régions liées par SEP3-AG à partir des ChIP-Seq réalisés par Kaufmann et al. (2009)

À partir de ces observations, nous avons fait 2 groupes :

— Les régions liés dans le mutant *ag* (5867 pics)
— Les régions uniquement liées par SEP3 chez la plante sauvage (455 pics)

À partir des deux groupes, nous avons reproduit les analyses d'espacement du paragraphe 3.2.3.1. Les figures ne montrant pas de préférences particulières dans les deux groupes de régions elles ne sont pas présentées ici.

Nous avons ensuite analysé les données ChIP-Seq de AG publiées par Ó'Maoiléidigh et al. (2013). Rien de notable n'a pu être observé sur la totalité des régions. En sélectionnant la sous partie également liée en DAP-Seq par SEP3-AG, on obtient les mêmes enrichissements qu'en prenant la totalité des régions liées en DAP-Seq par SEP3-AG.

La recherche de distances privilégiées entre CArGbox a été reproduite sur les données de Pajoro et al. (2014). Ici, nous souhaitions voir si les enrichissements observés en DAP-Seq coexistaient avec d'autres

distances, qui auraient pu indiquer des préférences différentes pour les autres quartets. Ces données n'ont elles non plus pas permis d'observer des espacements privilégiés.

3.3.2 Bilan et discussions

Ainsi, nous n'avons pas été capables d'observer de caractéristiques particulières dans l'ensemble des régions liées *in vivo*. Les données traitées n'ont pas permis non plus de mettre en évidence d'autres enrichissements qui auraient pu attester de spécificités différentes pour d'autres quartets de gènes à boîte MADS. Si notre modèle ne permettait pas de discriminer aussi bien les régions liées en ChIP-Seq que les régions liées en DAP-Seq, l'aire sous la courbe ROC dans les ChIP-Seq ($\simeq 0.7$) confirme bien la présence de CArGbox. Cela indique que l'absence de configuration observée n'est pas du à l'utilisation de la TFFM construite sur les régions liées en DAP-Seq par SEP3-AG.

Il reste que nous pouvons nuancer ces résultats en discutant des facteurs qui ont pu nous empêcher d'observer *in vivo* les configurations mises en évidences par DAP-Seq.

En effet, dans le premier jeu de données (Kaufmann et al., 2009), l'absence de réplicat pour le ChIPSeq dans le mutant *ag* nous a peut-être conduit à analyser des données peu fiables. De plus, les régions liées par SEP3 dans le mutant *ag* ne sont pas nécessairement des cibles de AG dans la plante sauvage. En effet, l'étude de ces régions a montré que la couverture normalisée du signal ChIP-Seq y était très faible. Ceci suggère que ce sont les pics les moins fiables, le filtrage IDR supprimant préférentiellement les pics avec la couverture normalisée la plus faible. Ensuite, la mutation *ag* a pu entrainer des modifications du contexte chromatinien ou du protéome capable d'interagir avec SEP3.

Dans le deuxième jeu de données, le signal des régions liées par SEP3-AG a pu être noyé dans celui des régions co-liées par SEP3 et d'autres TF. Le nombre de pics n'était pas suffisant pour que nous puissions croiser avec les données de DAP-Seq, afin de n'extraire que les régions dont nous sommes sûrs qu'elles contiennent un site de liaison pour SEP3-AG.

L'absence de configuration remarquable dans le ChIP-Seq de Ó'Maoiléidigh et al. (2013) est en accord avec le fait que peu de régions sont partagées entre le DAP-Seq de SEP3-AG et le ChIP-Seq de AG. En effet, la table ci-dessous détaille la répartition des pics. Notons que dans les 349 pics communs au DAP-Seq et au ChIP-Seq, nous avons été capables de retrouver les configurations observées dans la figure 43. Ceci montre que le ChIP-Seq de AG contient d'autres cibles que celles du complexe $(\text{SEP3-AG})_2$. En effet, le possible tétramère contrôlant la formation des étamines (SEP3-AG-PI-AP3) aura pu être capturé.

Pics SEP3-AG (DAP-Seq)	Pics communs	Pics AG (ChIP-Seq)
3723	349	1072

Le nombre de pics propres au DAP-Seq est attribuable à l'état de la chromatine et à d'autres TF qui agissent en compétition avec SEP3-AG. L'étude comparative du ChIP-Seq et du DAP-Seq introduisant les états chromatiniens serait une idée réalisable de suite pour expliquer ces différences.

3.3.3 Spécificité dans les gènes liés régulées par AG

En ChIP-Seq, peu de pics sont généralement associés à des modifications du transcriptome. Comme nous l'avons montré dans l'article du chapitre1, des modifications du niveau de transcription des gènes peuvent être associés à des configurations de liaison particulières. Ici, nous chercherons à répondre à la

troisième question posée en début de section, à savoir, si on observe des spécificités d'espacement particulières sur les gènes régulés par SEP3-AG. À cette fin, nous nous sommes aidés des données d'expression publiées par Ó'Maoiléidigh et al. (2013), ces données rassemblant les gènes directement ou indirectement régulés par AG.

En premier lieu nous avons réalisé l'étude sur des promoteurs de 1000 paires de bases s'arrêtant au $5'UTR$ de chaque gène. Afin de rechercher des espacements privilégiés, nous avons fait le choix de comparer les promoteurs des gènes régulés à ceux des gènes non régulés. Cette étude n'a montré aucun espacement privilégié.

Ceci n'est pas vraiment surprenant : les données contiennent des gènes indirectement régulés, qui ne sont dont pas nécessairement liées par AG. De plus, les promoteurs sont de grande taille et un signal quelconque peut être noyé dans le bruit.

En croisant avant le ChIP-Seq, nous pouvons obtenir la sous partie des gènes à la fois liés et régulés par AG, resserrant en même temps les promoteurs autour du site de liaison (ce qui devrait diminuer le bruit). L'ensemble des promoteurs de gènes non régulés a à nouveau été utilisé comme base de comparaison. À nouveau cette étude n'a montré aucun espacement privilégié.

3.4 Discussion du chapitre

3.4.1 Importance de la tétramérisation

Dans ce chapitre nous avons montré que la capacité de SEP3-AG à former des tétramères joue un rôle important sur les préférences du complexe. Malgré le phénotype visible du mutant SEP3$_d$*el*-AG (figure 23), nous n'avons pas été capables d'observer ces préférences *in vivo*. Compte tenu du petit nombre de pics partagés par le ChIP-Seq et le DAP-Seq (paragraphe 3.3.2), et compte tenu du fait que ces pics ne montrent aucune sur-représentation d'une configuration particulière, il est possible que la tétramérisation ne concerne qu'une très faible sous partie des gènes régulés par AG.

Il est également possible que nos modèles soient incapables de capturer cette spécificité. L'étude de Smaczniak et al. (2017) montre qu'à l'aide de K-mer, une méthode différente des PWM classiques, des espacements différents entre CArGbox pour les dimères SEP3-AG, SEP3-AP1 et SEP3-SEP3 sont observables *in vivo*. Il est intéressant de noter que leurs analyses du dimère SEP3-AG leur a permis de trouver un enrichissement pour un espacement autour de 35 paires de bases, que nous observons aussi.

Les données de Hugouvieux et al. (2018) suggérant que le gène *CRABS CLAW* (*CRC*) est différentiellement exprimé lorsque le mutant *sep1,sep2,sep3* est complémenté par SEP3AG et SEP3$_d$*el*-AG, nous avons cherché une des configurations observées en DAP-Seq dans la région du promoteur liée co-liée par SEP3 et AG (Hugouvieux et al., 2018). Nous avons été capables de mettre en évidence deux sites de liaison distants de 26 paires de base ce qui correspond au sommet d'un pic sur la figure 43.

Néanmoins, puisque la tétramérisation semble avoir peu d'importance *in vivo* (elle semble affecter un poignée de gènes seulement et les phénotypes obtenus par Hugouvieux et al. (2018) sont faibles), il est difficile de comprendre pourquoi le domaine de tétramérisation est conservé dans tous les TF à boîte MADS. Je pense au contraire que la tétramérisation a une fonction mais que nos analyses n'ont pas permis de la saisir.

3.4.2 À propos des outils bioinformatiques

Compte tenu de leur simplicité, nous avons pu voir que les PWM sont des outils très performants. Il demeure que la description des CArGbox gagne en précision par l'utilisation d'outils qui tiennent compte des dépendances entre les positions adjacentes. Devant les qualités de la TFFM face à ce problème, nous souhaitons qu'elle puisse s'affirmer comme un standard dans le domaine des gènes à boîte MADS. Cependant, devant les résultats obtenus par Smaczniak et al. (2017) grâce aux K-mer, une étude comparative de ces modèles serait nécessaire.

Conclusions

À ce stade, j'espère avoir convaincu le lecteur que comprendre la liaison des TF était une étape clé dans la compréhension des phénomènes de régulation. Dans cette thèse, nous nous sommes donc attachés à construire des modèles bioinformatiques prédictifs de la liaison pour des TF clés impliqués dans la floraison et la formation des organes floraux.

En premier lieu, nous avons étudié ARF2 et ARF5, un répresseur et un activateur , qui appartiennent à une famille de 20 TF impliqués dans des phénomènes développementaux orchestrés par l'auxine. Grâce au DAP-Seq, et à des PWM, nous avons vu que ces 2 TF qui présentaient à priori des préférences de liaison monomérique communes, semblent se lier en duos sur des paires de sites monomériques qui diffèrent dans leurs configurations. En raison des configurations observés nous suggérons également que ARF5 puisse former une chaîne oligomérique sur l'ADN. Pour tester si certaines configurations de liaison favorisaient une régulation, nous avons confronté ces données de liaison *in vitro* avec une liste de gènes régulés *in vivo*. La configuration IR13, pourtant encore jamais décrite, semblait être favorable à une régulation des gènes adjacents. D'une manière générale, les résultats semblent indiquer que les ARF suivent une syntaxe particulière, les permettant de se lier sur des sites différents ou identiques et d'en activer certains plus que d'autres, ce qui permet de coordonner avec finesse les processus développementaux dont le pilier est l'auxine.

Dans un second temps, nous avons étudié LFY. Ce TF a été exhaustivement étudié par les membres de mon d'accueil, qui avaient déjà consacré du temps à modéliser sa liaison. En utilisant de nouvelles données et de nouvelles méthodes de prédiction, nous avons réussi à approfondir des connaissances pourtant déjà bien établies. D'une part, nous avons sensiblement amélioré le modèle de liaison existant grâce à l'utilisation des TFFM, d'autre part, nous avons montré qu'en combinant DAP-Seq et DNase-Seq, nous pouvions également améliorer la prédiction des régions liées en ChIP-Seq. Enfin, l'étude des régions liées uniquement en ChIP-Seq pourrait nous avoir permis d'identifier des co-facteurs importants dans la régulation de gènes d'intérêt, comme *AP3*.

Enfin, nous nous sommes penchés sur les gènes qui confèrent aux organes floraux leur identité. Alors que les TF à boîte MADS partagent une préférence de liaison pour les mêmes CArGbox, nous avons montré au travers l'exemple du complexe SEP3/AG que le domaine de tétramérisation confère à ce complexe une spécificité de liaison qui dépend de l'espacement entre les CArGbox. Ainsi, nous suggérons que la spécificité des sites des TF régulant la formation des différentes couronnes de la fleur ne repose non pas sur la séquence des CArGbox mais sur l'espacement de celles-ci dans le génome.

D'une manière générale, nous sommes parvenus à mettre en évidence des particularités de liaison de certains facteurs de transcription impliqués dans la floraison en travaillant sur l'amélioration de nos modèles de liaison et en combinant des données génomiques de natures différentes.

Discussion

IV.1 À propos des modèles de liaison utilisés

Dans le travail présenté dans les 3 chapitres de résultats, j'ai analysé 3 types de facteurs de transcription appartenant à 3 familles différentes : LFY les ARF et les MADS-box TF. Au cours de ces analyses, nous avons pu voir que les malgré leur simplicité, les PWM sont des outils très performants. Ainsi, elles se prêtent très bien à la plupart des analyses que nous avons réalisées : elles affichent un très bon pouvoir prédictif – en particulier dans le cadre des DAP-Seq de LFY et des complexes SEP3-AG – et celui-ci est suffisant pour rechercher des espacements entre sites de liaisons, dans le cas où le facteur de transcription d'intérêt se lie de façon coopérative sur plusieurs sites. Par rapport au consensus, elle présente en plus l'avantage de capturer une plus grande variété de sites et de leur donner des scores relatifs à leurs affinités.

Cependant, la liaison de certains facteurs de transcription est mieux modélisée par des outils qui tiennent compte des interactions entre nucléotides proches. C'est le cas des facteurs de transcription à boîte MADS et de LFY, dont la TFFM décrit mieux le site de liaison. Si ses performances sont meilleures que celles de la PWM, elle ne les transcende pas. L'attrait de la TFFM vient plus du fait qu'elle donne un regard nouveau sur les particularités du site de liaison propre à un TF. On soulignera également sa grande simplicité d'utilisation, qui permet d'aller au delà de son apparente complexité, grâce à des outils très bien développés.

L'implémentation de la structure de l'ADN sur le site de liaison a été restreinte au site de LFY, pour lequel elle n'a pas amélioré les prédictions. Je pense cependant que l'utilisation du programme DNAshapedTFBS pourrait être étendue à l'étude des facteurs de transcription à boîte MADS pour lesquels l'importance de la structure de l'ADN sur le site de liaison a été mise en évidence. Ce programme présente néanmoins plusieurs inconvénients. D'abord, il est beaucoup plus lourd que les programmes qui implémentent les PWM et les TFFM. Le temps de calcul nécessaire est considérablement plus long. De plus, alors que nous avons souhaité comprendre les règles qui régissent les sites de liaison dimériques des ARF, son utilisation ne s'est pas montrée fructueuse. En effet, les différents gaps (ER7-8 par exemple) empêchent le programme de construire un modèle de structure fiable.

IV.2 À propos des méthodes utilisées pour tester les modèles de liaison

Dans chacune de nos analyses, nous avons évalué la capacité de notre modèle à discriminer un set de régions liées d'un set de régions témoins. Ici, plusieurs éléments peuvent être remis en question.

IV.2.1 Choisir un set de régions témoins

La construction d'un set de témoins est difficile. Je pense que le critère essentiel est le suivant : les régions liées doivent ressembler aux régions témoins. Cela dit, il est assez difficile de définir les critères de ressemblance. Il semblerait que le contenu nucléotidique ne suffise pas, des expériences menées dans le laboratoire ont montré que des modèles parviennent plus facilement à discriminer de l'ADN lié et de l'ADN "mélangé" (la composition nucléotidique est donc la même) que de l'ADN lié et de l'ADN génomique, de même composition nucléotidique. *BiasAway* (Hunt et al., 2014) permet en plus de conserver le contenu en dinucléotides de la région liée dans la région témoin, mais nous n'avons jamais testé nos modèles sur des séquences témoins répondant à ce critère. Afin de capturer au mieux la complexité de l'ADN génomique, nous avons fait le choix de piocher dans le génome des régions non liées ressemblant aux régions liées. Nos critères de ressemblance ont été :

— l'origine des régions (intron, exon, promoteur)

— le contenu de nucléotides GC

— La taille des régions

En atténuant le problème de la ressemblance, je pense que nous ne parvenons pas à l'éviter complètement. En effet, on ne peut pas être sûr que ces critères reflètent réellement les similitudes entre deux régions ; on peut par exemple évoquer les propriétés du génome au frontières des régions liées, qui ont certainement de l'importance pour attirer le TF jusqu'à un site de liaison spécifique, et que nous ne prenons pas en compte. Dans ce cadre, je pense que le DAP-Seq offre de nouvelles perspectives ; il peut permettre d'établir un inventaire des propriétés génomiques auxquelles sont sensibles les facteurs de transcription, mieux que toutes les autres techniques.

IV.2.2 Définir un critère pour évaluer nos modèles

Pour évaluer nos modèles, nous avons fait le choix du critère de l'aire sous la ROC. Ici, nous allons discuter de ce choix et d'une possible alternative, la *Precision recall curve* (PRC). La figure 47 illustre comment sont tracées les deux courbes.

Le principe de la ROC se conçoit assez naturellement. Pour un seuil donné, on trace le pourcentage de positifs sélectionnés parmi tous les positif (TPR) en fonction du pourcentage de négatifs sélectionnés parmi tous les négatifs (FPR). La signification de la PRC est plus difficile à appréhender. La précision peut être vue comme une mesure de la pertinence des cas sélectionnés (parmi les éléments sélectionnés, quel est le pourcentage de positif) alors que le rappel n'est autre que le pourcentage de positifs sélectionné parmi tous les positifs, qui correspond au TPR de la ROC. Ainsi, on mesure la qualité en fonction de l'exhaustivité. Dans chacun des cas, on peut obtenir une mesure globale des performances d'un modèle donné en calculant l'aire sous la courbe. Cette mesure globale doit néanmoins être considérée avec du recul, car elle efface certaines informations. Pour cette raison, la ROC et la PRC possèdent chacune des caractéristiques propres qui les rendent respectivement plus adaptées à des cas particuliers. Les figures suivantes, qui permettent d'illustrer ce propos, ont été obtenues à partir du code téléchargé à l'adresse `https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation`.

La figure 48 montre l'influence des tailles relatives des sets positifs et négatifs. Alors que entre (a) et (b), seule la taille du set positif change (pas sa distribution), l'aire sous la ROC reste, elle inchangée. À l'inverse, l'aire sous la PRC est drastiquement modifiée. Dans de telles conditions, la PRC peut être utilisée pour comparer les performances de deux modèles sur les mêmes échantillons, sans qu'elle puisse

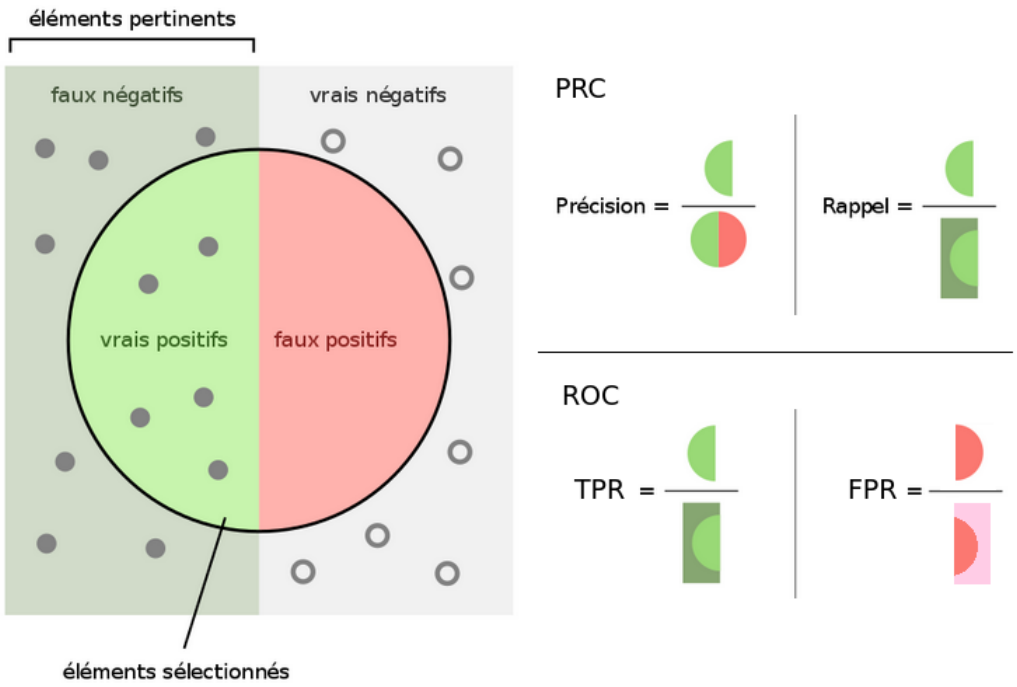


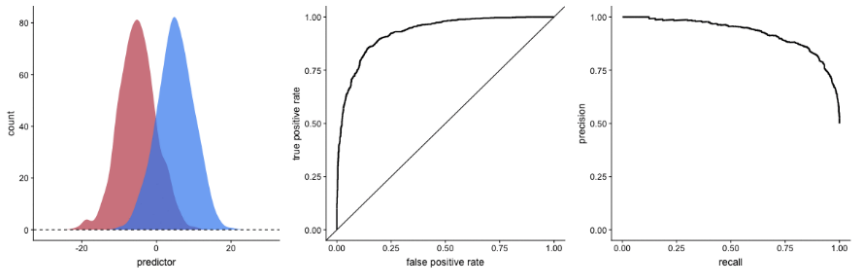
FIGURE 47: Un seuil sur le score permet de définir les éléments sélectionnés. Alors que la ROC exprime le *True Positive Rate* (TPR) en fonction du *False Positive Rate* (FPR), la courbe PRC trace la précision en fonction du TPR, aussi appelé rappel. Adapté d'après Wikipédia.

donner de mesure absolue sur un modèle. Ce cas ne s'applique pas à nos expériences puisque nous avons veillé à utiliser des sets positifs et négatifs de même taille. Cependant, si nos modèles venaient à être utilisés pour prédire les régions liées dans le génome entier, ceci devrait être pris en considération.

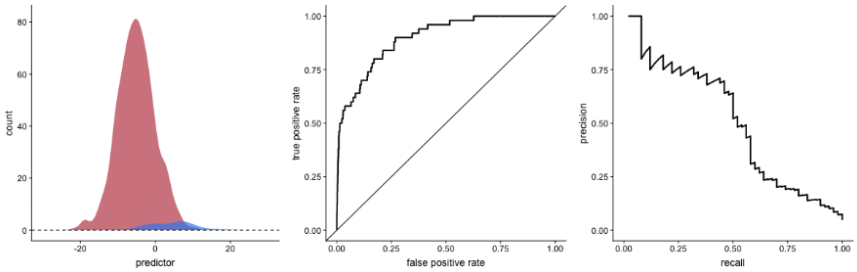
Un cas plus délicat est l'influence de la variance des distributions. Dans l'exemple montré en figure 49, en (b), une partie de la distribution des scores du set positif (en bleu) se détache complètement de la distribution du set négatif (en rouge) alors qu'en (a), les deux distributions sont complètement confondues. À l'inverse de l'aire sous la PRC, l'aire sous la ROC est supérieure en (a) qu'en (b). Ici, la nature du problème doit être clairement définie. Si on souhaite améliorer l'efficacité d'un moteur de recherche, les résultats affichés doivent être pertinents, l'exhaustivité de ces résultats n'est pas un critère important. Dans ces conditions, le cas (b) est clairement meilleur puisqu'une partie des deux distributions – qui correspond aux résultats retournés par le moteur de recherche – est complètement disjointe. Ainsi, dans le cas où on souhaite obtenir une sous-partie qui ne contient pas de faux positifs, l'aire sous la PRC est plus indiquée. On remarquera cependant que la pente de la courbe ROC à l'origine joue un rôle d'indicateur similaire que celui de l'aire sous la PRC. Si le problème est de nature différente et que l'on souhaite faire deux groupes – on détient des photographies de bouquetins et de chamois et on souhaite faire un tas de photographies pour chaque espèce – alors le cas (a) est meilleur puisqu'il occasionne moins de mélange entre les distributions. Dans ce cas, on préférera l'aire sous la courbe ROC.

Ainsi, pour faire la mesure de l'efficacité de nos modèles à classer les région liées et les régions non liées en deux groupes, je pense que la ROC est un outil adapté. Cependant, si nous devions appliquer nos modèles à prédire les sites de liaison d'un TF dans le génome, il faudrait peut-être préférer la PRC. En effet, les modèles que l'on utilise sont limités par le bruit et les conditions expérimentales des expériences

sur lesquels ils sont construits et sur lesquels on les teste. En DAP-Seq, par exemple, en raison du bruit, une partie des pics ne sont pas reproduits d'un réplicat à un autre. La quantité de protéine utilisée joue également certainement un rôle sur les résultats mais ce paramètre ne pourra jamais être pris en compte par nos modèles. Comme les modèles ne pourront jamais parfaitement refléter le comportement des TF, je pense qu'il est plus pertinent d'essayer de prédire avec fiabilité peu de régions mais d'être sûr qu'elles seront liées.

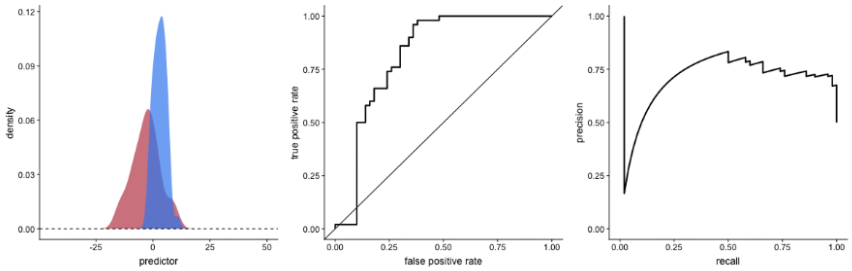


(a) Le set positif et le set négatif ont des tailles identiques

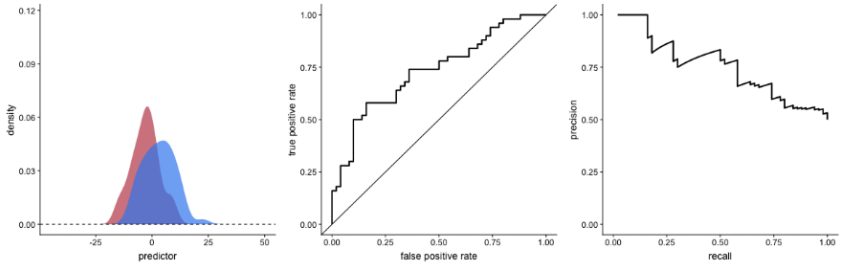


(b) Le set positif et le set négatif ont des tailles différentes

FIGURE 48: Influence des tailles relatives du set positif et du set négatif. À gauche, les distributions des scores du set positif et du set négatif sont respectivement représentées en bleu et en rouge. La ROC et la PRC sont respectivement tracées au milieu et à droite.



(a) Le set positif a une variance faible



(b) Le set positif a une variance importante

FIGURE 49: Influence de la variance des distributions. À gauche, les distributions des scores du set positif et du set négatif sont respectivement représentées en bleu et en rouge. La ROC et la PRC sont respectivement tracées au milieu et à droite.

IV.3 La liaison des TF définit un modèle de promoteur

Dans la discussion du chapitre 1, sur les ARF, nous avons pu voir que les gènes régulés par l'auxine sont liés en DAP-Seq entre 0 et 600 paires de bases du site d'initiation de la transcription (figure 30).

Cette observation est en accord avec l'étude publiée par Yu et al. (2016). Dans celle-ci, les auteurs utilisent 478 PWM de TF accessibles sur TRANSFAC, JASPAR, Athamap et CIS-BP (Bülow et al., 2006; Khan et al., 2017; Matys et al., 2006; Weirauch et al., 2014). Les sites de liaison de ces TF dans 3 génomes ont ainsi pu être inférés. En ne gardant que ceux conservés dans 3 espèces différentes (comme gage de leur fonctionnalité), il apparaît que la très grande majorité des sites de liaison conservés se situent 300 paires de bases avant le site d'initiation de la transcription. En parallèle, les auteurs de l'article ont étudié la position des pics de ChIP-Seq de 27 TF dans Arabidopsis thaliana, par rapport aux sites d'initiation de la transcription des gènes les plus proches. Bien que semblant de localiser également quelques centaines de paires de bases en amont du site d'initiation de la transcription, la position de ces pics est plus diffuse que celles des sites de liaison conservés, ou que dans notre méthode combinant DAP-Seq et RNA-Seq. Alors que les auteurs de l'article invoquent la grande taille des pics ChIP-Seq, qui complique la localisation du pic, il est également probable que beaucoup de régions liées en ChIP-Seq ne soient pas fonctionnelles, ce qui empêche de faire un lien direct avec leur position par rapport à un gène. En raison du nombre de ChIP-Seq utilisés (27), la robustesse de l'analyse y est également critiquée. En comparaison, de très nombreuses expériences de DAP-Seq ont été réalisées pour des TF d'*Arabidopsis thaliana*. Notre approche, qui combine DAP-Seq et données d'expression pourrait permettre de dépasser cette limite. À l'inverse, si en raison du grand nombre de PWM disponibles, l'étude des sites de liaison conservés permet de réaliser des analyses plus complètes, elle n'est pas – contrairement au DAP-Seq – un gage de la liaison du TF. Par exemple, tous les ARF semblent lier le même motif alors que les régions réellement liées diffèrent.

En utilisant la liste de gènes régulés par AG fournie par Ó'Maoiléidigh et al. (2013), nous avons souhaité voir si nous pouvions trouver des configurations enrichies en cas de régulation. Mais il n'est pas possible de conclure de manière affirmative (figure 50). En effet, il semble que les promoteurs des gènes non régulés soient enrichis de la même manière que ceux des gènes régulés par AG. On peut cependant nuancer ce résultat : il est possible que les gènes liés par SEP3/AG et non régulés par AG soient en fait régulés par d'autres facteurs de transcription à boîte MADS. La liaison de SEP3/AG sur ces gènes en DAP-Seq viendrait simplement du fait que les sites de liaison de ces gènes sont des CArGbox. Si ces sites sont fonctionnels, alors cette figure semble confirmer notre observation sur les ARF : les TF fonctionnels semblent se lier très proches et en amont des sites d'initiation de la transcription des gènes qu'ils régulent. Alors que dans un génome, un grand nombre de régions liées par un TF n'a pas d'influence sur la régulation, cette observation pourrait permettre d'éliminer un grand nombre de pics de ChIP-Seq sans importance.

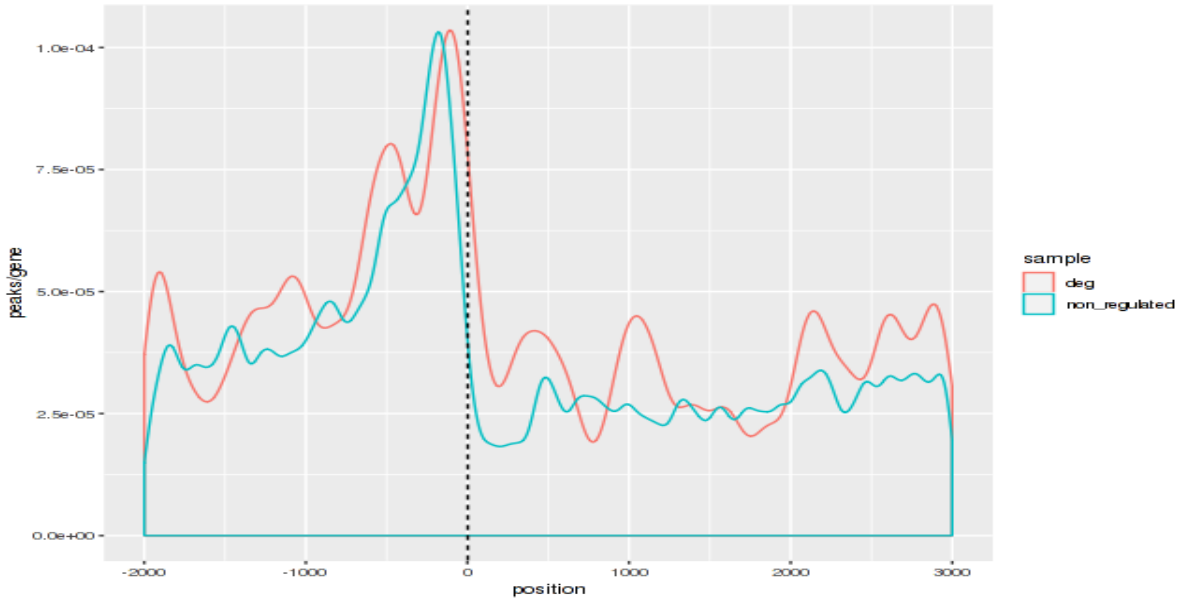


FIGURE 50: Position des pics de DAP-Seq du complexe SEP3AG par rapport aux gènes dans *A. thaliana*. L'axe y donne la densité de pics par gène et l'axe x donne la position des pics sur les gène. Les positions négatives correspondent aux positions avant le $5'UTR$, le sens des gènes ayant été pris en compte. Les gènes régulés par AG sont en rouge alors que les gènes non régulés sont en bleu

Cependant, le promoteur n'est pas nécessairement la seule région régulatrice d'un gène. Ainsi, la régulation du gène *AG* est opérée par la liaison de LFY dans son deuxième intron (Moyroud et al., 2011b) alors que l'activité du gène *TFL1* est sensible à la liaison de LFY plusieurs milliers de paires de basese en amont du $3'$ du gène. En raison de la taille variable des différents éléments qui composent un gène, ce type de région régulatrice n'est pas observable dans notre analyse. Elle mériterait donc d'être reprise en ne prenant plus comme point de référence le site d'initiation de la transcription (on pourrait prendre le début de chaque exon ou le $3'UTR'$).

IV.4 À propos de la bioinformatique

Durant ma thèse, il m'a semblé pouvoir faire la distinction entre deux manières différentes d'effectuer les analyses en bioinformatique.

Ainsi, dans mon équipe d'accueil à Grenoble, les questions sont directement rattachées à des phénomènes biologiques incompris, précis. Ici, le mot précis est important : dans chaque partie, nous avons répondu à une question propre à un phénomène, à un TF ou à une famille de TF. D'abord, nous avons cherché à éclaircir le comportement des ARF, puis de LFY et enfin des facteurs de transcriptions à boîte MADS.

Cependant, j'ai pu découvrir une autre approche. Dans le cadre de ma thèse, j'ai été amené à faire un stage de 3 mois dans l'équipe d'Anthony Mathelier à Oslo en Norvège, financé par l'IDEX *graduate school*. Son équipe, *Computational Biology & Gene Regulation*, est uniquement tournée vers la bioinformatique. Les recherches de celle-ci se concentrent sur la dérégulation des gènes dans les maladies cancéreuses. Ce stage représentait donc l'occasion pour que je me forme à la fois aux méthodes généralement employées dans le donaine de la biologie computationnelle et d'autre part à l'utilisation de modèles plus pointus que les PWM. Lors de ce séjour, les questions posées étaient générales, elles cherchent à établir des règles valides à l'échelle du génome. Ainsi nous avons cherché à voir si les modifications de structure de l'ADN induites par la méthylation des cytosines pouvaient affecter la liaison des TF sur l'ADN. Cela entraîne de grandes différences dans la manière d'aborder les problèmes.

Dans le deuxième cas, la problématique demande des ressources de calcul plus importantes. De plus, les questions font souvent appel à des modèles novateurs complexes. Pour étudier l'influence de la méthylation sur la structure, il faut auparavant un modèle qui calcule la structure, puis le modifier pour qu'il tienne compte de la méthylation. Pour voir à quel point cela affecte la liaison des TF, il faut ensuite un modèle qui puisse déduire la probabilité de liaison à partir de la structure calculée. Il me semble donc que la part de programmation y est extrêmement importante. Les programmes doivent êtres optimisés pour être suffisamment rapides et codés avec beaucoup de soin. Je dirais que la part d'informatique y est prépondérante comparée à la part de biologie.

Dans le premier cas, on cherchera plus à appliquer les modèles développés par la deuxième catégorie de bioinformaticiens. Ces modèles serviront cependant à éclairer des problèmes biologiques précis. Ainsi, l'utilisation des PWM et des TFFM nous aura permis d'établir les préférences de configuration coopérative pour les TF à boîte MADS et les ARF. Ainsi, les programmes développés par la première catégorie de bioinformaticiens servent à l'origine à répondre à des questions pour des TF précis. Il est cependant remarquable de constater que ces questions peuvent être étendues à une plus grande échelle. Si le programme de calcul d'enrichissement de distances a été initialement développé pour les ARF, nous avons pu l'appliquer au TF à boîte MADS. Le stage de M1 de Line Andresen a même eu pour objectif d'étendre cette question à tous les TF d'*A. thaliana* pour lesquels l'analyse était possible. Même si nous n'avons pas eu le temps d'approfondir les analyses, celles-ci ont permis de mettre en évidence des TF potentiels se liant en coopérativité.

Les deux approches sont donc complémentaires, chacune profitant à l'autre.

IV.5 Réflexion sur la portée des modèles et de la génomique

Nous clôturerons cette thèse par une réflexion sur l'utilité des modèles et de la génomique dans la compréhension des phénomènes biologiques. Ici, nous discuterons de l'impact et des limites de la génomique associée à la bioinformatique.

La génomique est apparue au début du XXI^{e} siècle, et son essor n'a pas subi de ralentissement. Les techniques de séquençage ont progressé, donnant accès à des informations de plus en plus pointues et coûtent de moins en moins cher. Alors que les génomes délivrent de plus en plus d'informations et que

le monde de la génomique s'enrichit des bases de données de plus en plus fournies, nous souhaitons ici prendre la mesure de ces avancées et de ce qu'elles impliquent en biologie, dans la compréhension des phénomènes de régulation des gènes.

D'une manière certaine, ces avancées donnent l'espoir de décrypter les génomes. En effet, il est aujourd'hui possible de connaître la position des TF sur l'ADN, savoir dans quels tissus ils sont exprimés, connaître l'état de l'ADN et de la chromatine dans ces tissus. Le but ultime est maintenant de comprendre comment s'orchestrent tous ces éléments. On peut poser simplement le problème de la manière suivante : être capable de prédire quels sont les gènes régulés par chaque TF. En effet, en capturant cette information, on peut deviner les cascades de régulation. Heureusement pour moi et pour tous les gens dont le métier est de répondre à des questions, poser des questions, même simples n'occasionne que rarement des réponses évidentes.

Le premier obstacle, qui est l'objet principal de cette thèse, est d'être capable de prédire la liaison d'un TF. Sur ce point, les modèles existants parviennent assez bien à séparer un set de régions liées d'un set de régions non liées. Cependant, leurs pouvoirs prédictifs à l'échelle d'un génome occasionneraient un trop grand nombre de faux positifs pour permettre des prédictions solides. Le deuxième obstacle vient de la nature des régions liées par un TF. En ChIP-Seq, toutes les régions liées ne sont pas fonctionnelles, elles ne régulent pas toutes des gènes. Ainsi, même un modèle de liaison solide ne saurait pas donner l'accès aux gènes régulés par un TF. Un troisième obstacle vient de la nature intrinsèque des données. Certaines techniques, comme le ChIP-Seq, peuvent être très difficiles à mettre en œuvre ; par exemple le nombre de cellules d'un tissu peut ne pas suffire. Ceci entraîne que la quantité de données ChIP-Seq disponible ne pourra jamais refléter le comportement de chaque TF dans chaque tissu.

Je pense cependant qu'il existe des moyens de surmonter la plupart de ces difficultés en croisant des données génomiques de natures différentes. On peut vérifier la fonctionnalité des sites de liaison capturés par un modèle en étudiant leur conservation dans d'autres espèces. De plus, dans le chapitre 1, nous avons vu qu'en croisant DAP-Seq et RNA-Seq, la fonctionnalité des gènes liés semble dépendre de la configuration de la liaison (IR13) et de la position de ces sites vis à vis des gènes qu'ils régulent, ce qui peut expliquer que des régions soient liées sans êtres fonctionnelles. Enfin, dans le chapitre sur LFY, nous avons montré qu'une partie des régions liées par ChIP-Seq pouvait être inférée à partir des données de DAP-Seq et des données de DNaseI dans le tissu correspondant, permettant ainsi de surmonter le troisième obstacle. Au lieu de réaliser $n \cdot m$ ChIP-Seq ou n et m correspondent respectivement au nombre de TF et au nombre de type cellulaire, on réalise n DAP-Seq, beaucoup plus simples dans leur mise en œuvre que le ChIP-Seq, et m DNase-Seq.

Ainsi, l'utilisation combinée et réfléchie des données disponibles pourra améliorer de manière notable la compréhension des génomes. Cependant, il restera à mon avis des zones d'ombre difficiles à éclaircir. Parmi celles-ci, on peut citer les modifications post-traductionnelles ou les co-facteurs qui donnent sa fonctionnalité à un TF. Ainsi, lorsqu'un co-facteur ne contacte pas l'ADN, les expériences de séquençage peineront à le mettre en évidence. Pour cette raison, la génomique et la bioinformatique doivent rester en contact avec la biologie et la biochimie.

Je pense donc que la compréhension des génomes et des réseaux de régulation nécessitera une part de modélisation, mais que cette modélisation ne saurait englober une partie des phénomènes, qui devront être traités au cas par cas dans un laboratoire de génétique.

Bibliographie

Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite : tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2) :W202–W208, 2009.

Ruth Bastow, Joshua S Mylne, Clare Lister, Zachary Lippman, Robert A Martienssen, and Caroline Dean. Vernalization requires epigenetic silencing of flc by histone methylation. *Nature*, 427(6970) :164, 2004.

Otto G Berg and Peter H von Hippel. Selection of dna binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4) : 723–743, 1987.

D Roeland Boer, Alejandra Freire-Rios, Willy AM van den Berg, Terrens Saaki, Iain W Manfield, Stefan Kepinski, Irene López-Vidrieo, Jose Manuel Franco-Zorrilla, Sacco C de Vries, Roberto Solano, et al. Structural basis for dna binding specificity by the auxin-dependent arf transcription factors. *Cell*, 156 (3) :577–589, 2014.

Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12) :1213, 2013.

Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. Atac-seq : a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1) :21–29, 2015.

Lorenz Bülow, Nils Ole Steffens, Claudia Galuschka, Martin Schindler, and Reinhard Hehl. Athamap : from in silico data to real transcription factor binding sites. *In silico biology*, 6(3) :243–252, 2006.

Thomas S Carroll, Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. Impact of artifact removal on chip quality metrics in chip-seq and chip-exo data. *Frontiers in genetics*, 5 :75, 2014.

Barry Causier, Mary Ashworth, Wenjia Guo, and Brendan Davies. The topless interactome : a framework for gene repression in arabidopsis. *Plant physiology*, 158(1) :423–438, 2012.

Hicham Chahtane. *Etude fonctionnelle et évolutive de LEAFY, un facteur de transcription clé dans la formation des fleurs*. PhD thesis, Grenoble, 2014.

Hicham Chahtane, Gregoire Denay, Julia Engelhorn, Marie Monniaux, Edwige Moyroud, Fanny Moreau, Cristel Carles, Gabrielle Tichtinsky, Chloe Zubieta, and François Parcy. Floral development : an integrated view. In *Les Houches Summer School*, volume 102, page np. Oxford Scholarship Online, 2014.

J. Cheneby, M. Gheorghe, M. Artufel, A. Mathelier, and B. Ballester. ReMap 2018 : an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, 46(D1) :D267–D275, Jan 2018.

Youfa Cheng and Yunde Zhao. A role for auxin in flower development. *Journal of integrative plant biology*, 49(1) :99–104, 2007.

Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. Dnashaper : an r/bioconductor package for dna shape prediction and feature encoding. *Bioinformatics*, 32(8) : 1211–1213, 2015.

Enrico S Coen and Elliot M Meyerowitz. The war of the whorls : genetic interactions controlling flower development. *nature*, 353(6339) :31, 1991.

Melanie Cole, John Chandler, Dolf Weijers, Bianca Jacobs, Petra Comelli, and Wolfgang Werr. Dornröschen is a direct target of the auxin response factor monopteros in the arabidopsis embryo. *Development*, 136(10) :1643–1651, 2009.

ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414) :57, 2012.

Filomena De Lucia, Pedro Crevillen, Alexandra ME Jones, Thomas Greb, and Caroline Dean. A phd-polycomb repressive complex 2 triggers the epigenetic silencing of flc during vernalization. *Proceedings of the National Academy of Sciences*, 105(44) :16831–16836, 2008.

Weiwei Deng, Hua Ying, Chris A Helliwell, Jennifer M Taylor, W James Peacock, and Elizabeth S Dennis. Flowering locus c (flc) regulates development pathways throughout the life cycle of arabidopsis. *Proceedings of the National Academy of Sciences*, 108(16) :6680–6685, 2011.

Marcos Egea-Cortines, Heinz Saedler, and Hans Sommer. Ternary complex formation between the mads-box proteins squamosa, deficiens and globosa is involved in the control of floral architecture in antirrhinum majus. *The EMBO journal*, 18(19) :5370–5379, 1999.

José M Franco-Zorrilla, Irene López-Vidriero, José L Carrasco, Marta Godoy, Pablo Vera, and Roberto Solano. Dna-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, 111(6) :2367–2372, 2014.

Nowlan H Freese, David C Norris, and Ann E Loraine. Integrated genome browser : visual analytics platform for genomics. *Bioinformatics*, 32(14) :2089–2095, 2016.

Mary Galli, Arjun Khakhar, Zefu Lu, Zongliang Chen, Sidharth Sen, Trupti Joshi, Jennifer L Nemhauser, Robert J Schmitz, and Andrea Gallavotti. The dna binding landscape of the maize auxin response factor family. *Nature communications*, 9(1) :4526, 2018.

Valérie Gaudin, Marc Libault, Sylvie Pouteau, Trine Juul, Gengchun Zhao, Delphine Lefebvre, and Olivier Grandjean. Mutations in like heterochromatin protein 1 affect flowering time and plant architecture in arabidopsis. *Development*, 128(23) :4847–4858, 2001.

Kevin Goslin, Beibei Zheng, Antonio Serrano-Mislata, Liina Rae, Patrick T Ryan, Kamila Kwaśniewska, Bennett Thomson, Diarmuid S Ó'Maoiléidigh, Francisco Madueño, Frank Wellmer, et al. Transcription factor interplay between leafy and apetala1/cauliflower during floral initiation. *Plant physiology*, 174 (2) :1097–1109, 2017.

Tom J Guilfoyle. The pb1 domain in auxin response factor and aux/iaa proteins : a versatile protein interaction module in the auxin response. *The Plant Cell*, 27(1) :33–43, 2015.

Y. Guo, S. Mahony, and D. K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, 8(8) :e1002638, 2012.

Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome biology*, 8(2) :R24, 2007.

Cécile Hamès, Denis Ptchelkine, Clemens Grimm, Emmanuel Thevenon, Edwige Moyroud, Francine Gérard, Jean-Louis Martiel, Reyes Benlloch, François Parcy, and Christoph W Müller. Structural basis for leafy floral switch function and similarity with helix-turn-helix proteins. *The EMBO journal*, 27(19) : 2628–2637, 2008.

Timothy E Hayes, Piali Sengupta, and Brent H Cochran. The human c-fos serum response factor and the yeast factors grm/prtf have related dna-binding specificities. *Genes & development*, 2(12b) :1713–1722, 1988.

Guangming He, Axel A Elling, and Xing Wang Deng. The epigenome and plant development. *Annual review of plant biology*, 62 :411–435, 2011.

Takashi Honma and Koji Goto. Complexes of mads-box proteins are sufficient to convert leaves into floral organs. *Nature*, 409(6819) :525, 2001.

Hai Huang, Yukiko Mizukami, Yi Hu, and Hong Ma. Isolation and characterization of the binding sequences for the product of the arabidopsis floral homeotic gene agamous. *Nucleic Acids Research*, 21(20) :4769–4776, 1993.

Kai Huang, John M Louis, Logan Donaldson, Fei-Ling Lim, Andrew D Sharrocks, and G Marius Clore. Solution structure of the mef2a–dna complex : structural basis for the modulation of dna bending and specificity by mads-box transcription factors. *The EMBO Journal*, 19(11) :2615–2628, 2000.

Véronique Hugouvieux, Catarina S Silva, Agnes Jourdain, Arnaud Stigliani, Quentin Charras, Vanessa Conn, Simon J Conn, Cristel C Carles, François Parcy, and Chloe Zubieta. Tetramerization of mads family transcription factors sepallata3 and agamous is required for floral meristem determinacy in arabidopsis. *Nucleic acids research*, 46(10) :4966–4977, 2018.

Rebecca Worsley Hunt, Anthony Mathelier, Luis Del Peso, and Wyeth W Wasserman. Improving analysis of transcription factor binding sites within chip-seq data based on topological motif enrichment. *BMC genomics*, 15(1) :472, 2014.

James P Jackson, Anders M Lindroth, Xiaofeng Cao, and Steven E Jacobsen. Control of cpnpg dna methylation by the kryptonite histone h3 methyltransferase. *Nature*, 416(6880) :556, 2002.

James P Jackson, Lianna Johnson, Zuzana Jasencakova, Xing Zhang, Laura PerezBurgos, Prim B Singh, Xiaodong Cheng, Ingo Schubert, Thomas Jenuwein, and Steven E Jacobsen. Dimethylation of histone h3 lysine 9 is a critical mark for dna methylation and gene silencing in arabidopsis thaliana. *Chromosoma*, 112(6) :308–315, 2004.

Seonghoe Jang, Virginie Marchal, Kishore CS Panigrahi, Stephan Wenkel, Wim Soppe, Xing-Wang Deng, Federico Valverde, and George Coupland. Arabidopsis cop1 shapes the temporal pattern of co accumulation conferring a photoperiodic flowering response. *The EMBO journal*, 27(8) :1277–1288, 2008.

Kerstin Kaufmann, Rainer Melzer, and Günter Theißen. Mikc-type mads-domain proteins : structural modularity, protein interactions and network evolution in land plants. *Gene*, 347(2) :183–198, 2005.

Kerstin Kaufmann, Jose M Muino, Ruy Jauregui, Chiara A Airoldi, Cezary Smaczniak, Pawel Krajewski, and Gerco C Angenent. Target genes of the mads transcription factor sepallata3 : integration of developmental and hormonal pathways in the arabidopsis flower. *PLoS biology*, 7(4) :e1000090, 2009.

Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Cheneby, Shubhada R Kulkarni, Ge Tan, et al. Jaspar 2018 : update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1) :D260–D266, 2017.

Benjamin L Kidder, Gangqing Hu, and Keji Zhao. Chip-seq : technical considerations for obtaining high-quality data. *Nature immunology*, 12(10) :918, 2011.

Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, page 1, 2019.

Xuelei Lai, Leonie Verhage, Veronique Hugouvieux, and Chloe Zubieta. Pioneer factors in animals and plants—colonizing chromatin for gene regulation. *Molecules*, 23(8) :1914, 2018.

Rebecca S Lamb, Theresa A Hill, Queenie K-G Tan, and Vivian F Irish. Regulation of apetala3 floral homeotic gene expression by meristem identity genes. *Development*, 129(9) :2079–2086, 2002.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4) :357–359, Mar 2012.

Sascha Laubinger, Virginie Marchal, Jose Gentilhomme, Stephan Wenkel, Jessika Adrian, Seonghoe Jang, Carmen Kulajta, Helen Braun, George Coupland, and Ute Hoecker. Arabidopsis spa proteins regulate photoperiodic flowering and interact with the floral inducer constans to regulate its stability. *Development*, 133(16) :3213–3222, 2006.

Ilha Lee, Diana S Wolfe, Ove Nilsson, and Detlef Weigel. A leafy co-regulator encoded by unusual floral organs. *Current Biology*, 7(2) :95–104, 1997.

Jeong Hwan Lee, Seong Jeon Yoo, Soo Hyun Park, Ildoo Hwang, Jong Seob Lee, and Ji Hoon Ahn. Role of svp in the control of flowering time by ambient temperature in arabidopsis. *Genes & development*, 21(4) :397–402, 2007.

Chenlong Li, Lianfeng Gu, Lei Gao, Chen Chen, Chuang-Qi Wei, Qi Qiu, Chih-Wei Chien, Suikang Wang, Lihua Jiang, Lian-Feng Ai, et al. Concerted genomic targeting of h3k27 demethylase ref6 and chromatin-remodeling atpase brm in arabidopsis. *Nature genetics*, 48(6) :687, 2016a.

Dan Li, Chang Liu, Lisha Shen, Yang Wu, Hongyan Chen, Masumi Robertson, Chris A Helliwell, Toshiro Ito, Elliot Meyerowitz, and Hao Yu. A repressor complex governs the integration of flowering signals in arabidopsis. *Developmental cell*, 15(1) :110–120, 2008.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14) :1754–1760, Jul 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16) :2078–2079, Aug 2009.

Jinsen Li, Jared M Sagendorf, Tsu-Pei Chiu, Marco Pasi, Alberto Perez, and Remo Rohs. Expanding the repertoire of dna shape features for genome-scale studies of transcription factor binding. *Nucleic acids research*, 45(22) :12877–12887, 2017.

Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, 5(3) :1752–1779, 09 2011. doi : 10.1214/11-AOAS466. URL `https://doi.org/10.1214/11-AOAS466`.

Si-Bei Li, Zong-Zhou Xie, Chun-Gen Hu, and Jin-Zhi Zhang. A review of auxin response factors (arfs) in plants. *Frontiers in plant science*, 7 :47, 2016b.

Xiaoying Lin, Samir Kaul, Steve Rounsley, Terrance P Shea, Maria-Ines Benito, Christopher D Town, Claire Y Fujii, Tanya Mason, Cheryl L Bowman, Mary Barnstead, et al. Sequence and analysis of chromosome 2 of the plant arabidopsis thaliana. *Nature*, 402(6763) :761, 1999.

Chunyan Liu, Falong Lu, Xia Cui, and Xiaofeng Cao. Histone methylation in higher plants. *Annual review of plant biology*, 61 :395–420, 2010.

Liansen Liu, Michael J White, and Thomas H MacRae. Transcription factors and their genes in higher plants : functional domains, evolution and regulation. *European Journal of Biochemistry*, 262(2) :247–257, 1999.

Xiao Liu, David M Noll, Jason D Lieb, and Neil D Clarke. Dip-chip : rapid and accurate determination of dna-binding specificity. *Genome research*, 15(3) :421–427, 2005.

Zhan-Bin Liu, Tim Ulmasov, Xiangyang Shi, Gretchen Hagen, and Tom J Guilfoyle. Soybean gh3 promoter contains multiple auxin-inducible elements. *The Plant Cell*, 6(5) :645–657, 1994.

Jan U Lohmann, Ray L Hong, Martin Hobe, Maximilian A Busch, François Parcy, Rüdiger Simon, and Detlef Weigel. A molecular link between stem cell regulation and floral patterning in arabidopsis. *Cell*, 105(6) :793–803, 2001.

Alexis Maizel, Maximilian A Busch, Takako Tanahashi, Josip Perkovic, Masahiro Kato, Mitsuyasu Hasebe, and Detlef Weigel. The floral regulator leafy evolves by substitutions in the dna binding domain. *Science*, 308(5719) :260–263, 2005.

Erik G Marklund, Anel Mahmutovic, Otto G Berg, Petter Hammar, David van der Spoel, David Fange, and Johan Elf. Transcription-factor binding and sliding on dna studied using micro-and macroscopic models. *Proceedings of the National Academy of Sciences*, 110(49) :19796–19801, 2013.

Raquel Martin-Arevalillo. *Aspects fonctionnels, structuraux et évolutifs de la réponse transcriptionnelle à l'auxine*. PhD thesis, Grenoble Alpes, 2017.

Raquel Martin-Arevalillo, Max H Nanao, Antoine Larrieu, Thomas Vinos-Poyo, David Mast, Carlos Galvan-Ampudia, Géraldine Brunoud, Teva Vernoux, Renaud Dumas, and François Parcy. Structure of the arabidopsis topless corepressor provides insight into the evolution of transcriptional repression. *Proceedings of the National Academy of Sciences*, 114(30) :8107–8112, 2017.

Anthony Mathelier and Wyeth W Wasserman. The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9) :e1003214, 2013.

Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W Wasserman. Dna shape features improve transcription factor binding site predictions in vivo. *Cell systems*, 3(3) :278–286, 2016.

Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. Transfac® and its module transcompel® : transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl_1) :D108–D110, 2006.

Rainer Melzer and Günter Theißen. Reconstitution of 'floral quartets' in vitro involving class b and class e floral homeotic proteins. *Nucleic acids research*, 37(8) :2723–2736, 2009.

Rainer Melzer, Wim Verelst, and Günter Theißen. The class e floral homeotic protein sepallata3 is sufficient to loop dna in 'floral quartet'-like complexes in vitro. *Nucleic Acids Research*, 37(1) :144–157, 2008.

Eugenio Gómez Minguet, Stéphane Segard, Céline Charavay, and François Parcy. Morpheus, a webtool for transcription factor binding analysis using position weight matrices with dependency. *PLoS One*, 10 (8) :e0135586, 2015.

Nobutaka Mitsuda and Masaru Ohme-Takagi. Functional analysis of transcription factors in arabidopsis. *Plant and Cell Physiology*, 50(7) :1232–1248, 2009.

Edwige Moyroud, Elske Kusters, Marie Monniaux, Ronald Koes, and François Parcy. Leafy blossoms. *Trends in plant science*, 15(6) :346–352, 2010.

Edwige Moyroud, Eugenio Gómez Minguet, Felix Ott, Levi Yant, David Posé, Marie Monniaux, Sandrine Blanchet, Olivier Bastien, Emmanuel Thévenon, Detlef Weigel, et al. Prediction of regulatory interactions from genome sequences using a biophysical model for the arabidopsis leafy transcription factor. *The Plant Cell*, 23(4) :1293–1306, 2011a.

Edwige Moyroud, Eugenio Gómez Minguet, Felix Ott, Levi Yant, David Posé, Marie Monniaux, Sandrine Blanchet, Olivier Bastien, Emmanuel Thévenon, Detlef Weigel, et al. Prediction of regulatory interactions from genome sequences using a biophysical model for the arabidopsis leafy transcription factor. *The Plant Cell*, 23(4) :1293–1306, 2011b.

Jose M Muino, Cezary Smaczniak, Gerco C Angenent, Kerstin Kaufmann, and Aalt DJ van Dijk. Structural determinants of dna recognition by plant mads-domain transcription factors. *Nucleic acids research*, 42 (4) :2138–2146, 2013.

Joshua S Mylne, Lynne Barrett, Federico Tessadori, Stéphane Mesnage, Lianna Johnson, Yana V Bernatavichute, Steven E Jacobsen, Paul Fransz, and Caroline Dean. Lhp1, the arabidopsis homologue of heterochromatin protein1, is required for epigenetic silencing of flc. *Proceedings of the National Academy of Sciences*, 103(13) :5012–5017, 2006.

N. T. T. Nguyen, B. Contreras-Moreira, J. A. Castro-Mondragon, W. Santana-Garcia, R. Ossio, C. D. Robles-Espinoza, M. Bahin, S. Collombet, P. Vincens, D. Thieffry, J. van Helden, A. Medina-Rivera, and M. Thomas-Chollier. RSAT 2018 : regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.*, 46(W1) :W209–W214, Jul 2018.

Ove Nilsson, Ilha Lee, Miguel A Blázquez, and Detlef Weigel. Flowering-time genes modulate the response to leafy activity. *Genetics*, 150(1) :403–410, 1998.

Kiyotaka Okada, Junichi Ueda, Masako K Komaki, Callum J Bell, and Yoshiro Shimura. Requirement of the auxin polar transport system in early stages of arabidopsis floral bud formation. *The Plant Cell*, 3 (7) :677–684, 1991.

Ronan C O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G Lewsey, Anna Bartlett, Joseph R Nery, Mary Galli, Andrea Gallavotti, and Joseph R Ecker. Cistrome and epicistrome features shape the regulatory dna landscape. *Cell*, 165(5) :1280–1292, 2016.

Diarmuid S Ó'Maoiléidigh, Samuel E Wuest, Liina Rae, Andrea Raganelli, Patrick T Ryan, Kamila Kwaśniewska, Pradeep Das, Amanda J Lohan, Brendan Loftus, Emmanuelle Graciet, et al. Control of reproductive floral organ identity specification in arabidopsis by the c function regulator agamous. *The Plant Cell*, 25(7) :2482–2503, 2013.

Alice Pajoro, Pedro Madrigal, Jose M Muiño, José Tomás Matus, Jian Jin, Martin A Mecchia, Juan M Debernardi, Javier F Palatnik, Salma Balazadeh, Muhammad Arif, et al. Dynamics of chromatin accessibility and gene regulation by mads-domain transcription factors in flower development. *Genome biology*, 15(3) :R41, 2014.

Sebastien Paque and Dolf Weijers. Q&a : Auxin : the plant molecule that influences almost anything. *BMC biology*, 14(1) :67, 2016.

Francois Parcy. Flowering : a time for integration. *International Journal of Developmental Biology*, 49 (5-6) :585–593, 2004.

Francois Parcy, Ove Nilsson, Maximilian A Busch, Ilha Lee, and Detlef Weigel. A genetic framework for floral patterning. *Nature*, 395(6702) :561, 1998.

Steven Passmore, Randolph Elble, and Bik-Kwoon Tye. A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. *Genes & Development*, 3(7) : 921–935, 1989.

Wendy Ann Peer. From perception to attenuation : auxin signalling and responses. *Current opinion in plant biology*, 16(5) :561–568, 2013.

Soraya Pelaz, Gary S Ditta, Elvira Baumann, Ellen Wisman, and Martin F Yanofsky. B and c floral organ identity functions require sepallata mads-box genes. *Nature*, 405(6783) :200, 2000.

Craig S Pikaard and Ortrun Mittelsten Scheid. Epigenetic regulation in plants. *Cold Spring Harbor perspectives in biology*, 6(12) :a019315, 2014.

Roy Pollock and Richard Treisman. A sensitive method for the determination of protein-dna binding specificities. *Nucleic Acids Research*, 18(21) :6197–6204, 1990.

Aaron R. Quinlan and Ira M. Hall. BEDTools : a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6) :841–842, 01 2010.

Mahe Raccaud, Elias T Friman, Andrea B Alber, Harsha Agarwal, Cedric Deluz, Timo Kuhn, J Christof M Gebhardt, and David M Suter. Mitotic chromosome binding predicts transcription factor properties in interphase. *Nature communications*, 10(1) :487, 2019.

Didier Reinhardt, Therese Mandel, and Cris Kuhlemeier. Auxin regulates the initiation and radial position of plant lateral organs. *The Plant Cell*, 12(4) :507–518, 2000.

José Luis Riechmann, Beth Allyn Krizek, and Elliot M Meyerowitz. Dimerization specificity of arabidopsis mads domain homeotic proteins apetala1, apetala3, pistillata, and agamous. *Proceedings of the National Academy of Sciences*, 93(10) :4793–4798, 1996.

Michel Ruiz Rosquete, Elke Barbez, and Jürgen Kleine-Vehn. Cellular auxin homeostasis : gatekeeping is housekeeping. *Molecular plant*, 5(4) :772–786, 2012.

François Roudier, Ikhlak Ahmed, Caroline Bérard, Alexis Sarazin, Tristan Mary-Huard, Sandra Cortijo, Daniel Bouyer, Erwann Caillieux, Evelyne Duvernois-Berthet, Liza Al-Shikhley, et al. Integrative epigenomic mapping defines four main chromatin states in arabidopsis. *The EMBO journal*, 30(10) : 1928–1938, 2011.

Camille Sayou. *Structure, fonction et évolution de LEAFY, facteur de transcription clé du développement floral*. PhD thesis, Université de Grenoble, 2013.

Camille Sayou, Marie Monniaux, Max H Nanao, Edwige Moyroud, Samuel F Brockington, Emmanuel Thévenon, Hicham Chahtane, Norman Warthmann, Michael Melkonian, Yong Zhang, et al. A promiscuous intermediate underlies the evolution of leafy dna binding specificity. *Science*, 343(6171) :645–648, 2014.

Camille Sayou, Max H Nanao, Marc Jamin, David Posé, Emmanuel Thévenon, Laura Grégoire, Gabrielle Tichtinsky, Grégoire Denay, Felix Ott, Marta Peirats Llobet, et al. A sam oligomerization domain shapes the genomic binding landscape of the leafy transcription factor. *Nature communications*, 7, 2016.

Joana Sequeira-Mendes, Irene Aragüez, Ramón Peiró, Raul Mendez-Giraldez, Xiaoyu Zhang, Steven E Jacobsen, Ugo Bastolla, and Crisanto Gutierrez. The functional topography of the arabidopsis genome is organized in a reduced number of linear motifs of chromatin states. *The Plant Cell*, 26(6) :2351–2366, 2014.

Susan Shannon and D Ry Meeks-Wagner. Genetic interactions that regulate inflorescence development in arabidopsis. *The Plant Cell*, 5(6) :639–655, 1993.

Rahul Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites : generalizing the position weight matrix. *PloS one*, 5(3) :e9722, 2010.

Cezary Smaczniak, Jose M Muiño, Dijun Chen, Gerco C Angenent, and Kerstin Kaufmann. Differences in dna binding specificity of floral homeotic protein complexes predict organ-specific target genes. *The Plant Cell*, 29(8) :1822–1835, 2017.

Young Hun Song, Shogo Ito, and Takato Imaizumi. Flowering time regulation : photoperiod-and temperature-sensing in leaves. *Trends in plant science*, 18(10) :575–583, 2013.

François Spitz and Eileen EM Furlong. Transcription factors : from enhancer binding to developmental control. *Nature reviews genetics*, 13(9) :613, 2012.

Gary D Stormo. Modeling the specificity of protein-dna interactions. *Quantitative biology*, 1(2) :115–130, 2013.

Gary D Stormo, Thomas D Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, 10(9) :2997–3011, 1982.

Takako Tanahashi, Naomi Sumikawa, Masahiro Kato, and Mitsuyasu Hasebe. Diversification of gene function : homologs of the floral regulator flo/lfy control the first zygotic cell division in the moss physcomitrella patens. *Development*, 132(7) :1727–1736, 2005.

Günter Theißen and Annette Becker. Gymnosperm orthologues of class b floral homeotic genes and their impact on understanding flower origin. *Critical Reviews in Plant Sciences*, 23(2) :129–148, 2004.

Günter Theissen and Heinz Saedler. Plant biology : floral quartets. *Nature*, 409(6819) :469, 2001.

Günter Theißen, Jan T Kim, and Heinz Saedler. Classification and phylogeny of the mads-box multigene family suggest defined roles of mads-box gene subfamilies in the morphological evolution of eukaryotes. *Journal of Molecular Evolution*, 43(5) :484–516, 1996.

Shiv B Tiwari, Gretchen Hagen, and Tom Guilfoyle. The roles of auxin response factor domains in auxin-responsive transcription. *The Plant Cell*, 15(2) :533–543, 2003.

Maria Tsompana and Michael J Buck. Chromatin accessibility : a window into the genome. *Epigenetics & chromatin*, 7(1) :33, 2014.

Franziska Turck, François Roudier, Sara Farrona, Marie-Laure Martin-Magniette, Elodie Guillaume, Nicolas Buisine, Séverine Gagnot, Robert A Martienssen, George Coupland, and Vincent Colot. Arabidopsis tfl2/lhp1 specifically associates with genes marked by trimethylation of histone h3 lysine 27. *PLoS genetics*, 3(6) :e86, 2007.

Tim Ulmasov, Gretchen Hagen, and Tom J Guilfoyle. Arf1, a transcription factor that binds to auxin response elements. *Science*, 276(5320) :1865–1868, 1997a.

Tim Ulmasov, Jane Murfett, Gretchen Hagen, and Tom J Guilfoyle. Aux/iaa proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *The Plant Cell*, 9(11) :1963–1971, 1997b.

Tim Ulmasov, Gretchen Hagen, and Tom J Guilfoyle. Dimerization and dna binding of auxin response factors. *The Plant Journal*, 19(3) :309–319, 1999.

Federico Valverde, Aidyn Mouradov, Wim Soppe, Dean Ravenscroft, Alon Samach, and George Coupland. Photoreceptor regulation of constans protein in photoperiodic flowering. *Science*, 303(5660) :1003–1006, 2004.

Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276, 2004.

Detlef Weigel and Ove Nilsson. A developmental switch sufficient for flower initiation in diverse plants. *Nature*, 377(6549) :495, 1995.

Dolf Weijers and Doris Wagner. Transcriptional responses to the auxin hormone. *Annual review of plant biology*, 67 :539–574, 2016.

Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6) :1431–1443, 2014.

Philip A Wigge, Min Chul Kim, Katja E Jaeger, Wolfgang Busch, Markus Schmid, Jan U Lohmann, and Detlef Weigel. Integration of spatial and temporal information during floral induction in arabidopsis. *Science*, 309(5737) :1056–1059, 2005.

Cara M Winter, Ryan S Austin, Servane Blanvillain-Baufume, Maxwell A Reback, Marie Monniaux, Miin-Feng Wu, Yi Sang, Ayako Yamaguchi, Nobutoshi Yamaguchi, Jane E Parker, et al. Leafy target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. *Developmental cell*, 20 (4) :430–443, 2011.

Christopher T Workman, Yutong Yin, David L Corcoran, Trey Ideker, Gary D Stormo, and Panayiotis V Benos. enologos : a versatile web tool for energy normalized sequence logos. *Nucleic acids research*, 33 (suppl_2) :W389–W392, 2005.

Miin-Feng Wu, Nobutoshi Yamaguchi, Jun Xiao, Bastiaan Bargmann, Mark Estelle, Yi Sang, and Doris Wagner. Auxin-regulated chromatin switch directs acquisition of flower primordium founder fate. *Elife*, 4 :e09269, 2015.

Judy Wynne and Richard Treisman. Srf and mcm1 have related but distinct dna binding specificities. *Nucleic acids research*, 20(13) :3297–3303, 1992.

Jun Xiao, Run Jin, and Doris Wagner. Developmental transitions : integrating environmental cues with hormonal signaling in the chromatin landscape in plants. *Genome biology*, 18(1) :88, 2017.

Nobutoshi Yamaguchi, Miin-Feng Wu, Cara M Winter, Markus C Berns, Staci Nole-Wilson, Ayako Yamaguchi, George Coupland, Beth A Krizek, and Doris Wagner. A molecular framework for auxin-mediated initiation of flower primordia. *Developmental cell*, 24(3) :271–282, 2013.

Kazuhiko Yamasaki, Takanori Kigawa, Makoto Inoue, Satoru Watanabe, Masaru Tateno, Motoaki Seki, Kazuo Shinozaki, and Shigeyuki Yokoyama. Structures and evolutionary origins of plant-specific transcription factor dna-binding domains. *Plant Physiology and Biochemistry*, 46(3) :394–401, 2008.

Levi Yant, Johannes Mathieu, Thanh Theresa Dinh, Felix Ott, Christa Lanz, Heike Wollmann, Xuemei Chen, and Markus Schmid. Orchestration of the floral transition and floral development in arabidopsis by the bifunctional transcription factor apetala2. *The Plant Cell*, 22(7) :2156–2170, 2010.

Chun-Ping Yu, Jinn-Jy Lin, and Wen-Hsiung Li. Positional distribution of transcription factor binding sites in arabidopsis thaliana. *Scientific reports*, 6 :25164, 2016.

Eva Zažímalová, Angus S Murphy, Haibing Yang, Klára Hoyerová, and Petr Hošek. Auxin transporters—why so many ? *Cold Spring Harbor perspectives in biology*, 2(3) :a001552, 2010.

Elena V Zemlyanskaya, Daniil S Wiebe, Nadezhda A Omelyanchuk, Victor G Levitsky, and Victoria V Mironova. Meta-analysis of transcriptome data identified tgtcnn motif variants associated with the response to plant hormone auxin in arabidopsis thaliana l. *Journal of bioinformatics and computational biology*, 14(02) :1641009, 2016.

Xiaoyu Zhang, Sophie Germann, Bartlomiej J Blus, Sepideh Khorasanizadeh, Valerie Gaudin, and Steven E Jacobsen. The arabidopsis lhp1 protein colocalizes with histone h3 lys27 trimethylation. *Nature structural & molecular biology*, 14(9) :869, 2007.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9) : R137, 2008.

Heng Zhu, Guohua Wang, and Jiang Qian. Transcription factors as readers and effectors of dna methylation. *Nature Reviews Genetics*, 17(9) :551, 2016.

Zecheng Zuo, Hongtao Liu, Bin Liu, Xuanming Liu, and Chentao Lin. Blue light-dependent interaction of cry2 with spa1 regulates cop1 activity and floral initiation in arabidopsis. *Current Biology*, 21(10) : 841–847, 2011.

Zheng Zuo, Basab Roy, Yiming Kenny Chang, David Granas, and Gary D Stormo. Measuring quantitative effects of methylation on transcription factor–dna binding affinity. *Science Advances*, 3(11) :eaao1799, 2017.

Annexes

V.1 Revue

Molecular Plant
Review Article

CellPress
PARTNER JOURNAL

Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants

Xuelei Lai[1,*], Arnaud Stigliani[1], Gilles Vachon[1], Cristel Carles[1], Cezary Smaczniak[2], Chloe Zubieta[1], Kerstin Kaufmann[2] and François Parcy[1,*]

[1]CNRS, Univ. Grenoble Alpes, CEA, INRA, BIG-LPCV, 38000 Grenoble, France

[2]Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, Berlin, Germany

***Correspondence: Xuelei Lai (xuelei.lai@cea.fr), François Parcy (francois.parcy@cea.fr)**

https://doi.org/10.1016/j.molp.2018.10.010

ABSTRACT

Transcription factors (TFs) are key cellular components that control gene expression. They recognize specific DNA sequences, the TF binding sites (TFBSs), and thus are targeted to specific regions of the genome where they can recruit transcriptional co-factors and/or chromatin regulators to fine-tune spatiotemporal gene regulation. Therefore, the identification of TFBSs in genomic sequences and their subsequent quantitative modeling is of crucial importance for understanding and predicting gene expression. Here, we review how TFBSs can be determined experimentally, how the TFBS models can be constructed *in silico*, and how they can be optimized by taking into account features such as position interdependence within TFBSs, DNA shape, and/or by introducing state-of-the-art computational algorithms such as deep learning methods. In addition, we discuss the integration of context variables into the TFBS modeling, including nucleosome positioning, chromatin states, methylation patterns, 3D genome architectures, and TF cooperative binding, in order to better predict TF binding under cellular contexts. Finally, we explore the possibilities of combining the optimized TFBS model with technological advances, such as targeted TFBS perturbation by CRISPR, to better understand gene regulation, evolution, and plant diversity.

Key words: Transcription factor binding site, Gene regulation, flower development

Lai X., Stigliani A., Vachon G., Carles C., Smaczniak C., Zubieta C., Kaufmann K., and Parcy F. (2019). Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. Mol. Plant. **12**, 743–763.

INTRODUCTION

Transcription factors (TFs) are sequence-specific DNA binding proteins that regulate gene expression in all organisms (Lelli et al., 2012; Lambert et al., 2018). They constitute a large number of protein-coding genes (between 4% and 10%) in the genomes of all species (Babu et al., 2004). For example, in the model plant *Arabidopsis thaliana*, 2492 genes encode TFs, accounting for more than 9% of its total protein-coding genes (Swarbreck et al., 2008; Pruneda-Paz et al., 2014). TFs orchestrate gene regulation by binding to their cognate DNA binding sites (TFBS) that are usually located in *cis*-regulatory regions. Upon binding to a TFBS, some TFs are able to recruit epigenetic factors, such as chromatin remodelers (e.g., BRAHMA and SPLAYED in plants [Bezhani et al., 2007]) or modifiers (e.g., polycomb repressive complexes [PRC] [Xiao and Wagner, 2015]) to alter chromatin states. TFs can also interact with components of transcriptional machineries, such as co-factor (e.g., Mediator and SAGA complexes in animals [Allen and Taatjes, 2015; Baptista et al., 2017]), general transcriptional factors (Müller et al., 2010), and RNA polymerase II for regulation of transcriptional initiation. The interplay between TFs and these factors together leads to robust and dynamic gene expression regulation (Spitz and Furlong, 2012; Voss and Hager, 2013).

TFs recognize TFBS in a sequence-specific manner as revealed by structural studies of protein–DNA complexes (Paillard and Lavery, 2004; Rohs et al., 2010) and next-generation sequencing (NGS) techniques such as SELEX-seq and ChIP-seq (Table 1). In the last decade, these NGS techniques have revolutionized the exploration of the TF binding landscape both *in vitro* and *in vivo* (Koboldt et al., 2013). This has resulted in many databases for TFBS deposition and profiling, such as TRANSFAC (Matys et al., 2006), JASPAR (Khan et al., 2018), UniPROBE (Hume et al., 2015), HOCOMOCO (Kulakovskiy et al., 2013), CIS-BP (Weirauch et al., 2014), and SwissRegulon (Pachkov et al., 2013). Such efforts have substantially boosted our understanding of interactions between TFs and TFBSs in different species, tissues, and developmental stages. While a

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

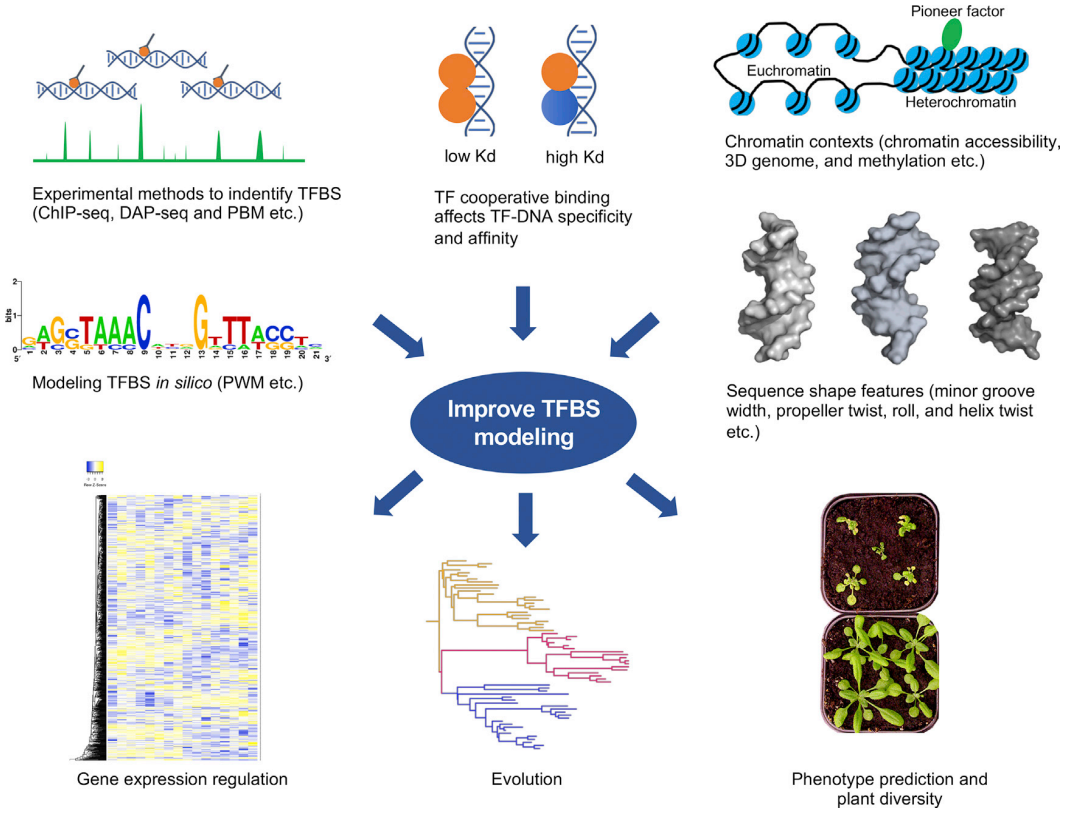


Figure 1. Schematic Overview of the TFBS Modeling and Application.

great deal of progress has been made to map TFBS, the resulting models are often poorly predictive of actual gene regulation. This can be due to poor modeling and prediction of TFBS, nonproductive TF binding (i.e., that does not result in gene regulation), or a combination of the two. Here, we focus on the more tractable question of how to model TFBS. As transcriptional regulation is a highly dynamic process that occurs in a cell- and tissue-specific manner, to better understand such a complex process unbiased quantitative modeling of TFBS with improved prediction power of TF binding is highly demanded. This includes taking into account of variables such as nucleosome positioning, chromatin states, methylation patterns, and the 3D structure of the genome, all of which greatly affect TF binding and, for a subset of these binding events, gene expression. Therefore, these variables need to be incorporated into any model to better describe functional TF binding *in vivo* and the concomitant gene regulation.

In this review, we address how TFBS are identified experimentally, how the TFBS models can be built *in silico*, and their optimization strategies. We further integrate context variables into the TFBS model in order to better understand gene regulation networks, evolution, and plant diversity (for an outline of the review, refer to Figure 1). Throughout the review, we use examples from case studies of TFs that are involved in flower development, a developmental transition that involves the activation of a wealth of genes that are otherwise silent, and the concomitant repression of a subset of genes.

TFBS MODELING

TFs read genomic DNA sequences in three fundamental ways, namely base readout, indirect readout, and shape readout (Rohs et al., 2010; Slattery et al., 2014). In base readout, TFs recognize a given nucleotide sequence by physical interactions between amino acid side chains and accessible edges of the base pairs of DNA. These interactions include hydrogen bonding, hydrophobic interactions, and the formation of salt bridges. Indirect readout involves mostly interactions between the TF and the DNA phosphate backbone, whose position is influenced by the nature of the base but not as strongly as in base readout. In shape readout (Abe et al., 2015; Yang et al., 2017), TFs recognize the structural features of DNA, such as DNA bending, groove width, and unwinding (Stella et al., 2010; Chen et al., 2013; Hancock et al., 2013). Although once considered as mutually exclusive driving forces for DNA recognition, recent studies have shown that most TFs likely combine base, indirect, and shape readout to recognize their TFBSs. Indeed, the integration of these features has been shown to improve TFBS prediction (Zhou et al., 2015; Mathelier et al., 2016).

EXPERIMENTAL METHODS FOR IDENTIFYING TFBS

With the emergence of NGS technologies, many NGS-based methodologies, both *in vitro* and *in vivo*, have been developed for determining TFBSs. Here we concentrate on some of the most widely used and recently developed methods and discuss their advantages and limitations (Table 1).

Experimental method	Description	*In vivo* or *in vitro*	DNA ligand	Unique features	TF source	References
PBM (protein binding microarrays) and variants	PBM uses microarrays of randomized DNA to which TF binding can be assayed by fluorescent antibodies to the TF	*In vitro*	Synthetic and randomized	High-throughput	Recombinant, usually fused with tags, such as GST	Berger et al., 2006; Berger and Bulyk, 2009
ChIP-seq (chromatin immunoprecipitation followed by sequencing) and variants	ChIP-seq couples chromatin immunoprecipitation with massively parallel sequencing, is capable of mapping genome-wide TFBSs *in vivo*	*In vivo*	Genomic	High-throughput and most widely used protocol for TFBS mapping *in vivo*	Endogenous TF or fused with an epitope tag	Johnson et al., 2007; Robertson et al., 2007; Kaufmann et al., 2010; Rhee and Pugh, 2011
SELEX-seq (systematic evolution of ligands by exponential enrichment followed by sequencing) and variant.	SELEX-seq uses recombinant TF to IP randomized DNA sequence in one or more cycles. The enriched DNA are sequenced by NGS and used to infer a model of specificity, typically a PWM, for a TF	*In vitro*	Synthetic and randomized	High-throughput, widely used, and allows to detect effects of DNA methylation (methyl-SELEX) and TF cooperative binding (CAP-SELEX)	Recombinant or *in vitro* translated TF	Jolma et al., 2010, 2015; Yin et al., 2017
ORGANIC (occupied regions of genomes from affinity-purified naturally isolated chromatin)	ORGANIC applies MNase to digest non-crosslinked chromatin then perform affinity purification by TF followed by NGS sequencing	*In vivo*	Genomic	Avoiding sonication and crosslinking	Endogenous TF or fused with an epitope tag	Kasinathan et al., 2014
BunDLE-seq (binding to designed library, extracting, and sequencing)	BunDLE-seq provides quantitative measurements of TF binding to thousands of fully designed sequences of 200 bp in length within a single experiment	*In vitro*	Synthetic and randomized	Allows comprehensive characterization of TF binding determinants within and outside of core binding sites	Recombinant TF	Levo et al., 2015
ChEC-seq (chromatin endogenous cleavage followed by sequencing)	ChEC-seq uses fusion of a TF to MNase to target calcium-dependent cleavage to specific genomic loci *in vivo*	*In vivo*	Genomic	Rapidly inducible nature of ChEC-seq allows separation of TFBSs based on their recognition by DNA sequence and shape or shape alone	Fused with MNase, produced *in vivo* under native or high inducible promotor	Zentner et al., 2015
ChIPmentation	ChIPmentation introduces sequencing-compatible adaptors in a single-step reaction directly on bead-bound chromatin, which reduces time, cost, and input requirements, thus providing a convenient and broadly useful alternative to existing ChIP-seq protocols	*In vivo*	Genomic	Avoids sequencing adaptor dimers which are common in standard ChIP-seq protocol, and requires only a single DNA purification step before library amplification	Endogenous TF	Schmidl et al., 2015

Table 1. Experimental Methods to Identify TFBS.

(Continued on next page)

Experimental method	Description	*In vivo* or *in vitro*	DNA ligand	Unique features	TF source	References
(amp)DAP-seq (DNA affinity purification followed by sequencing)	DAP-seq uses recombinant TF to affinity-purify genomic DNA fragments followed by NGS sequencing, capable of derive cistrome; ampDAP-seq applies PCR amplification to remove methylation patterns of fragmented genomic DNA before affinity purification, capable of deriving epicistrome	*In vitro*	Genomic	Allows low-cost and high-throughput generation of cistrome and epicistrome maps for hundreds of TFs of an organism	Recombinant or *in vitro* translated TF	O'Malley et al., 2016; Bartlett et al., 2017
SMiLE-seq (selective microfluidics-based ligand enrichment followed by sequencing)	SMiLE-seq applies microfluidics-based technology to perform a rigorous on-chip isolation of interacting TF–DNA complexes, allows robust identification of DNA binding specificities of TF monomers, homodimers and heterodimers	*In vitro*	Synthetic and randomized	Distinguish TF binding specificity from TF monomers and dimers of a TF (or hetero-/oligo-dimers of TFs) by microfluidics	Recombinant or *in vitro* translated TF	Isakova et al., 2017
SLIM-ChIP (short-fragment-enriched, low-input, indexed MNase ChIP)	SLIM-ChIP combines enzymatic fragmentation of chromatin and on-bead indexing to map high-resolution binding landscape of a TF	*In vivo*	Genomic	Low material input and allows mapping DNA binding proteins and charting the surrounding chromatin occupancy landscape at a single-cell level	Endogenous TF	Gutin et al., 2018
CUT&RUN (cleavage under targets and release using nuclease)	CUT&RUN is an epigenomic profiling strategy in which antibody-targeted controlled cleavage by MNase releases specific protein–DNA complexes into the supernatant for NGS sequencing	*In vivo*	Genomic	Avoids crosslinking and solubilization issues, and requires less sequencing depth	Endogenous TF	Skene and Henikoff, 2017, 2018)
Methyl-Spec-seq	Methyl-Spec-seq measures the effects of CpG methylation (mCPG) on TF binding affinity, allowing quantitative assessment of the effects at every position in a binding site.	*In vitro*	Synthetic and randomized	Facilitates the quantitative modeling of mCpG effects on gene regulation	Recombinant TF	Zuo et al., 2017
ChIP-STARR-seq	ChIP-STARR-seq combines ChIP with a massively parallel reporter assay to identify functional enhancers genome-wide in a quantitative manner. This method is potentially applicable in plant system	*In vivo*	Genomic	ChIP-STARR-seq allows high-throughput identification of functional enhancer *in vivo*	Endogenous TF	Barakat et al., 2018

Table 1. ***Continued***

Chromatin immunoprecipitation sequencing (ChIP-seq) has long been the gold standard for detecting genome-wide TFBSs bound by a given TF *in vivo* (Johnson et al., 2007; Robertson et al., 2007; Kaufmann et al., 2010). In a standard ChIP-seq protocol, sample tissues are treated with a crosslinking reagent and subjected to nuclei purification to isolate chromatin containing TF–DNA complexes. Generally, an additional step of chromatin shearing by sonication is applied before the final step of chromatin immunoprecipitation (IP) using a TF-specific antibody. The IP product containing enriched DNA fragments that are recognized by the TF of interest is then subjected to NGS sequencing. ChIP-seq has been successfully used routinely in many laboratories. However, standard ChIP-seq protocols have intrinsic limitations and technical drawbacks (Park, 2009). One of the limitations comes from sonication, a process that is highly irreproducible and produces variable DNA fragment sizes that are difficult to sequence. The other limitation is crosslinking, which produces low signal-to-noise ratio and many false positives. To overcome such limitations, many ChIP-seq variant methods have been developed, including ORGANIC (Kasinathan et al., 2014), ChEC-seq (Zentner et al., 2015), CUT&RUN (Skene and Henikoff, 2017), and SLIM-seq (Gutin et al., 2018) (see Table 1 for unique features and details of these methods), all of which use micrococcal nuclease (MNase) to fragment chromatin, therefore avoiding sonication. Due to the mild conditions of DNA fragmentation by MNase, these methods do not denature or disrupt TF–DNA complexes and eliminate the requirement of crosslinking. These protocols also require substantially lower amounts of input materials and are thus feasible for low-input applications. Processing of IP enriched DNA fragments for downstream NGS application poses another challenge in ChIP-seq and is often time consuming. To simplify the process, ChIPmentation applies Tn5 transposase directly to bead-bound chromatin, allowing single-step NGS-compatible DNA library preparation (Schmidl et al., 2015) (Table 1).

ChIP-seq and its variants not only identify TFBSs *in vivo*, but also provide a wealth of information such as detection of binding sites bound by co-binders of the TF. As such, however, this also poses a challenge in distinguishing the true TFBSs from indirect binding mediated by a TF partner. ChIP-seq can be complemented by DNA binding assays performed *in vitro* using recombinant TFs. Among the most widely used *in vitro* techniques are protein binding microarrays (PBM) (Berger et al., 2006; Berger and Bulyk, 2009) and SELEX-seq (Jolma et al., 2010) (Table 1). Both methods allow high-throughput identification of TF binding specificities *in vitro*, with such information useful to predict TFBS in genomic sequences; however, they employ synthetic randomized DNA that lack at least some genomic DNA sequence properties known to affect TF binding, including nonphysiological primary sequences, core motif flanking regions, and lack of chemical modifications, such as cytosine methylation. To overcome these biases, DAP-seq (DNA affinity purification sequencing) has been recently developed, which uses fragmented genomic DNA as substrates for IP and recombinant TFs (O'Malley et al., 2016; Bartlett et al., 2017) (Table 1). As DNA methylation patterns are conserved in genomic DNA, DAP-seq allows genome-wide mapping of the epicistrome and the discovery of TF binding specificity from genomic DNA. Furthermore, when combined with ampDAP-seq, which uses amplified and thus demethylated genomic DNA as substrates, a comprehensive mapping of both the cistrome and the epicistrome can be derived for a given TF. Compared with ChIP-seq and its variants, DAP-seq can be performed in a high-throughput manner with lower costs, as recently demonstrated (O'Malley et al., 2016). Despite these advantages, DAP-seq has its limitations; for example, some TFs are not stable when recombinantly expressed and are thus not compatible with DAP-seq, while others require interacting partners for their DNA binding activity, and many TFs have distinct DNA binding properties in the presence of co-factors. These limitations have to be taken into account during experimental design and data analysis. Moreover, it has to be noted that DAP-seq lacks cellular chromatin context, therefore, combination of *in vitro* DAP-seq and *in vivo* ChIP-seq would be an informative approach regarding TFBS modeling and *in vivo* TF binding prediction.

MODELING TFBS *IN SILICO*

To make accurate *de novo* prediction of binding sites of a given TF in the genome, a quantitative TFBS model that is representative of TF–DNA binding affinity is required. This could be derived from a set of known TFBSs using computational methods. Here we discuss how conventional modeling methods could be improved by integrating complex features, such as sequence position dependencies and DNA shape features, which have been shown to play a role in determining TF–DNA specificity. We focus on the most recent and representative TFBS modeling algorithms (Table 2); other algorithms have been extensively reviewed elsewhere (Tompa et al., 2005; Hombach et al., 2016).

Position Weight Matrix

Position weight matrix (PWM) is the most widely used model to represent TF–DNA binding specificity (Schneider and Stephens, 1990; Stormo and Zhao, 2010). In brief, from a collection of TFBSs a matrix is built that gives the frequency of each nucleotide at each position of the motif. Based on these frequencies, a PWM or position specific scoring matrix can be computed that gives a log-scale value to each nucleotide at each position. Based on the PWM, a score can be calculated for any sequence corresponding to the sum of all values at each position. The logo representation of a PWM illustrates the information content at each position and represents the four bases depending on their frequency (Figure 2).

Dependencies

PWMs provide good approximation of TF–DNA interactions in most cases, and can be generated from various datasets, ranging from a small set of known TFBSs to TF–DNA binding data derived from high-throughput assays. However, standard PWM assumes that each position within a TFBS contributes to binding affinity independent of other positions, and is thus unable to represent inter-base dependencies, which have been observed for some TFs (Bulyk, 2002; Tomovic and Oakeley, 2007; Badis et al., 2009). Various models that take into account these dependencies have been shown to outperform standard PWM in *de novo* prediction. For example, the MORPHEUS program allows to introduce di- and trinucleotide position dependencies in PWM and has been successfully applied to plant TFs with, in some cases, improved predictive power (Moyroud et al., 2011; Minguet et al., 2015) (Table 2).

Several approaches can be taken with respect to how and what dependencies are to be integrated into the modeling algorithm. Some consider pairwise dependencies between adjacent and/or distal positions, such as the binding energy model (BEM) (Zhao et al., 2012), dinucleotide weight matrices (DWM) (Siddharthan, 2010), and TF Flexible Model (TFFM) (Mathelier and Wasserman, 2013) (Table 2). Others introduce higher-order *k*-mer features that take into account all possible sequences with length *k*, such as the feature motif model (Sharon et al., 2008) (Table 2). In some cases, model complexity can increase dramatically when arbitrary positions or unconstrained *k*-mer features are used and become prone to be overfitting. Alternative approaches start from a model without dependencies, and use a greedy algorithm to improve

TFBS modeling methods	Description	Features integrated	Web server or source code	Motif representation	References
PWM (position weight matrix)	PWMs are normalized representations of the position-specific log-likelihoods of a nucleotide's probability to occur at each position in a sequence	NA (not applicable)	NA	PWM logo	Stormo et al., 1982; Schneider and Stephens, 1990
DWM (dinucleotide weight matrix)	DWM considers the 16 combinations of dinucleotide instead of the 4 nucleotides used for PWM	Dinucleotides	NA	DWM logo	Siddharthan, 2010
BEM (binding energy model)	BEM introduces energy parameters of adjacent nucleotides to the binding affinity quantification	Dependencies (adjacent positions) and binding affinity data	http://stormo.wustl.edu/TF-BEMs/	Binding energy logo	Zhao et al., 2012
TFFM (TF Flexible Model)	TFFM model integrates a Markov model to take dependencies and background into account	Dependencies (adjacent position) and background	http://cisreg.cmmt.ubc.ca/cgi-bin/TFFM/TFFM_webapp.py?rm=start	TFFM logo	Mathelier and Wasserman, 2013
PIM (pairwise interaction model)	PIM is based on the principle of maximum entropy and describes pairwise correlations between nucleotides at different positions	Dependencies between all positions	https://github.com/msantolini/PIM	PWM mixture model	Santolini et al., 2014
gkm-SVM (gapped *k*-mer support vector machine)	gkm-SVM predicts regulatory sequence using gapped *k*-mer features	*k*-mers supporting gaps	http://www.beerlab.org/gkmsvm/	NA	Ghandi et al., 2014
SeqGL	SeqGL is a *de novo* motif discovery algorithm to identify multiple TF sequence signals from ChIP-seq, DNase-seq, and ATAC-seq profiles	*k*-mer, chromatin accessibility	http://cbio.mskcc.org/public/Leslie/SeqGL/	NA	Setty and Leslie, 2015
MORPHEUS	MORPHEUS is a webtool for TF binding analysis using PWM with dependencies	Dependencies between all positions	http://biodev.cea.fr/morpheus/	PWM logo with dependencies	Minguet et al., 2015
FeatureREDUCE	FeatureREDUCE provides a flexible framework for building sequence-to-affinity models from PBM data	Dependencies between all positions	http://software.bussemakerlab.org	NA	Riley et al., 2015
Cytomod	Cytomod models methyl-sensitive TF motifs with an expanded epigenetic alphabet	DNA methylation	NA	PWM logo with an extended alphabet (e.g., 5mC stands for 5-methylcytosine)	Viner et al., 2016
DeepBind	DeepBind can learn several motifs to predict binding sites of DNA and RNA binding proteins	NA	http://tools.genes.toronto.edu/deepbind/	Weighted ensemble of PWM logos	Alipanahi et al., 2015

Table 2. TFBS Modeling Methods.

(Continued on next page)

TFBS modeling methods	Description	Features integrated	Web server or source code	Motif representation	References
DeepSEA (deep learning-based sequence analyzer)	DeepSEA predicts effects of noncoding variants with deep learning-based sequence model	Integrate DNase I hypersensitivity data and histone-mark profiles	http://deepsea.princeton.edu/job/analysis/create/	NA	Zhou and Troyanskaya, 2015
DWT (dinucleotide weight tensor)	DWT is a regulatory motif model that incorporates arbitrary pairwise dependencies for TFBS prediction	Dependencies between all positions	http://dwt.unibas.ch/fcgi/dwt	DWT "dilogo" motifs	Omidi et al., 2017
TFImpute	TFImpute predict cell-specific TF binding trained by deep learning	NA	https://bitbucket.org/feeldead/tfimpute	NA	Qin and Feng, 2017
BEESEM (short for Binding Energy Estimation on SELEX with Expectation Maximization)	BEESEM estimates BEMs using SELEX-seq data based on expectation maximization method	NA	http://stormo.wustl.edu/beeml/	NA	Ruan et al., 2017
DeFine	DeFine quantifies TF–DNA binding affinity and facilitate evaluation of functional noncoding variants in the genome based on deep learning algorithms	Integrate Hi-C data	http://define.cbi.pku.edu.cn	PWM logo	Wang et al., 2018a
DFIM (Deep Feature Interaction Maps)	DFIM estimates pairwise interactions between features (such as nucleotides or subsequences) in any input DNA sequences by a neural network	Dependencies between all positions, interaction between motifs, core motif flanking region, and chromatin accessibility	https://github.com/kundajelab/dm.	DFIM with feature importance scores	Greenside et al., 2018
NRLB (No Read Left Behind)	NRLB provides scalable and quantitative method to identify functional *in vivo* binding sites of TF and to define relative binding affinities for any TF–DNA complex	NA	NA	Energy logo representation	Rastogi et al., 2018
KSM model (*k*-mer set memory)	A *k*-mer finder (KMAC) finds *k*-mers that are over-represented in TFBSs, and KSM allow accurate regulatory variant prediction	*k*-mers	https://github.com/gifford-lab/GEM3	KSM (*k*-mer set memory)	Guo et al., 2018
SelexGLM	SelexGLM incorporates core motif flanking region for TFBS binding quantification	Core motif flanking region	https://www.bioconductor.org	Energy logo representation	Zhang et al., 2018a

Table 2. ***Continued***

A

Site 1	T	T	T	G	T	C	G	C	A	T
Site 2	T	T	T	G	T	C	G	A	A	T
Site 3	C	A	T	G	T	C	G	G	T	C
Site 4	C	A	T	G	T	C	G	G	T	C
…	…	…	…	…	…	…	…	…	…	…
Site N	G	T	T	G	T	C	T	G	G	A
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

B

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	68	102	2	3	1	4	2	53	131	239
C	182	49	4	2	3	596	4	25	173	65
G	65	26	44	597	2	2	467	494	145	67
T	292	430	557	5	601	5	134	35	158	236

C

$$W(b,i) = \ln\frac{f(b,i)}{f_{exp}(b)}$$

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	-0.8	-0.4	-4.3	-3.9	-5.0	-3.6	-4.3	-1.1	-0.1	0.5
C	0.2	-1.1	-3.6	-4.3	-3.9	1.4	-3.6	-1.8	0.1	-0.8
G	-0.8	-1.8	-1.2	1.4	-4.3	-4.3	1.1	1.2	0.0	-0.8
T	0.7	1.0	1.3	-3.4	1.4	-3.4	-0.1	-1.5	0.0	0.4

D

$$S = \sum_{i}^{n} W(b_i.i)$$

A	T	T	A	G	G	C	C	C	G
-0.8	1.0	1.3	-3.4	-4.3	-4.3	3.6	-1.8	0.1	-0.8

∑ = -9.4

E

Figure 2. Workflow to Generate PWM for a Set of Known TFBS.
Here, ARF5 bound sequences are retrieved from DAP-seq (O'Malley et al., 2016). The sequences of *N* different binding sites are aligned **(A)**. The nucleotide frequency is computed at each position of the binding site to yield the position frequency matrix **(B)**, which is then converted to a PWM **(C)**. In the formula, *W*(*b*, *i*) stands for the weight of a nucleotide *b* at position *i*, *f*(*b*,*i*) is the frequency of this nucleotide, and f_{exp} is the expected background frequency of the given nucleotide. If each nucleotide appearance is equal, one can take $f_{exp} = N/4$ **(C)**. One can score a given sequence by summing the corresponding PWM weights **(D)**. The TFBS logo represents the preference of the TF at each position of the binding site **(E)**. This calculation represents one possible method among several to calculate a sequence score (Stormo, 2000, 2013; Wasserman and Sandelin, 2004).

the model by adding dependency features iteratively (Hu et al., 2010; Santolini et al., 2014). Thus, dependency features are iteratively added until no further feature could be found to improve the model. Others use Bayesian Markov models (BaMM) of order *k* that take into account dependencies between one nucleotide and the *k* previous positions (Kiesel et al., 2018). Complex models integrating dependency features generally outperform simple PWM models; however, some of these models require more expertise to apply and repeated manual attempts to be trained correctly and are thus not facile to use. This constitutes one of the limiting factors that restricts these models from being used routinely in the community. In Table 2 we summarize features of some of the most recent models.

Shape Features

Sequence-based models provide accurate estimation of base readout, however, it cannot explain why some TFs, which have highly conserved DNA binding domains, bind different sequences genome-wide. For example, the TF paralogs, androgen and glucocorticoid receptors, which bind similar DNA motifs by a set of identical amino acids (Shaffer et al., 2004; Meijsing et al., 2009), share only a third of their genomic binding sites (Zhang et al., 2018a). It turns out that DNA shape features contribute significantly to distinguish bona fide TFBSs from others. In the last decade, many studies have revealed that DNA shape features indeed play an important role for determining TF–DNA binding specificity (Rohs et al., 2009; Abe et al., 2015; Yang et al., 2017). A most recent example is the MADS-box TF, SEPALLATA3 (SEP3), a key regulator of flower organ specification (Muiño et al., 2014; Hugouvieux et al., 2018). MADS-box TFs bind to CArG-boxes with consensus sequence of 5′-CC(A/T)$_6$GG-3′, yet only a fraction of the CArG-boxes available genome-wide is bound by SEP3. Käppel et al. (2018) showed that SEP3-DNA binding affinity correlates well with the width of the minor groove of CArG-box probes, a shape readout mechanism involving a conserved arginine residue that contacts minor groove. Although shape features are mainly determined by the TFBS core motif, it can also be affected by flanking regions. In the past, these regions have been overlooked in characterizations of TF binding due to their low sequence information. Now both bioinformatics analysis and biochemical evidence have accumulated pointing toward their importance for TF binding. For example, Dror et al. (2015) showed a widespread role of the motif environment in TF binding by analyzing binding data from SELEX-seq and ChIP-seq for some 300 TFs, and that the preference for a specific environment differs between distinct TF families. Selective binding of core motifs with different flanking sequences have also been observed by *in vitro* assays for several TFs (Gordân et al., 2013; White et al., 2013; Afek et al., 2014; Levo et al., 2015).

Introducing shape features into TFBSs modeling requires integrating several distinct shape parameters, including Minor Groove Width, Propeller Twist, Roll and Helix Twist. These features have been shown to be distinguished by different TFs (Yang et al., 2014). Very recently, nine additional shape features were introduced to the repertoire in order to better describe the unique 3D structure encoded in a given DNA sequence (Li et al., 2017). Apart from "naked" DNA shape features, DNA methylation on cytosine residues also affects DNA structure, making it a unique type of shape feature that could be recognized by many TFs (Lazarovici et al., 2013; Yin et al., 2017; Rao et al., 2018). Several TFBS modeling methods that take into account shape features (some combined with sequence-based features) have been developed, and show improvement compared with only sequence feature-based models (Table 2). However, these models use DNA shape information generated from computational simulations, such as Monte Carlo or molecular dynamics, and potential biases exist. Improvements have already been obtained by integrating

DNA shape information derived from experimental data, such as X-ray crystallography (Li et al., 2017). Thus, one major challenge regarding incorporating shape features into TFBS modeling is to derive unbiased DNA structural data in a high-throughput manner that has been robustly verified experimentally, which currently is a challenge. The other challenge is that both DNA conformation (Azad et al., 2018) and TF conformation (Patel et al., 2018) could be changed in an adaptation mechanism upon interacting with each other due to both protein and DNA plasticity. This makes integrating shape features even more difficult as they changes dynamically.

Energy-Based and Deep Learning-Based Models

Energy-based biophysical models are a powerful alternative to probabilistic models such as PWM. They use the action mass law to characterize amino acid and DNA interaction and are valid on a wider range of protein concentrations than probabilistic models, which in fact represent an approximation of energy-based models. Whenever they can be built, and several methods exist based on PBM or SELEX-seq, for example, they should be preferred without disadvantages except that PWM is the simplest to build (Zhao et al., 2009; Stormo, 2013; Ruan and Stormo, 2017).

Machine learning methods, such as deep learning, are able to leverage very large datasets to discover intricate connections within them and make accurate predictions (Lecun et al., 2015). In the last few years, deep learning has been increasingly applied to resolve complex biological problems, including those from regulatory genomics (Angermueller et al., 2016). Several methods based on deep learning have been developed to model TF–DNA binding specificity or to predict TF *in vivo* binding, including DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou and Troyanskaya, 2015), TFImpute (Qin and Feng, 2017), DeFind (Wang et al., 2018a) and DFIM (Greenside et al., 2018) (Table 2). Advantages of these models include, for example: (1) they can be trained from various types of sequencing data either alone or in integrated manner, and can be further combined with other information, such as DNase I hypersensitivity data, for better *in vivo* TFBSs prediction (Zhou and Troyanskaya, 2015); (2) they can tolerate a certain degree of noise stemming from either data acquisition technology or sequencing biases; (3) they can train predictive models fully automatically, alleviating the need for time-consuming manual intervention and expertise; (4) they can accurately identify genomic variants in the regulatory region, and indicate how variations affect TF binding within a specific sequence. However, one of the yet to be tackled difficulties of deep learning models is that they are more difficult to interpret than PWMs given the hidden layers in the networks. More information about these models and their unique properties can be found in Table 2. To conclude, it is worth mentioning that, until now, no single model has been identified to be the best for all TFs and the nature of the most adequate model depends on the individual TF.

Link between Models and TF 3D Structure

TFBS models derived from NGS allow a broad overview of where TFs are able to bind and their sequence specificity. Structures of TF–DNA complexes provide complementary information by identifying the amino acids and specific bases involved in TF–DNA interactions. These structural data not only explain base and shape readout at the residue and even atomic level, but also allow for the prediction of how amino acid mutations and/or changes in a given *cis* element will affect TF binding. Indeed, many diseases resulting from gene misregulation are due to either mutations in a TF or alterations in its binding site. Combining the "go broad" NGS approach with the "go deep" structural approach provides a powerful tool in refining TFBS and gene regulation models.

Recent modeling tools have attempted to use 3D structural data for improving predictions of TF–DNA binding, and structure-based databases for TFBS data are currently available (Turner et al., 2012; Lin and Chen, 2013; Xu et al., 2013). Structure-based TFBS methods rely on different energy functions to score TF–DNA interactions. Such energy functions are used to describe all possible physiochemical interactions such as Van der Waals interactions, hydrogen bonding, electrostatic interactions, and solvation energy. Energy functions can be divided into physics-based molecular mechanics force fields (Liu et al., 2009a; Marcovitz and Levy, 2011; Yin et al., 2015) and knowledge-based potentials (Liu et al., 2005; Zhang et al., 2005; Takeda et al., 2013). While physics-based energy functions are able to accurately describe TF–DNA interactions, they have a high computational cost and thus are less often applied than knowledge-based potentials. In knowledge-based potentials, statistical analysis is used to describe TF–DNA interactions at the atom or residue scale using known TF–DNA structures. These are simpler and less computationally expensive than physics-based energy models. Recent work combining aspects of both types of models to derive an "integrative energy" function have also been applied to TF–DNA modeling and have been shown to further improve, in some cases, the predictive power of structure-based TFBS models (Farrel et al., 2016; Farrel and Guo, 2017).

A second way that 3D structural data can be used to help refining TFBS models is through the prediction of protein–protein interactions (PPIs), which may affect TF binding to DNA. Often *in vitro* TFBS models are relatively poor predictors of *in vivo* TF binding due to the added complexity of interacting proteins *in vivo*. Pull-down assays, mass spectrometry, and yeast two-hybrid assay allow the generation of at least a partial interacting network for a given TF (Yazaki et al., 2016; Trigg et al., 2017). These methods have limitations and often generate incomplete models due to the difficulty in determining true interaction partners and in detecting rare or transient interactions. Structural data can be incorporated to improve PPI models by providing quantitative parameters to determine whether a putative interaction is likely to occur based on energy calculations or homology modeling (Aloy and Russell, 2006; Beltrao et al., 2007). By adding partners to the simple TF–DNA model, differences between *in vitro* and *in vivo* binding are better accounted for and perturbations due to mutations, for example, can be more easily modeled as has been shown for mammalian TFs (Guturu et al., 2013). To our knowledge a full integration of structural data with TFBS models has not been implemented for plant TFs; however, as many TF families are conserved across the kingdom of life, these methods may be applicable to plant TFs.

IMPROVING THE PREDICTIVE POWER OF TFBS MODELS: GENOME CONTEXT

Eukaryotic genomes contain numerous potential binding sites for a given TF; however, only a small fraction is actually bound *in vivo*, and these sites vary substantially depending on contexts, such as cell types, developmental stages, and environmental or cellular conditions. In addition, only a subset of the bound sites drive transcription (Wasserman and Sandelin, 2004; Hu et al., 2007; Fisher et al., 2012; Whiteld et al., 2012). Therefore, various contexts have to be taken into account to predict functional TFBSs precisely. This includes chromatin states (such as accessibility and epigenetic marks), methylation states, nucleosome positioning, genome 3D structures, and combinatorial binding of TFs.

Nucleosome Positioning, Chromatin States, and 3D Genome

In the nucleus of eukaryotic cells, DNA wraps around histone proteins to form nucleosomes (McGinty and Tan, 2015), which can be further compacted into a highly condensed structure called heterochromatin by various mechanisms (Allshire and Madhani, 2018). This involves factors such as linker histones (Fyodorov et al., 2017), repressive histone marks (Allis and Jenuwein, 2016) such as H3K27me1/3 and H3K9me2, and DNA methylation on cytosine residues (Kim and Zilberman, 2014; Zhu et al., 2016), among others. Thus chromatin structure is intrinsically repressive, a mechanism that helps to establish stable gene expression and prevents unwanted cell fate transitions. For gene activation, eukaryotic cells evolved various counter mechanisms for each of the chromatin compacting factors to create accessible chromatin, such as active histone marks (e.g., H3K4me2/3 and H3K27ac), chromatin remodelers (Ho and Crabtree, 2010), and demethylation machineries (Wu and Zhang, 2014). The interplay between all these factors results in a highly dynamic chromatin environment, in which TFs have to find their cognate DNA binding sites.

Nucleosome Positioning

In general, TFs preferentially bind to TFBSs in accessible chromatin regions, where nucleosomes are depleted (nucleosome-depleted region [NDR]). This is evidenced by large-scale *cis*-element studies, which showed that the vast majority of the active *cis* elements reside in the NDR in different species (Thurman et al., 2012; Weber et al., 2016), including *Arabidopsis* and maize (Zhang et al., 2012; Vera et al., 2014). Therefore, a precise *in vivo* TFBS prediction model could integrate NDR as its first layer of filter to leave out sites/regions with well-positioned nucleosomes. Indeed, several TFBS modeling methods that integrate DNase I hypersensitivity datasets have shown increased prediction power for *in vivo* binding (Zhou and Troyanskaya, 2015; Kelley et al., 2016; Wang et al., 2018b). Thus, it is essential to generate datasets representing chromatin accessibility. To address this, recent technological advances have become available, such as DNase-seq, MNase-seq, FAIRE-seq, and ATAC-seq (Meyer and Liu, 2014). Among these, ATAC-seq is a rising star method as it requires a minimal amount of input sample and even can be used at the single-cell level (Buenrostro et al., 2013, 2015; Corces et al., 2017). This is particularly attractive for the plant biology community, where some plant tissues are extremely scarce, such as flower meristem cells, organ primordia, and root tips. Furthermore, when combined with INTACT (isolation of nuclei tagged in specific cell types), which allows isolation of nuclei from individual cell types of a tissue by affinity-based purification, ATAC-seq is able to map chromatin accessibility with high resolution and low noise from a specific cell type (Deal and Henikoff, 2011; Sijacic et al., 2018).

Although a majority of TFs favor binding in NDRs, exceptions exist. A special group of TFs, so-called pioneer factors, are able to bind TFBSs even when nucleosomes are present (Iwafuchi-Doi and Zaret, 2016; Zaret and Mango, 2016; Zaret, 2018). As exemplified by FoxA1 and GATA4, pioneer factors are able to outcompete nucleosomes or create NDR through various mechanisms, such as mimicking linker histones, recruiting chromatin remodelers, and/or depositing active epigenetic marks (Mayran and Drouin, 2018). Therefore, pioneer factors have to be considered with care while modeling their *in vivo* binding. One of the first reported plant pioneer factors was LEAFY COTYLEDON1 (LEC1), a seed-specific TF and a master regulator of embryogenesis. Tao et al. (2017) showed that LEC1 can target mitotically silenced chromatin at the loci of floral repressor *FLOWERING LOCUS C* (*FLC*) and promote the initial establishment of an active chromatin state. This activates FLC expression *de novo* in the pro-embryo and leads to the reversal of the silenced chromatin state inherited from gametes. Three TFs, LEAFY (LFY), APETALA1 (AP1), and SEP3, which are key factors in floral development in *A. thaliana*, have been shown to be likely pioneer factors. A combination of ChIP-seq and DNase-seq data suggested that LFY is able to bind its TFBSs in closed chromatin, and this activity is highly correlated with its oligomerization activity. This is a potential novel driving force for pioneer activity, which has not been reported in other organisms (Sayou et al., 2016). For AP1 and SEP3, it has been shown that upon binding to their TFBSs both TFs are able to confer chromatin accessibility near those sites (Pajoro et al., 2014). Interestingly, both factors are able to form higher-order homo- and hetero-oligomers, with such activity essential for their function *in vivo*. Therefore, an attractive hypothesis is that oligomerization likely confers high binding affinity in order for them to bind TFBSs that are otherwise inaccessible due to the occupancy of histones at these sites. Although further evidence of pioneer activity of these TFs, including both genome-wide and biochemical studies, are required, modeling their *in vivo* binding requires examination of both closed and open chromatin regions.

Chromatin States: Histone Modifications, Histone Variants, and Chromatin Remodelers

TF binding *in vivo* confronts various chromatin states that are established by various types of histone modifications, histone variants, and remodelers. Histone modifications act as either active or repressive marks, corresponding to transcriptionally competent and inactive chromatin, respectively. These marks are deposited by epigenetic writers and removed by epigenetic erasers. For instance, the PRC2 is a writer responsible for H3K27me3 deposition while the REF6 demethylase erases this mark (Hennig and Derkacheva, 2009; Li et al., 2016). For some, if not all, epigenetic marks there is a corresponding epigenetic reader that reads the specific mark and confers downstream responses. Histone variants are also

determinants of chromatin states and affect transcription. For instance, the H3.3 and H3.1 variants differ only by four amino acids (Ingouff and Berger, 2010), and while H3.1 is enriched in heterochromatin and preferentially carries repressive H3K27 methylation marks, H3.3 is enriched in transcriptionally active regions and preferentially carries active H3K36 methylation marks (Johnson et al., 2004; Stroud et al., 2012). Chromatin remodelers, which use ATP energy to evict, disassemble, or slide nucleosomes, are also landmarks affecting TF binding. The increasing datasets for genome-wide profiling of histone variants, marks, writers, erasers, readers, and chromatin remodelers thus constitutes a highly informative resource to improve TFBS prediction.

Crosstalk between TFs and chromatin factors co-regulate chromatin accessibility and exposure of *cis* elements (Vachon et al., 2018). In these processes, TFs operate either by recruiting chromatin factors or directly competing with them for target sites. There are several examples of TF-mediated recruitment of chromatin factors in plants, such as that of REF6 by NF-Y TFs for H3K27 demethylation at *SOC1*, inducing flowering (Hou et al., 2014), or Polycomb mark reader TFL2/LHP1 recruitment by SHORT VEGETATIVE PHASE at *SEP3* for flower patterning (Liu et al., 2009b), or BRAHMA and SPLAYED ATPase recruitment by LFY and SEP3 at flower morphogenetic genes (Wu et al., 2012). By contrast, several TFs were shown to compete with Polycomb complexes at target genes, such as NF-YC, which prevents PRC2 binding to *FLOWERING LOCUS T* for floral transition (Liu et al., 2018) and AG, which evicts PRC2 from *KNUCKLES* for flower meristem termination (Sun et al., 2014). Interestingly, at the time of flower termination, AG also has the opposite effect at *WUS*, promoting PRC2 recruitment for deposition of H3K27me3 (Liu et al., 2011). Differences in TF behavior for eviction versus recruitment of PRC2 may depend on the distance between the Polycomb recognition element (PRE) and the TFBS. In this regard, large-scale analyses of ChIP-seq data revealed TFBSs in plant PREs, thereby expanding the repertoire of TF–chromatin factor interactions and providing resources for further exploration of the relationship between *cis* elements and TF/chromatin factor binding (Wang et al., 2016; Xiao et al., 2017). Taken together, intricate and dynamic interplays among TFs and chromatin factors have to be carefully examined for TF binding *in vivo* as they define the chromatin state of a region, where TFs in turn have to engage with.

Methylation State

DNA methylation at the 5′ position of cytosine plays an essential role in gene regulation and genome stability in plants and animals (Zhang et al., 2018b). Precise patterns of DNA methylation are crucial for plant growth and development, including flowering (Finnegan et al., 1998). Unlike animals, in which DNA methylation are predominantly found in the CG context, plant DNA methylation occurs in contexts including CG, CHG, and CHH (H represents A, T, or C) (Zhang et al., 2006; Lister et al., 2008). Most TFs favor not to bind to methylated TFBSs due to the fact that DNA methylation affects shape features of TFBSs and that methyl groups often clash with residues that form direct interactions with otherwise unmethylated DNA motifs. Interestingly, recent studies have revealed that some TFs preferentially bind to methylated DNA (Zhu et al., 2016; Yin et al., 2017; Zuo et al., 2017). In addition, these TFs seemed to be enriched in embryonic and organismal development, such as homeodomain TFs and pluripotent factors (e.g., OCT4), which are well-characterized pioneer factors. Although proteins that specifically bind to methylated DNA are found in plants as exemplified by methyl-CpG binding domain proteins (Zemach and Grafi, 2007), they are not classified as TFs but epigenetic modifiers. To our knowledge, TFs that are insensitive to methylation have not yet been reported in plants, however, it is appealing to investigate whether the aforementioned potential plant pioneer factors (i.e., LEC1, LFY, SEP3, and AP1) are insensitive to methylation. Another mechanism that affects TF binding is that widespread DNA methylation promotes repressive histone modifications such as H3K9me2 and inhibits permissive histone modifications such as histone acetylation, resulting in highly compacted heterochromatin (Zhang et al., 2018b), thus inaccessible to vast majority of the TFs except pioneer factors. Taken together, traditional views suggested that methylation seem to inhibit TF binding to TFBSs; however, there are likely at least a subset of TFs, such as pioneer factors, that can target methylated sites. Therefore, their DNA binding affinity and specificity need to be carefully examined with regard to prediction of their *in vivo* binding. There are several methods available to model the effects of DNA methylation, such as Cytomod (Viner et al., 2016) (Table 2), which uses the classical PWM approach with an extended alphabet (e.g., 5mC representing methylated cytosine). In some practices, multiple PWM logos are given for the same TF, for which the enriched methylated and nonmethylated sequences are represented separately (Yin et al., 2017). In addition, apart from the aforementioned DAP-seq, two additional experimental approaches are now available to investigate the effects of DNA methylation to TF binding *in vitro*, namely Methyl-Spec-seq (Zuo et al., 2017) and methyl-SELEX (Yin et al., 2017) (for details see Table 1).

3D Genome- and TF-Mediated Long-Range Gene Interactions

The linear nucleotide sequences are folded into highly organized 3D architectures in the nucleus of higher eukaryotes. Chromosome conformation capture (3C) techniques, such as Hi-C (Eagen, 2018), revealed widespread existence of long-range gene interactions within the so-called topologically associating domains (TAD). Within the TAD, distal and proximal *cis* elements relative to transcription start sites (TSS) form cell-type-specific long-range interactions that in many cases are established by architectural proteins such as cohesin (Yan et al., 2013; Rao et al., 2017), CTCF (Phillips and Corces, 2009; Ren et al., 2017), Yin Yang 1 (Weintraub et al., 2018), and others (Rada-Iglesias et al., 2018). Such interaction is a highly conserved mechanism for eukaryotes to achieve spatiotemporal gene expression (Sanyal et al., 2012; Harmston and Lenhard, 2013; Dekker and Misteli, 2015). TADs therefore form territories within which more frequent gene interaction occurs, whereas less interaction happens beyond these territories. Disruption of TAD boundaries can lead to ectopic activation of gene expression and eventually to noticeable phenotypes (Lupiáñez et al., 2015, 2016; Franke et al., 2016).

Although it seems that *A. thaliana* does not form TADs likely due to not having the architectural proteins such as CTCF

that is important for TAD maintenance (Liu et al., 2017), in many other plant species, such as maize, rice, and tomato, TADs are clearly detected according to Hi-C data (Dong et al., 2017). Nevertheless, long-range gene interactions are still widespread in the *Arabidopsis* genome but in a less compartmentalized manner compared with other plant species (Liu et al., 2016). Apart from architectural proteins, TFs are usually the links mediating cell-type-specific long-range gene interactions, for which the transactivation domains (TDs) found in a majority of TFs appear to play an essential role. In general, TDs are enriched with acidic and hydrophobic residues, and residues that are able to form intrinsically disorder structures, such as serine, glycine, and proline (Staller et al., 2018). These properties appear to allow TDs to interact with or recruit various factors with modest affinity but high specificity under various contexts. One such factor is Mediator, a mega protein complex that can be recruited by divergent TFs to connect distal and proximal *cis* elements (Soutourina, 2017). Another factor is the SAGA complex, which has recently been shown to be a general factor that is required for the construction of the pre-initiation complex at the TSS for transcription initiation in animal systems (Baptista et al., 2017). Despite being less well characterized in plants, homolog protein components for both factors are well conserved in plants (Elfving et al., 2011; Mathur et al., 2011; Moraga and Aquea, 2015).

With the accumulation of datasets from Hi-C and related methods, it is now possible to predict spatiotemporal TF binding more precisely. In plants, Hi-C has been carried out from species including *A. thaliana* (Liu et al., 2016), rice (*Oryza sativa*) (Dong et al., 2018), barley (*Hordeum vulgare*) (Mascher et al., 2017), tomato (*Solanum lycopersicum*), maize (*Zea mays*), sorghum (*Sorghum bicolor*), foxtail millet (*Setaria italica*) (Dong et al., 2017) and cotton (*Gossypium* spp.) (Wang et al., 2017, 2018c). Considering that chromosome conformation is highly cell-type specific, these Hi-C datasets have to be carefully examined when applied to other cell types. For example, in the process of flowering initiation it is more relevant to perform Hi-C using flower meristems at a certain stage in order to map long-range gene interactions of the stage. 3C assays can also be performed for a specific TF using chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (Fullwood et al., 2009; Li et al., 2014) and HiChIP (Mumbach et al., 2016). These methods combine ChIP with 3C to produce a direct view of long-range interactions associated with a TF of interest. To our knowledge, ChIA-PET or HiChIP has not yet been applied in floral TFs despite its high potential to correlate chromatin 3D structure with TF binding. For example, MADS-box TF homo- or heterotetramer complex has been shown to bind to two CArG boxes in short linear distance to form loops that are essential for target gene expression (Melzer et al., 2009; Mendes et al., 2013). However, it is not clear whether MADS-box TFs (or other oligomeric TFs) also enable long-range looping or even cause 3D chromatin structural rearrangement, such as breaking TAD boundaries as shown for Yamanaka factors during cell fate reprogramming (Stadhouders et al., 2018). These are potential mechanisms that could explain the functional diversity of MADS-box TFs and their potential pioneer activity in flower organ specification, respectively, for which ChIA-PET or HiChIP might provide valuable insight.

TF Cooperative Binding

Cooperative binding affects TF–DNA affinity and specificity. This is a widespread mechanism in eukaryotes for maximizing TF functional complexity by utilizing the minimum number of TFs. For example, Hox TFs in *Drosophila* bind highly similar sequences as monomers, whereas heterodimerization with the co-factor Extradenticle from the same TF family evokes significant differences in DNA binding affinity and specificity as revealed by SELEX-seq (Slattery et al., 2011). In plants, MADS-box TFs are prominent examples of cooperative binding. They form heterotetrametric complexes, so-called floral quartets, to regulate a distinct set of genes in the processes of flower formation and flower organ specification (Ruelens et al., 2017; Hugouvieux and Zubieta, 2018). It has been shown, both *in vivo* and *in vitro*, that different combinations of MADS-box TFs confer unique DNA binding specificity and affinity (Smaczniak et al., 2012, 2017; Muiño et al., 2014; Hugouvieux et al., 2018).

In some cases, a co-factor can be a non-DNA binding protein. For instance, two non-DNA binding co-factors in yeast, MET4 and MET28, enhance DNA binding specificity of TF Cbf1 through forming MET4-MET28-Cbf1 complex, which is required for activation of downstream genes (Siggers et al., 2011). In plants, the Evening Complex (EC), consisting of ARRHYTHMO (LUX), EARLY FLOWERING 3 (ELF3), and ELF4, is a key component of the circadian clock (Greenham and McClung, 2015; Huang and Nusinow, 2016). While only LUX is a TF, the *in vivo* functioning of the EC in the process of temperature and circadian clock-dependent flowering pathway requires non-DNA binding co-factors ELF3 and ELF4 (Nusinow et al., 2011). Furthermore, ChIP-seq data showed that G-box motifs are enriched adjacent to LUX binding sites, indicating additional co-factors that likely co-bind with EC to obtain further specificity and cooperativity for transcriptional regulation (Ezer et al., 2017). Similar mechanisms have also been proposed for PHYTOCHROME INTERACTING FACTOR 4 (PIF4), a key TF involved in thermoresponsive flowering in *Arabidopsis*. Its DNA binding activity can be sequestered by ELF3 (Nieto et al., 2015) or abrogated by DELLA proteins (Lucas et al., 2008), both through direct physical interactions. Taken together, it is crucial to take into account the presence or absence of TF co-factors for *in vivo* binding prediction.

INCORPORATING TFBS MODELS IN CURRENT AND FUTURE ANALYSES OF GENE REGULATION

The capacity to detect TFBS, both *in vitro* and *in vivo*, in increasingly reliable ways offers the opportunity to better answer various types of biological questions. For example, it is now possible to manipulate TFBS with genome editing, study the way how TFBSs are evolving, better predict gene regulation, and understand the DNA recruitment of chromatin regulators.

From DNA Binding to Gene Regulation and Regulatory Networks

Once TFBSs are identified or reliably predicted, the next challenge is to understand whether, how, and in which cellular context TF binding results in changes of target gene expression. Here, one can distinguish dedicated analyses of individual

binding events and potential target genes (“bottom up”) or make use of genome-wide expression data followed by mathematical modeling (“top down”).

A classical and powerful way to identify regulators of a given biological process or developmental transition consists of building lists of co-regulated genes and identifying *cis* elements overrepresented in their promoters. Once identified, these motifs can be compared with motifs in TFBS databases to identify TFs or TF families that are candidate regulators. Combined with detailed TF expression data, this represents a way to identify regulators. Several bioinformatics tools were developed based on this approach, such as Cistome (Austin et al., 2016), PlantRegMap (Jin et al., 2017), and TF2Network (Kulkarni et al., 2017). These tools take as an input a set of genes for which predicted regulators are searched. As a result, a set of potential regulators is identified and can be further validated using experimental approaches. In the example of TF2Network, using a standard dataset based on experimental TF binding data revealed that it recovers 92% of the true regulators using the long region promoter definition and the overall of 56% of the correct regulators when fed with a set of differentially expressed genes (Kulkarni et al., 2017). In a related approach, mathematical modeling using gene expression data can reveal gene network modules, and knowledge from known TF binding preferences can be used to validate predicted key gene regulatory interactions (Ichihashi et al., 2014).

TFBS models are now widely applied not only to characterize gene regulatory networks (GRN), but also to understand mechanisms underlying gene activation or repression. For example, TFBS prediction helped to identify TFs that mediate recruitment of repressive Polycomb protein complexes to specific genomic locations (Xiao et al., 2017; Zhou et al., 2018). A major challenge is still to identify and validate cell-type-specific gene regulatory interactions, which can now be addressed by combining cell-type selection by fluorescence activated cell/nuclei sorting or INTACT with ChIP-seq or other epigenomic technologies (see review by Wang et al., 2012). Another promising technology is single-cell approaches (see review by Grün and Van Oudenaarden, 2015), which at the moment are still under development for plant tissues (Brennecke et al., 2013).

Targeted TFBS Perturbation

Testing the functional impact of a predicted TFBS usually involves targeted mutagenesis in a transgenic context, e.g., using reporter assays, or in an endogenous context using CRISPR-Cas9-based systems. By using reporter genes (e.g., GFP or luciferase) under control of the target gene regulatory regions with modified TFBSs, it is possible to dissect spatiotemporal and quantitative changes in target gene expression depending on the presence or absence of a TFBS (Benn and Dehesh, 2016; Díaz-Triviño et al., 2017). For example, the tissue specificity of the *AP3* promoter was altered by replacing native TFBS with the ones of predicted specificity toward SEP3-AG or SEP3-AP1 floral homeotic protein complexes (Smaczniak et al., 2017). Combining results from TFBS prediction and reporter gene assays also helps reveal the mechanisms of TFBS recruitment in a native promoter context. For example, in *Drosophila* the combination of SELEX and reporter gene expression (lacZ and GFP) experiments revealed that clusters of low-affinity binding sites are maintained and required for the proper tissue-specific expression of the Hox genes, homeotic genes crucial for segment specification (Crocker et al., 2015). To which extent the clusters of TFBSs regulate gene expression in plants is yet to be determined through similar approaches.

With the advent of new technologies such as CRISPR-Cas9, mutations can now be introduced in endogenous genomic locations. For example, when a regulatory region of *AG* gene located in the second intron was deleted by the CRISPR-Cas9, mutant plants show partial homeotic transformations of stamens to petals, supporting an important role of this regulatory region (Yan et al., 2016). One of the challenges for the CRISPR-Cas9 strategy is that it requires TFBSs that contain a PAM motif 5′-NGG-3′ for efficient cleavage. Thus, the generation of new versions of Cas9 with different sequence requirement or the possibility to perform directed mutagenesis with a template DNA will open new avenues for precise TFBS perturbation. Moreover, the modifications of the CRISPR-Cas9 system by fusing cytidine deaminases to catalytically inactive Cas9 allow for the targeted, programmable single-nucleotide changes within a TFBS of interest (Yan et al., 2016). This is a paradigm shift, as genetics until now has mainly challenged regulatory networks by modifying their nodes, i.e., the protein-coding genes, and TFBS mutations will allow challenging the links without compromising all functions of a potentially pleiotropic TF. Recently, Barakat et al. (2018) reported an assay that combines ChIP and a massively parallel reporter assay (ChIP-STARR-seq) to identify functional TFBS in primed and naive human embryonic stem cells (Table 1). The resulting functional TFBSs of a given TF were further validated by CRISPR-Cas9 followed by a transient expression assay, proving the robustness of such method. This method is potentially applicable in plants, but with a limitation that maintaining and transfection of plant cells in culture (e.g., leaf protoplast) of a stage of interest is more challenging than that of the mammalian system.

Evolution of TFBS and Plant Diversity

Studying the conservation of *cis* elements containing TFBSs between different species or between promoters of closely related paralogous genes in a genome can shed light on the evolution of a GRN. How changes in *cis* elements relate to alterations in the expression pattern of a gene and subsequently lead to novel gene functions is not yet fully understood. At another level, understanding the evolutionary dynamics of gene regulatory interactions can provide deeper insights into how developmental programs evolve. The first step into this direction is to develop experimental approaches to study TFBSs in different species. In a genome-wide comparative ChIP-seq study, Muino et al. (2016) studied binding of SEP3 in two closely related *Arabidopsis* species with similar flower morphology. They found that TF binding conservation was associated with sequence conservation of CArG-box motifs and with the relative position of the TFBS to its potential target gene, and that loss/gain of binding sites tended to be associated with changes in gene expression. Their study revealed clear differences in SEP3-bound regions between the two species. A high level of binding divergence (13% overlap) was also reported for two orthologous MADS-box TFs, FLC in *Arabidopsis* and PERPETUAL FLOWERING1 in *Arabis alpine* (Mateos et al., 2017). Therefore,

comparative ChIP-seq studies can indicate conserved core target gene networks of developmental TFs in plants and distinguish plant lineage-specific functions and potentially less relevant binding sites. Interestingly, a similar observation was also reported in animals, as revealed by ChIP-seq on livers of five vertebrates (Schmidt et al., 2010). The authors observed a highly conserved TFBS motif for two TFs, but highly divergent binding events on conserved genes of different species.

Besides genome-wide approaches, targeted analysis of individual promoters or regulatory regions can elucidate regulatory divergence after speciation or gene duplication. For example, absence/presence of a single CArG-box in the promoter regions of the two MADS-box TF paralogs, *AP1* and *CAL*, determines spatiotemporal and quantitative differences in gene activity (Ye et al., 2016). Studying the TFBSs of homologs from different species can also reveal the evolution of their molecular function. For example, analyzing TF binding specificity of the LFY homologs from different plant species, land plants, mosses, and algae has revealed subtle changes in their preferred TFBS motifs, suggesting that LFY DNA binding specificity changed during land plant evolution (Sayou et al., 2014). Thus, the combination of numerous TFBS models with novel genome sequences could ultimately unlock mechanisms of GRN evolution.

FUTURE PERSPECTIVES

Technological advances such as NGS have revolutionized TFBS identification. It is now possible to identify TFBSs for hundreds of TFs for any organism in a short time with limited costs. However, accurate quantitative TFBS modeling is not a trivial process and seems to lag behind the pace of NGS dataset generation. Innovative analyses will be required to better extract valuable information hidden in datasets and compensate for the biases and drawbacks inherent to each particular method. As an example, by combining DAP-seq data for auxin response factor 5 (ARF5), a list of auxin-induced genes, and PWM model for dimeric ARF binding, we have uncovered a new ARF binding site configuration (inverted repeat 13) that seems to favor regulation and was not noticed before (Stigliani et al., 2018). This experience highlights that re-examination of the publicly available datasets could lead to novel findings, and that TFBS modeling requires careful planning and implementation. Therefore, an important future goal is full automation of TFBS modeling. In this regard, emerging artificial intelligence and machine learning are projected to make important contributions. Indeed, machine learning approaches have already played an important role in TFBS modeling and have been shown to be more powerful than conventional algorithms in several aspects, as discussed earlier. Finally, integrating datasets from chromatin environments, such as chromatin accessibility and 3D genome maps, is of great importance to better predict TF binding and the concomitant transcriptional events. A dedicated and accessible tool that could integrate such complex datasets in a combinatorial way is still lacking and will likely be an important focus of future investigations.

FUNDING

This work was supported by the Agence Nationale de la Recherche (project FloPiNet to C.Z., K.K., and F.P.), the Grenoble Alliance for Cell and Structural Biology (ANR-10-LABX-49-01 to F.P., C.Z., and A.S.), and Action Thématique et Incitative sur Programme (ATIP)-Avenir (to C.Z.).

AUTHOR CONTRIBUTIONS

F.P., C.Z., and X.L. planned the review outline and content. X.L. wrote the manuscript with input from all authors. All authors contributed to the reviewing and editing of the manuscript.

ACKNOWLEDGMENTS

No conflict of interest declared.

Received: June 29, 2018
Revised: September 20, 2018
Accepted: October 30, 2018
Published: November 14, 2018

REFERENCES

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. (2015). Deconvolving the recognition of DNA shape from sequence. Cell **161**:307–318.

Afek, A., Schipper, J.L., Horton, J., Gordân, R., and Lukatsky, D.B. (2014). Protein-DNA binding in the absence of specific base-pair recognition. Proc. Natl. Acad. Sci. U S A **111**:17140–17145.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**:831–838.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat. Rev. Mol. Cell Biol. **16**:155–166.

Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. **17**:487–500.

Allshire, R.C., and Madhani, H.D. (2018). Ten principles of heterochromatin formation and function. Nat. Rev. Mol. Cell Biol. **19**:229–244.

Aloy, P., and Russell, R.B. (2006). Structural systems biology: Modelling protein interactions. Nat. Rev. Mol. Cell Biol. **7**:188–197.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. **12**:878.

Austin, R.S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T.T., Fan, J., Foong, C., Breit, R., Desveaux, D., et al. (2016). New BAR tools for mining expression data and exploring Cis-elements in *Arabidopsis thaliana*. Plant J. **88**:490–504.

Azad, R.N., Zafiropoulos, D., Ober, D., Jiang, Y., Chiu, T.-P., Sagendorf, J.M., Rohs, R., and Tullius, T.D. (2018). Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. Nucleic Acids Res. **46**:2636–2647.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. **14**:283–291.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. Science **324**:1720–1723.

Baptista, T., Grünberg, S., Minoungou, N., Koster, M.J.E., Timmers, H.T.M., Hahn, S., Devys, D., and Tora, L. (2017). SAGA is a general cofactor for RNA polymerase II transcription. Mol. Cell **70**:1163–1164.

Barakat, T.S., Halbritter, F., Zhang, M., Rendeiro, A.F., Perenthaler, E., Bock, C., and Chambers, I. (2018). Functional dissection of the enhancer repertoire in human embryonic stem cells. Cell Stem Cell **23**:276–288.e8.

Bartlett, A., O'Malley, R.C., Huang, S.C., Galli, M., Nery, J.R., Gallavotti, A., and Ecker, J.R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat. Protoc. **12**:1659–1672.

Beltrao, P., Kiel, C., and Serrano, L. (2007). Structures in systems biology. Curr. Opin. Struct. Biol. **17**:378–384.

Benn, G., and Dehesh, K. (2016). Quantitative analysis of cis -regulatory element activity using synthetic promoters in transgenic plants. Methods Mol. Biol. **1482**:15–30.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. **4**:393–411.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. **24**:1429–1435.

Bezhani, S., Winter, C., Hershman, S., Wagner, J.D., Kennedy, J.F., Kwon, C.S., Pfluger, J., Su, Y., and Wagner, D. (2007). Unique, shared, and redundant roles for the *Arabidopsis* SWI/SNF chromatin remodeling ATPases BRAHMA and SPLAYED. Plant Cell **19**:403–416.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods **10**:1093–1098.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods **10**:1213–1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature **523**:486–490.

Bulyk, M.L. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. **30**:1255–1261.

Chen, Y., Zhang, X., Dantas Machado, A.C., Ding, Y., Chen, Z., Qin, P.Z., Rohs, R., and Chen, L. (2013). Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. Nucleic Acids Res. **41**:8368–8376.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods **14**:959–962.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer HOX specificity and regulatory robustness. Cell **160**:191–203.

Deal, R.B., and Henikoff, S. (2011). The INTACT method for cell type–specific gene expression and chromatin profiling in *Arabidopsis thaliana*. Nat. Protoc. **6**:56–68.

Dekker, J., and Misteli, T. (2015). Long-range chromatin interactions. Cold Spring Harb. Perspect. Biol. **7**:1–23.

Díaz-Triviño, S., Long, Y., Scheres, B., and Blilou, I. (2017). Analysis of a plant transcriptional regulatory network using transient expression systems. Methods Mol. Biol. **1629**:83–104.

Dong, P., Tu, X., Chu, P.Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D chromatin architecture of large plant genomes determined by local A/B compartments. Mol. Plant **10**:1497–1509.

Dong, Q., Li, N., Li, X., Yuan, Z., Xie, D., Wang, X., Li, J., Yu, Y., Wang, J., Ding, B., et al. (2018). Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. Plant J. **94**:1141–1156.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res. **25**:1268–1280.

Eagen, K.P. (2018). Principles of chromosome architecture revealed by Hi-C. Trends Biochem. Sci. **43**:469–478.

Elfving, N., Davoine, C., Benlloch, R., Blomberg, J., Brannstrom, K., Muller, D., Nilsson, A., Ulfstedt, M., Ronne, H., Wingsle, G., et al. (2011). The *Arabidopsis thaliana* Med25 mediator subunit integrates environmental cues to control plant development. Proc. Natl. Acad. Sci. U S A **108**:8245–8250.

Ezer, D., Jung, J.-H., Lan, H., Biswas, S., Gregoire, L., Box, M.S., Charoensawan, V., Cortijo, S., Lai, X., Stöckle, D., et al. (2017). The evening complex coordinates environmental and endogenous signals in *Arabidopsis*. Nat. Plants **3**:17087.

Farrel, A., and Guo, J. (2017). An efficient algorithm for improving structure-based prediction of transcription factor binding sites. BMC Bioinformatics **18**:342.

Farrel, A., Murphy, J., and Guo, J.T. (2016). Structure-based prediction of transcription factor binding specificity using an integrative energy function. Bioinformatics **32**:i306–i313.

Finnegan, E.J., Genger, R.K., Kovac, K., Peacock, W.J., and Dennis, E.S. (1998). DNA methylation and the promotion of flowering by vernalization. Proc. Natl. Acad. Sci. U S A **95**:5824–5829.

Fisher, W.W., Li, J.J., Hammonds, A.S., Brown, J.B., Pfeiffer, B.D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J.A., Eisen, M.B., et al. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proc. Natl. Acad. Sci. U S A **109**:21330–21335.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature **538**:265–269.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature **462**:58–64.

Fyodorov, D.V., Zhou, B.-R., Skoultchi, A.I., and Bai, Y. (2017). Emerging roles of linker histones in regulating chromatin structure and function. Nat. Rev. Mol. Cell Biol. **19**:192–206.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput. Biol. **10**:e1003711.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. **3**:1093–1104.

Greenham, K., and McClung, C.R. (2015). Integrating circadian dynamics with physiological processes in plants. Nat. Rev. Genet. **16**:598–610.

Greenside, P.G., Shimko, T., Fordyce, P., and Kundaje, A. (2018). Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. Bioinformatics **34**:i629–i637.

Grün, D., and Van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. Cell **163**:799–810.

Guo, Y., Tian, K., Zeng, H., Guo, X., and Gifford, D.K. (2018). A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. Genome Res. https://doi.org/10.1101/130815.

Gutin, J., Sadeh, R., Bodenheimer, N., Joseph-Strauss, D., Klein-Brill, A., Alajem, A., Ram, O., and Friedman, N. (2018). Fine-resolution

mapping of TF binding and chromatin interactions. Cell Rep. **22**:2601–2614.

Guturu, H., Doxey, A.C., Wenger, A.M., and Bejerano, G. (2013). Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. **368**:20130029.

Hancock, S.P., Ghane, T., Cascio, D., Rohs, R., Di Felice, R., and Johnson, R.C. (2013). Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. Nucleic Acids Res. **41**:6750–6760.

Harmston, N., and Lenhard, B. (2013). Chromatin and epigenetic features of long-range gene regulation. Nucleic Acids Res. **41**:7185–7199.

Hennig, L., and Derkacheva, M. (2009). Diversity of Polycomb group complexes in plants: same rules, different players? Trends Genet. **25**:414–423.

Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature **463**:474–484.

Hombach, D., Schwarz, J.M., Robinson, P.N., Schuelke, M., and Seelow, D. (2016). A systematic, large-scale comparison of transcription factor binding site models. BMC Genomics **17**:1–10.

Hou, X., Zhou, J., Liu, C., Liu, L., Shen, L., and Yu, H. (2014). Nuclear factor Y-mediated H3K27me3 demethylation of the SOC1 locus orchestrates flowering responses of *Arabidopsis*. Nat. Commun. **5**:1–14.

Hu, Z., Killion, P.J., and Iyer, V.R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet. **39**:683–687.

Hu, M., Yu, J., Taylor, J.M.G., Chinnaiyan, A.M., and Qin, Z.S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. Nucleic Acids Res. **38**:2154–2167.

Huang, H., and Nusinow, D.A. (2016). Into the evening: complex interactions in the *Arabidopsis* circadian clock. Trends Genet. **32**:674–686.

Hugouvieux, V., and Zubieta, C. (2018). MADS transcription factors cooperate: complexities of complex formation. J. Exp. Bot. **69**:1821–1823.

Hugouvieux, V., Silva, C.S., Jourdain, A., Stigliani, A., Charras, Q., Conn, V., Conn, S.J., Carles, C.C., Parcy, F., and Zubieta, C. (2018). Tetramerization of MADS family transcription factors SEPALLATA3 and AGAMOUS is required for floral meristem determinacy in *Arabidopsis*. Nucleic Acids Res. **46**:4966–4977.

Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., and Bulyk, M.L. (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. **43**:D117–D122.

Ichihashi, Y., Aguilar-Martinez, J.A., Farhi, M., Chitwood, D.H., Kumar, R., Millon, L.V., Peng, J., Maloof, J.N., and Sinha, N.R. (2014). Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. Proc. Natl. Acad. Sci. U S A **111**:E2616–E2621.

Ingouff, M., and Berger, F. (2010). Histone3 variants in plants. Chromosoma **119**:27–33.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P., and Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nat. Methods **14**:316–322.

Iwafuchi-Doi, M., and Zaret, K.S. (2016). Cell fate control by pioneer transcription factors. Development **143**:1833–1837.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. **45**:D1040–D1045.

Johnson, L., Mollah, S., Garcia, B.A., Muratore, T.L., Shabanowitz, J., Hunt, D.F., and Jacobsen, S.E. (2004). Mass spectrometry analysis of *Arabidopsis* histone H3 reveals distinct combinations of post-translational modifications. Nucleic Acids Res. **32**:6511–6518.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science **316**:1497–1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpa, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. **20**:861–873.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature **527**:384–388.

Käppel, S., Melzer, R., Rümpler, F., Gafert, C., and Theißen, G. (2018). The floral homeotic protein SEPALLATA3 recognizes target DNA sequences by shape readout involving a conserved arginine residue in the MADS-domain. Plant J. **95**:341–357.

Kasinathan, S., Orsi, G.A., Zentner, G.E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. Nat. Methods **11**:203–209.

Kaufmann, K., Muiño, J.M., Østerås, M., Farinelli, L., Krajewski, P., and Angenent, G.C. (2010). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat. Protoc. **5**:457–472.

Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. **26**:990–999.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., Van Der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. **46**:D260–D266.

Kiesel, A., Roth, C., Ge, W., Wess, M., Meier, M., and Söding, J. (2018). The BaMM web server for de-novo motif discovery and regulatory sequence analysis. Nucleic Acids Res. **46**:W215–W220.

Kim, M.Y., and Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. Trends Plant Sci. **19**:320–326.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013). The next-generation sequencing revolution and its impact on genomics. Cell **155**:27–38.

Kulakovskiy, I.V., Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B., and Makeev, V.J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. **41**:195–202.

Kulkarni, S.R., Vaneechoutte, D., Van de Velde, J., and Vandepoele, K. (2017). TF2Network: predicting transcription factor regulators and gene regulatory networks in *Arabidopsis* using publicly available binding site information. Nucleic Acids Res. **46**:e31.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell **172**:650–665.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A.C., Riley, T.R., Sandstrom, R., Sabo, P.J., Lu, Y., Rohs, R., Stamatoyannopoulos, J.A., et al. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. Proc. Natl. Acad. Sci. U S A **110**:6376–6381.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature **521**:436–444.

Lelli, K.M., Slattery, M., and Mann, R.S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. Annu. Rev. Genet. **46**:43–68.

Levo, M., Zalckvar, E., Sharon, E., Machado, A.C.D., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. Genome Res. **25**:1018–1029.

Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E.V., Kolchanov, N.A., and Ruan, Y. (2014). Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. BMC Genomics **15**:1–10.

Li, C., Gu, L., Gao, L., Chen, C., Wei, C.Q., Qiu, Q., Chien, C.W., Wang, S., Jiang, L., Ai, L.F., et al. (2016). Concerted genomic targeting of H3K27 demethylase REF6 and chromatin-remodeling ATPase BRM in *Arabidopsis*. Nat. Genet. **48**:687–693.

Li, J., Sagendorf, J.M., Chiu, T.P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res. **45**:12877–12887.

Lin, C.K., and Chen, C.Y. (2013). PiDNA: predicting protein-DNA interactions with structural models. Nucleic Acids Res. **41**:523–530.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell **133**:523–536.

Liu, Z., Mao, F., Guo, J., Yan, B., Wang, P., Qu, Y., and Xu, Y. (2005). Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. Nucleic Acids Res. **33**:546–558.

Liu, H., Shi, Y., Chen, X.S., and Warshel, A. (2009a). Simulating the electrostatic guidance of the vectorial translocations in hexameric helicases and translocases. Proc. Natl. Acad. Sci. U S A **106**:7449–7454.

Liu, C., Xi, W., Shen, L., Tan, C., and Yu, H. (2009b). Regulation of floral patterning by flowering time genes. Dev. Cell **16**:711–722.

Liu, X., Kim, Y.J., Muller, R., Yumul, R.E., Liu, C., Pan, Y., Cao, X., Goodrich, J., and Chen, X. (2011). AGAMOUS terminates floral stem cell maintenance in arabidopsis by directly repressing WUSCHEL through recruitment of polycomb group proteins. Plant Cell **23**:3654–3670.

Liu, C., Wang, C., Wang, G., Becker, C., Zaidem, M., and Weigel, D. (2016). Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. Genome Res. **26**:1057–1068.

Liu, C., Cheng, Y.J., Wang, J.W., and Weigel, D. (2017). Prominent topologically associated domains differentiate global chromatin packing in rice from *Arabidopsis*. Nat. Plants **3**:742–748.

Liu, X., Yang, Y., Hu, Y., Zhou, L., Li, Y., and Hou, X. (2018). Temporal-specific interaction of NF-YC and CURLY LEAF during the floral transition regulates flowering. Plant Physiol. **177**:105–114.

Lucas, M., Davière, J.-M., Rodríguez-Falcón, M., Pontin, M., Iglesias-Pedraz, J.M., Lorrain, S., Fankhauser, C., Blázquez, M.A., Titarenko, E., Prat, S., et al. (2008). A molecular framework for light and gibberellin control of cell elongation. Nature **451**:480–484.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell **161**:1012–1025.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: how alterations of chromatin domains result in disease. Trends Genet. **32**:225–237.

Marcovitz, A., and Levy, Y. (2011). Frustration in protein-DNA binding influences conformational switching and target search kinetics. Proc. Natl. Acad. Sci. U S A **108**:17957–17962.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature **544**:427–433.

Mateos, J.L., Tilmes, V., Madrigal, P., Severing, E., Richter, R., Rijkenberg, C.W.M., Krajewski, P., and Coupland, G. (2017). Divergence of regulatory networks governed by the orthologous transcription factors FLC and PEP1 in Brassicaceae species. Proc. Natl. Acad. Sci. U S A **114**:E11037–E11046.

Mathelier, A., and Wasserman, W.W. (2013). The next generation of transcription factor binding site prediction. PLoS Comput. Biol. **9**:e1003214.

Mathelier, A., Xin, B., Chiu, T.P., Yang, L., Rohs, R., and Wasserman, W.W. (2016). DNA shape features improve transcription factor binding site predictions in vivo. Cell Syst. **3**:278–286.e4.

Mathur, S., Vyas, S., Kapoor, S., and Tyagi, A.K. (2011). The mediator complex in plants: structure, phylogeny, and expression profiling of representative genes in a dicot (*Arabidopsis*) and a monocot (rice) during reproduction and abiotic stress. Plant Physiol. **157**:1609–1627.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. **34**:D108–D110.

Mayran, A., and Drouin, J. (2018). Pioneer transcription factors shape the epigenetic landscape. J. Biol. Chem. **293**:13795–13804.

McGinty, R.K., and Tan, S. (2015). Nucleosome structure and function. Chem. Rev. **115**:2255–2273.

Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. Science **324**:407–410.

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

Mendes, M.A., Guerra, R.F., Berns, M.C., Manzo, C., Masiero, S., Finzi, L., Kater, M.M., and Colombo, L. (2013). MADS domain transcription factors mediate short-range DNA looping that is essential for target gene expression in *Arabidopsis*. Plant Cell **25**:2560–2572.

Meyer, C.A., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet. **15**:709–721.

Minguet, E.G., Segard, S., Charavay, C., and Parcy, F. (2015). MORPHEUS, a webtool for transcription factor binding analysis using position weight matrices with dependency. PLoS One **10**:1–12.

Moraga, F., and Aquea, F. (2015). Composition of the SAGA complex in plants and its role in controlling gene expression in response to abiotic stresses. Front. Plant Sci. **6**:1–9.

Moyroud, E., Minguet, E.G., Ott, F., Yant, L., Posé, D., Monniaux, M., Blanchet, S., Bastien, O., Thévenon, E., Weigel, D., et al. (2011). Prediction of regulatory interactions from genome sequences using a biophysical model for the *Arabidopsis* LEAFY transcription factor. Plant Cell **23**:1293–1306.

Muiño, J.M., Smaczniak, C., Angenent, G.C., Kaufmann, K., and Van Dijk, A.D.J. (2014). Structural determinants of DNA recognition by plant MADS-domain transcription factors. Nucleic Acids Res. **42**:2138–2146.

Muino, J.M., De Bruijn, S., Pajoro, A., Geuten, K., Vingron, M., Angenent, G.C., and Kaufmann, K. (2016). Evolution of DNA-

binding sites of a floral master regulatory transcription factor. Mol. Biol. Evol. **33**:185–200.

Müller, F., Zaucker, A., and Tora, L. (2010). Developmental regulation of transcription initiation: more than just changing the actors. Curr. Opin. Genet. Dev. **20**:533–540.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods **13**:919–922.

Nieto, C., López-Salmerón, V., Davière, J.M., and Prat, S. (2015). ELF3-PIF4 interaction regulates plant growth independently of the evening complex. Curr. Biol. **25**:187–193.

Nusinow, D.A., Helfer, A., Hamilton, E.E., King, J.J., Imaizumi, T., Schultz, T.F., Farré, E.M., and Kay, S.A. (2011). The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. Nature **475**:398–402.

O'Malley, R.C., Huang, S., Shan, C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. Cell **166**:1598.

Omidi, S., Zavolan, M., Pachkov, M., Breda, J., Berger, S., and van Nimwegen, E. (2017). Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. PLoS Comput. Biol. **13**:1–22.

Pachkov, M., Balwierz, P.J., Arnold, P., Ozonov, E., and Van Nimwegen, E. (2013). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res. **41**:214–220.

Paillard, G., and Lavery, R. (2004). Analyzing protein-DNA recognition mechanisms. Structure **12**:113–122.

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., et al. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol. **15**:R41.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. **10**:669–680.

Patel, A., Yang, P., Tinkham, M., Pradhan, M., Sun, M.A., Wang, Y., Hoang, D., Wolf, G., Horton, J.R., Zhang, X., et al. (2018). DNA conformation induces adaptable binding by tandem zinc finger proteins. Cell **173**:221–233.e12.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell **137**:1194–1211.

Pruneda-Paz, J.L., Breton, G., Nagel, D.H., Kang, S.E., Bonaldi, K., Doherty, C.J., Ravelo, S., Galli, M., Ecker, J.R., and Kay, S.A. (2014). A genome-scale resource for the functional characterization of arabidopsis transcription factors. Cell Rep. **8**:622–632.

Qin, Q., and Feng, J. (2017). Imputation for transcription factor binding predictions based on deep learning. PLoS Comput. Biol. **13**:e1005403.

Rada-Iglesias, A., Grosveld, F.G., and Papantonis, A. (2018). Forces driving the three-dimensional folding of eukaryotic genomes. Mol. Syst. Biol. **14**:e8214.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin loss eliminates all loop domains. Cell **171**:305–320.e24.

Rao, S., Chiu, T.P., Kribelbauer, J.F., Mann, R.S., Bussemaker, H.J., and Rohs, R. (2018). Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding. Epigenetics Chromatin **11**:1–11.

Rastogi, C., Rube, H.T., Kribelbauer, J.F., Crocker, J., Loker, R.E., Martini, G.D., Laptenko, O., Freed-Pastor, W.A., Prives, C., Stern, D.L., et al. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. Proc. Natl. Acad. Sci. U S A **115**:E3692–E3701.

Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D.R., and Zhao, K. (2017). CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. Mol. Cell **67**:1049–1058.e6.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell **147**:1408–1419.

Riley, T.R., Lazarovici, A., Mann, R.S., and Bussemaker, H.J. (2015). Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using featureREDUCE. Elife **4**:1–14.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods **4**:651–657.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature **461**:1248–1253.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. **79**:233–269.

Ruan, S., and Stormo, G.D. (2017). Inherent limitations of probabilistic models for protein-DNA binding specificity. PLOS Comput. Biol. **13**:e1005638.

Ruan, S., Swamidass, S.J., and Stormo, G.D. (2017). BEESEM: estimation of binding energy models using HT-SELEX data. Bioinformatics **33**:2288–2295.

Ruelens, P., Zhang, Z., van Mourik, H., Maere, S., Kaufmann, K., and Geuten, K. (2017). The origin of floral organ identity quartets. Plant Cell **29**:229–242.

Santolini, M., Mora, T., and Hakim, V. (2014). A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites. PLoS One **9**:e99015.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature **489**:109–113.

Sayou, C., Monniaux, M., Nanao, M.H., Moyroud, E., Brockington, S.F., Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y., et al. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science **343**:645–648.

Sayou, C., Nanao, M.H., Jamin, M., Pose, D., Thevenon, E., Gregoire, L., Tichtinsky, G., Denay, G., Ott, F., Llobet, M.P., et al. (2016). A SAM oligomerization domain shapes the genomic binding landscape of the LEAFY transcription factor. Nat. Commun. **48**:829–834.

Schmidl, C., Rendeiro, A.F., Sheffield, N.C., and Bock, C. (2015). ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. Nat. Methods **12**:963–965.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science **328**:1036–1040.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18**:6097–6100.

Setty, M., and Leslie, C.S. (2015). SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. PLoS Comput. Biol. **11**:1–22.

Shaffer, P.L., Jivan, A., Dollins, D.E., Claessens, F., and Gewirth, D.T. (2004). Structural basis of androgen receptor binding to selective

androgen response elements. Proc. Natl. Acad. Sci. U S A **101**:4758–4763.

Sharon, E., Lubliner, S., and Segal, E. (2008). A feature-based approach to modeling protein-DNA interactions. PLoS Comput. Biol. **4**:e1000154.

Siddharthan, R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One **5**:e9722.

Siggers, T., Duyzend, M.H., Reddy, J., Khan, S., and Bulyk, M.L. (2011). Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Mol. Syst. Biol. **7**:1–14.

Sijacic, P., Bajic, M., McKinney, E.C., Meagher, R.B., and Deal, R.B. (2018). Changes in chromatin accessibility between *Arabidopsis* stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. Plant J. **94**:215–231.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife **6**:1–35.

Skene, J.P., and Henikoff, S. (2018). CUT&RUN: targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat. Protoc. **13**:1–28.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. Cell **147**:1270–1282.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. **39**:381–399.

Smaczniak, C., Immink, R.G.H., Muiño, J.M., Blanvillain, R., Busscher, M., Busscher-Lange, J., Dinh, Q.D.P., Liu, S., Westphal, A.H., Boeren, S., et al. (2012). Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. Proc. Natl. Acad. Sci. U S A **109**:1560–1565.

Smaczniak, C., Muiño, J.M., Chen, D., Angenent, G.C., and Kaufmann, K. (2017). Differences in DNA-binding specificity of floral homeotic protein complexes predict organ-specific target genes. Plant Cell **29**:1822–1835.

Soutourina, J. (2017). Transcription regulation by the Mediator complex. Nat. Rev. Mol. Cell Biol. **19**:262–274.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. **13**: 613–626.

Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., et al. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nat. Genet. **50**:238–249.

Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., and Cohen, B.A. (2018). A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. Cell Syst. **6**:444–455.e6.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. **24**:814–826.

Stigliani, A., Martin-Arevalillo, R., Lucas, J., Bessy, A., Vinos-Poyo, T., Mironova, V., Vernoux, T., Dumas, R., and Parcy, F. (2018). Capturing auxin response factors syntax using DNA binding models. Mol. Plant https://doi.org/10.1016/j.molp.2018.09.010.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics **16**:16–23.

Stormo, G.D. (2013). Modeling the specificity of protein-DNA interactions. Quant. Biol. **1**:115–130.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. **11**:751–760.

Stormo, G.D., Schneider, T.D., and Gold, L.M. (1982). Characterization of translational initiation sites in E. coli. Nucleic Acids Res. **10**:2971–2996.

Stroud, H., Otero, S., Desvoyes, B.B., Ramírez-Parra, E., Jacobsen, S.E., Gutierrez, C., Ramirez-Parra, E., Jacobsen, S.E., and Gutierrez, C. (2012). Genome-wide analysis of histone H3. 1 and H3. 3 variants in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U S A **109**:5370–5375.

Sun, B., Looi, L.-S., Guo, S., He, Z., Gan, E.-S., Huang, J., Xu, Y., Wee, W.-Y., and Ito, T. (2014). Timing mechanism dependent on cell division is invoked by polycomb eviction in plant stem cells. Science **343**:1248559.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., et al. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. **36**:1009–1014.

Takeda, T., Corona, R.I., and Guo, J.T. (2013). A knowledge-based orientation potential for transcription factor-DNA docking. Bioinformatics **29**:322–330.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y. (2017). Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. Nature **551**:124–128.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature **489**:75–82.

Tomovic, A., and Oakeley, E.J. (2007). Position dependencies in transcription factor binding sites. Bioinformatics **23**:933–941.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. **23**:137–144.

Trigg, S.A., Garza, R.M., MacWilliams, A., Nery, J.R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Huang, S.S.C., et al. (2017). CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. Nat. Methods **14**:819–825.

Turner, D., Kim, R.G., and Guo, J.T. (2012). TFinDit: transcription factor-DNA interaction data depository. BMC Bioinformatics **13**:220.

Vachon, G., Engelhorn, J., and Carles, C.C. (2018). Interactions between transcription factors and chromatin regulators in the control of flower development. J. Exp. Bot. **69**:2461–2471.

Vera, D.L., Madzima, T.F., Labonne, J.D., Alam, M.P., Hoffman, G.G., Girimurugan, S.B., Zhang, J., McGinnis, K.M., Dennis, J.H., and Bass, H.W. (2014). Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional DNA elements in maize. Plant Cell **26**:3883–3893.

Viner, C., Johnson, J., Walker, N., Shi, H., Sjöberg, M., Adams, D.J., Ferguson-Smith, A.C., Bailey, T.L., and Hoffman, M.M. (2016). Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. bioRxiv https://doi.org/10.1101/043794.

Voss, T.C., and Hager, G.L. (2013). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat. Rev. Genet. **15**:69–81.

Wang, D., Mills, E.S., and Deal, R.B. (2012). Technologies for systems-level analysis of specific cell types in plants. Plant Sci. **197**:21–29.

Wang, H., Liu, C., Cheng, J., Liu, J., Zhang, L., He, C., Shen, W.H., Jin, H., Xu, L., and Zhang, Y. (2016). *Arabidopsis* flower and embryo developmental genes are repressed in seedlings by different

combinations of polycomb group proteins in association with distinct sets of cis-regulatory elements. PLoS Genet. **12**:1–25.

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., Ye, Z., Shen, C., Li, J., Zhang, L., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. Nat. Genet. **49**:579–587.

Wang, M., Tai, C., E, W., and Wei, L. (2018a). DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic Acids Res. **46**:e69.

Wang, Z., Civelek, M., Miller, C., Sheffield, N., Guertin, M.J., and Zang, C. (2018b). BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. Bioinformatics **34**:2867–2869.

Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., and Zhang, X. (2018c). Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. Nat. Plants **4**:90–97.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. **5**:276–287.

Weber, B., Zicola, J., Oka, R., and Stam, M. (2016). Plant enhancers: a call for discovery. Trends Plant Sci. **21**:974–987.

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannet, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2018). YY1 is a structural regulator of enhancer-promoter loops. Cell **171**:1573–1588.e28.

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., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158**:1431–1443.

White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. U S A **110**:11952–11957.

Whiteld, T.W., Wang, J., Collins, P.J., Partridge, E.C., Aldred, S.F., Trinklein, N.D., Myers, R.M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. Genome Biol. **13**:R50.

Wu, H., and Zhang, Y. (2014). Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell **156**:45–68.

Wu, M., Sang, Y., Bezhani, S., Yamaguchi, N., Han, S., Li, Z., Su, Y., Slewinski, T.L., and Wagner, D. (2012). SWI2/SNF2 chromatin remodeling ATPases overcome polycomb repression and control floral organ identity with the LEAFY and SEPALLATA3 transcription factors. Proc. Natl. Acad. Sci. U S A **109**:3576–3581.

Xiao, J., and Wagner, D. (2015). Polycomb repression in the regulation of growth and development in *Arabidopsis*. Curr. Opin. Plant Biol. **23**:15–24.

Xiao, J., Jin, R., Yu, X., Shen, M., Wagner, J.D., Pai, A., Song, C., Zhuang, M., Klasfeld, S., He, C., et al. (2017). Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in *Arabidopsis*. Nat. Genet. **49**:1546–1552.

Xu, B., Schones, D.E., Wang, Y., Liang, H., and Li, G. (2013). A structural-based strategy for recognition of transcription factor binding sites. PLoS One **8**:e52460.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., et al. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell **154**:801–813.

Yan, W., Chen, D., and Kaufmann, K. (2016). Efficient multiplex mutagenesis by RNA-guided Cas9 and its use in the characterization of regulatory elements in the AGAMOUS gene. Plant Methods **12**:1–9.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W.W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. **42**:148–155.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. Mol. Syst. Biol. **13**:910.

Yazaki, J., Galli, M., Kim, A.Y., Nito, K., Aleman, F., Chang, K.N., Carvunis, A.-R., Quan, R., Nguyen, H., Song, L., et al. (2016). Mapping transcription factor interactome networks using HaloTag protein arrays. Proc. Natl. Acad. Sci. U S A **113**:E4238–E4247.

Ye, L., Wang, B., Zhang, W.-G., Shan, H., and Kong, H. (2016). Gain of an auto-regulatory site led to divergence of the arabidopsis APETALA1 and CAULIFLOWER duplicate genes in the time, space and level of expression and regulation of one paralog by the other. Plant Physiol. **171**:1055–1069.

Yin, Y., Sieradzan, A.K., Liwo, A., He, Y., and Scheraga, H.A. (2015). Physics-based potentials for coarse-grained modeling of protein-DNA interactions. J. Chem. Theor. Comput. **11**:1792–1808.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science **356**:eaaj2239.

Zaret, K.S. (2018). Pioneering the chromatin landscape. Nat. Genet. **50**:167–169.

Zaret, K.S., and Mango, S.E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev. **37**:76–81.

Zemach, A., and Grafi, G. (2007). Methyl-CpG-binding domain proteins in plants: interpreters of DNA methylation. Trends Plant Sci. **12**:80–85.

Zentner, G.E., Kasinathan, S., Xin, B., Rohs, R., and Henikoff, S. (2015). ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. Nat. Commun. **6**:8733.

Zhang, C., Liu, S., Zhu, Q., and Zhou, Y. (2005). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J. Med. Chem. **48**:2325–2335.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., et al. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell **126**:1189–1201.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. Plant Cell **24**:2719–2731.

Zhang, L., Martini, G.D., Tomas Rube, H., Kribelbauer, J.F., Rastogi, C., FitzPatrick, V.D., Houtman, J.C., Bussemaker, H.J., and Pufall, M.A. (2018a). SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. Genome Res. **28**:111–121.

Zhang, H., Lang, Z., and Zhu, J. (2018b). Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. **19**:489–506.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput. Biol. **5**:e1000590.

Zhao, Y., Ruan, S., Pandey, M., and Stormo, G.D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. Genetics **191**:781–790.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods **12**:931–934.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. U S A **112**:4654–4659.

Zhou, Y., Wang, Y., Krause, K., Yang, T., Dongus, J.A., Zhang, Y., and Turck, F. (2018). Telobox motifs recruit CLF/SWN-PRC2 for H3K27me3 deposition via TRB factors in *Arabidopsis*. Nat. Genet. **50**:638–644.

Zhu, H., Wang, G., and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. Nat. Rev. Genet. **17**:551–565.

Zuo, Z., Roy, B., Chang, Y.K., Granas, D., and Stormo, G.D. (2017). Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. Sci. Adv. **3**:eaao1799.

V.2 JASPAR

D260–D266 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 13 November 2017*
doi: 10.1093/nar/gkx1126

JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Aziz Khan[1,†], Oriol Fornes[2,†], Arnaud Stigliani[3,†], Marius Gheorghe[1], Jaime A. Castro-Mondragon[1], Robin van der Lee[2], Adrien Bessy[3], Jeanne Chèneby[4,5], Shubhada R. Kulkarni[6,7,8], Ge Tan[9,10], Damir Baranasic[9,10], David J. Arenillas[2], Albin Sandelin[11,*], Klaas Vandepoele[6,7,8], Boris Lenhard[9,10,12,*], Benoît Ballester[4,5], Wyeth W. Wasserman[2,*], François Parcy[3] and Anthony Mathelier[1,13,*]

[1]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [2]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 28th Ave W, Vancouver, BC V5Z 4H4, Canada, [3]University of Grenoble Alpes, CNRS, CEA, INRA, BIG-LPCV, 38000 Grenoble, France, [4]INSERM, UMR1090 TAGC, Marseille, F-13288, France, [5]Aix-Marseille Université, UMR1090 TAGC, Marseille, F-13288, France, [6]Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium, [7]VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium, [8]Bioinformatics Institute Ghent, Ghent University, Technologiepark 927, 9052 Ghent, Belgium, [9]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK, [10]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W12 0NN, UK, [11]The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, DK2200 Copenhagen N, Denmark, [12]Sars International Centre for Marine Molecular Biology, University of Bergen, N-5008 Bergen, Norway and [13]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 25, 2017; Revised October 17, 2017; Editorial Decision October 18, 2017; Accepted October 27, 2017

ABSTRACT

JASPAR (http://jaspar.genereg.net) is an open-access database of curated, non-redundant transcription factor (TF)-binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups. In the 2018 release of JASPAR, the CORE collection has been expanded with 322 new PFMs (60 for vertebrates and 262 for plants) and 33 PFMs were updated (24 for vertebrates, 8 for plants and 1 for insects). These new profiles represent a 30% expansion compared to the 2016 release. In addition, we have introduced 316 TFFMs (95 for vertebrates, 218 for plants and 3 for insects). This release incorporates clusters of similar PFMs in each taxon and each TF class per taxon. The JASPAR 2018 CORE vertebrate collection of PFMs was used to predict TF-binding sites in the human genome. The predictions are made available to the scientific community through a UCSC Genome Browser track data hub. Finally, this update comes with a new web framework with an interactive and responsive user-interface, along with new features. All the underlying data can be retrieved programmatically using a RESTful API and through the JASPAR 2018 R/Bioconductor package.

INTRODUCTION

Transcription factors (TFs) are sequence-specific DNA-binding proteins involved in the transcriptional regulation of gene expression (1). TFs bind to DNA through their DNA-binding domain(s) (DBDs), which are used for TF classification (2). DNA regions at which TFs bind are defined as TF-binding sites (TFBSs) and can be identified

*To whom correspondence should be addressed. Tel: +47 228 40 561; Email: anthony.mathelier@ncmm.uio.no
Correspondence may also be addressed to Albin Sandelin. Tel: +45 2245 6668; Fax: +45 3532 2128; Email: albin@binf.ku.dk
Correspondence may also be addressed to Boris Lenhard. Tel: +44 20 8383 8353; Email: b.lenhard@imperial.ac.uk
Correspondence may also be addressed to Wyeth W. Wasserman. Tel: +1 604 875 3812; Fax: +1 604 875 3840; Email: wyeth@cmmt.ubc.ca
†These authors contributed equally to the paper as first authors.

Table 1. Overview of the growth of the number of PFMs in the JASPAR 2018 CORE collection compared to the JASPAR 2016 CORE collection

Taxonomic group	Non-redundant PFMs in JASPAR 2016	New non-redundant PFMs in JASPAR 2018	Updated PFMs in JASPAR 2018	Total PFMs (non-redundant) in JASPAR 2018	Total PFMs (all versions) in JASPAR 2018
Vertebrates	519	60	24	579	719
Plants	227	262	8	489	501
Insects	133	0	1	133	140
Nematodes	26	0	0	26	26
Fungi	176	0	0	176	177
Urochordata	1	0	0	1	1
Total	**1082**	**322**	**33**	**1404**	**1564**

in vivo by methods such as chromatin immunoprecipitation (ChIP) or *in vitro* by methods based on binding of large pools of DNA fragments (e.g. Systematic evolution of ligands by exponential enrichment (SELEX) or protein-binding microarrays (PBM)) (reviewed in (3)). Analysis of TFBSs for a given TF provides models for its specific DNA-binding preferences, which in turn can be used to predict TFBSs in DNA sequences (4). This is important as experiments can only identify TFBSs that are bound in the cell and state analyzed.

The computational representation of TF binding preferences has evolved over the years, from simple consensus sequences to position frequency matrices (PFMs). A PFM summarizes experimentally determined DNA sequences bound by an individual TF by counting the number of occurrences of each nucleotide at each position within aligned TFBSs. Such matrices can be converted into position weight matrices (PWMs), also known as position-specific scoring matrices, which are probabilistic models that can be used to predict TFBSs in DNA sequences (reviewed in (5)).

PFMs/PWMs have been the standard models for describing binding preferences of TFs for many years. The JASPAR database is among the most popular and longest maintained databases for PFMs and a standard resource in the field. In particular, the JASPAR CORE collection of the database, which is the most used, stores non-redundant TF binding profiles, providing a single representative DNA binding model per TF decided by expert curators. Exceptionally, multiple TF-binding profiles are associated to a TF when it is known to interact with DNA with multiple distinct sequence preferences, due to differential splicing for example (6,7). JASPAR was created and persists under three guiding principles: (i) unrestricted open-access; (ii) manual curation and non-redundancy of profiles; and (iii) ease-of-use. The 2016 release of the JASPAR CORE collection stored 1082 non-redundant and manually curated TF-binding profiles as PFMs for TFs from six different taxonomic groups (vertebrates, plants, insects, nematodes, fungi and urochordata) (8).

An intrinsic limitation to PFMs/PWMs is that they ignore inter-nucleotide dependencies within TFBSs (9–13). TF–DNA interaction data derived from next-generation sequencing assays has improved the computational modeling of TF binding (14–19). For example, the TF flexible models (TFFMs) (14), based on first-order hidden Markov models, capture dinucleotide dependencies within TFBSs and were introduced in the 2016 release of the JASPAR database.

In this report, we describe the seventh release of JASPAR (8,20–24), which comes with a major expansion and update of the CORE collection of TF-binding profiles as PFMs and TFFMs. These models have been manually assessed by expert curators who reconciled recent high-throughput data with available literature and linked the models to the classification of their TF DBDs from TFClass (2). The CORE collection expansion is supported by a range of new functionalities and resources, including PFM clustering, genome-wide UCSC tracks of predicted TFBSs and fully redesigned user and programming interfaces.

EXPANSION AND UPDATE OF THE JASPAR CORE COLLECTION

In this 2018 release of the JASPAR database, we added 355 new PFMs for TFs from plants (270), vertebrates (84) and insects (1) to the JASPAR CORE collection (Table 1). Specifically, we added 322 PFMs (262 for plants, a 118% increase and 60 for vertebrates, an 11% increase) for TF monomers and dimers that were not previously present in JASPAR and updated 33 (8 in plants, 3% of JASPAR 2016, 24 in vertebrates, 5% of JASPAR 2016 and 1 in insects). The PFMs were manually curated using independent external literature supporting the candidate TF-binding preferences, as previously described in (23). The curated PFMs were derived from ChIP-seq (from ReMap (25) and (26–30)), DAP-seq (31), SMiLE-seq (32), PBM (33) and HT-SELEX (34) experiments. The JASPAR CORE collection now includes 1404 non-redundant PFMs (579 for vertebrates, 489 for plants, 176 for fungi, 133 for insects, 26 for nematodes and 1 for urochordata) (Table 1).

We continued with the incorporation of TFFM models, initiated in JASPAR 2016. In this release of JASPAR, we introduced 316 new TFFMs for vertebrates (95), plants (218) and *Drosophila* (3), which represents a 243% increase in the number of non-redundant TFFMs stored in the JASPAR CORE collection.

HIERARCHICAL CLUSTERING OF TF-BINDING PROFILES

While the non-redundancy of binding profiles is one of the guiding principles of JASPAR, TFs with similar DBDs often have similar binding preferences (35,36). To facilitate the exploration of similar profiles in the JASPAR CORE collection, we performed hierarchical clustering of PFMs using the RSAT matrix-clustering tool (37). Specifically, the tool was applied to PFMs in each taxon independently as

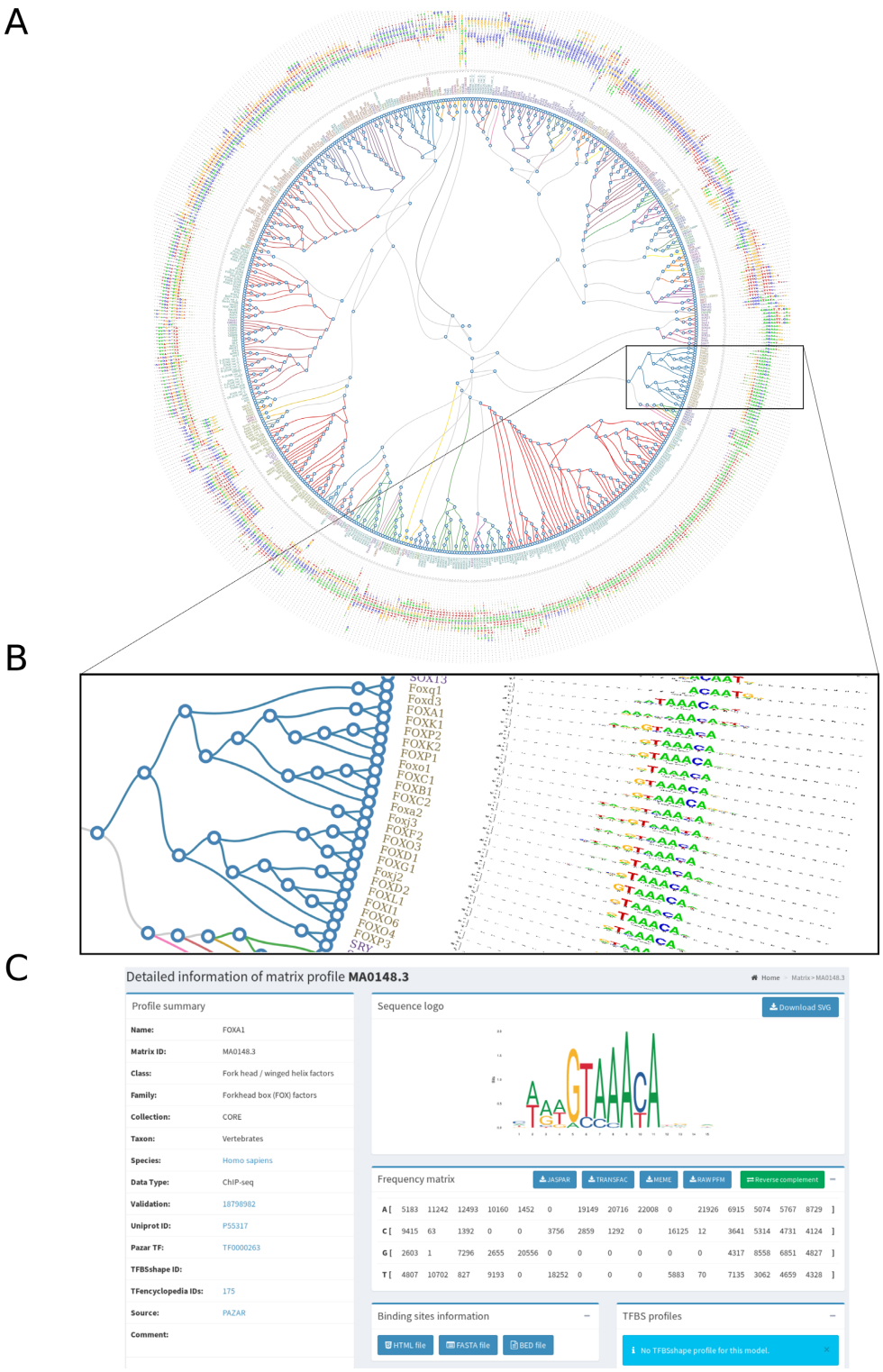


Figure 1. JASPAR PFM clustering. (**A**) Radial tree representing the clusterization of the JASPAR CORE vertebrate PFMs. (**B**) Zoom in view of the radial tree where the predicted clusters are highlighted at the branches and the TF classes are indicated with different colors at the leaves. (**C**) Clicking on a leaf in the radial tree will open a link to the corresponding motif description page on the JASPAR website (the MA0148.3 profile associated to FOXA1 is provided here as an example).

Figure 2. Overview of the JASPAR 2018 new web interface with interactive searching activity. (**A**) A quick and detailed search feature on the homepage. (**B**) A responsive table lists the searched profile(s), which can be further selected and added to the cart listed on the right panel for users to perform their own analyses. (**C**) A detailed page for the GATA3 matrix profile, which is divided into sub-panels including the profile summary, sequence logo, PFM, TF-binding information, external links, version information, ChIP-seq centrality, TFFM and other details. (**D**) The PFM for the GATA3 profile (MA0037.2) is downloaded in MEME format using the RESTful API.

well as in each TF class per taxon. The clustering results are provided as radial trees (Figure 1), which can further be explored through dedicated web pages (http://jaspar.genereg.net/matrix-clusters).

JASPAR UCSC TRACKS FOR GENOME-WIDE ANALYSES OF TFBSs

A typical application of JASPAR TF-binding profiles in gene regulation studies is the identification of TFBSs in DNA sequences for further analyses. Although, we recognize that genome-wide PWM-based predictions contain a high number false positives, we believe that they are a powerful resource for the research community in the context of a variety of genomic information, including transcription start site activity, DNA accessibility, histone marks, evolutionary conservation or *in vivo* TF binding (38–46). To facilitate such integrative analyses, we have performed TFBS predictions on the human genome using the JASPAR CORE vertebrate PFMs (see Supplementary Data for details on the computation). The predicted TFBSs are publicly available through a UCSC Genome Browser data hub (47) containing tracks for the human genome assemblies hg19 and hg38 (http://jaspar.genereg.net/genome-tracks/).

A NEW, POWERFUL AND USER-FRIENDLY WEB INTERFACE

A new web interface

The JASPAR 2018 release comes with a completely redesigned web interface that meets modern web standards. This interactive web framework is implemented using Django, a model-view-controller based web-framework for Python. We used MySQL as a backend database to store profile metadata and Bootstrap as a frontend template engine. We have greatly improved the visibility and usability of existing functionality, created easier navigation with semantic URLs, and enhanced browsing and searching. On the homepage, we provide a dynamic tour of JASPAR 2018, walking users through the main features of the new website. A video of the tour is available at http://jaspar.genereg.net/tour. The database can be browsed for individual collections by using the navigation links on the left sidebar. Moreover, it can be searched for each of the six different taxonomic groups included in the JASPAR CORE collection using the tabs available on the homepage (Figure 2). TF-binding profiles can be further filtered through the case insensitive search option available on the homepage. In addition, through the 'Advanced Options', the search criteria can be further restricted (Figure 2A). Search results are presented in a responsive and paginated table along with sequence logos of the PFMs, which can be selected for download or to perform a variety of analyses available on the right panel (Figure 2B). All information in the tables can be downloaded as comma-separated value files. Profile IDs and sequence logos can be clicked to view the detailed profile pages (Figure 2C). PFMs can be downloaded in several formats including JASPAR, TRANSFAC and MEME (Figure 2D). Furthermore, we have incorporated new features to the web interface, such as 'Add to Cart', where users can add TF profiles of interest for download or further analyses (Figure 2B). Finally, we have introduced semantic URLs to facilitate external linking to the detailed pages of individual profiles (e.g. http://jaspar.genereg.net/matrix/MA0059.1/). We have implemented a URL redirection mechanism to correctly direct the links pointing to previous JASPAR URL patterns from external resources.

RESTful API

In previous releases, the underlying data could be retrieved as flat files or by using programming language-specific modules. Associated with this release, we introduced a RESTful API to access the JASPAR database programmatically (see https://www.biorxiv.org/content/early/2017/07/06/160184 for details). The RESTful API enables programmatic access to JASPAR by most programming languages and returns data in seven widely used formats: JSON, JSONP, JASPAR, MEME, PFM, TRANSFAC and YAML. Further, it provides a browsable interface and access to the JASPAR motif inference tool for bioinformatics tool developers. The RESTful API is implemented in Python using the Django REST Framework and is freely accessible at http://jaspar.genereg.net/api/. The source code for the website and RESTful API are freely available at https://bitbucket.org/CBGR/jaspar under GPL v3 license.

CONCLUSION AND PERSPECTIVES

In this seventh release of the JASPAR database, we continue our commitment to provide the research community with high-quality, non-redundant TF-binding profiles for TFs in six taxa. As in previous releases, we have greatly expanded the number of available profiles in the database, both for PFMs and TFFMs. We also greatly improved user experience through a new easy-to-use website and a RESTful API that grants universal programmatic access to the database. Moreover, for the PFMs in the JASPAR CORE collection, we provide a hierarchical clustering and genome-wide TFBS predictions for the hg19 and hg38 human genome assemblies as UCSC tracks.

During the curation process, hundreds of PFMs were discarded because our curators failed to find any support from existing literature. As new experiments and data become available, binding preferences for these TFs will be considered for JASPAR incorporation. For instance, we re-examined data from (34) to incorporate seven previously excluded PFMs into JASPAR 2018. In the future, we would like to engage the scientific community in the curation process to increase our capacity to introduce new TF-binding profiles in JASPAR. We plan to dedicate a specific section of the website to hosting the profiles that were not introduced into JASPAR, to encourage researchers to perform experiments and/or point us to literature that our curators missed in order to support these profiles. We believe that the engagement of the scientific community to support JASPAR will further improve our capacity to expand the collection of high quality TF-binding profiles.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the scientific community for performing experimental assays of TF–DNA interactions and for publicly releasing the data. We thank Georgios Magklaras and his team for IT support. We thank José Manuel Franco for sharing the plant PBM data, Jens De Ceukeleire for help with plant ChIP-seq data processing and José Luis Villanueva-Cañas for sharing the *Drosophila* TFFMs prior to publication. We thank Rachelle Farkas for proofreading the manuscript.

FUNDING

Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., M.G., A.K.); Genome Canada and Canadian Institutes of Health Research (On-Target Grants) [255ONT and BOP-149430 to W.W.W., O.F., R.v.d.L., D.J.A.]; Natural Sciences and Engineering Research Council of Canada (Discovery Grant) [RGPIN-2017–06824 to W.W.W.]; Weston Brain Institute [20R74681 to O.F.]; Agence Nationale de la Recherche [ANR-10-LABX-49–01 to F.P., A.S.]; IDEX graduate schoool (to A.S.); CNRS (to A.B., F.P.); Research Foundation–Flanders Grant [G001015N to S.R.K.]; French Ministry

of Higher Education and Research (MESR) PhD Fellowship (to J.A.C.-M.); Lundbeck Foundation (to A.S.); Independent Research Fund Denmark (to A.S.); Innovation Fund Denmark (to A.S.); Elixir Denmark (to A.S.); Wellcome Trust [106954 to G.T., D.B., B.L.]; Biotechnology and Biological Sciences Research Council [BB/N023358/1 to G.T., D.B., B.L.]; Medical Research Council UK [MC_UP_1102/1 to G.T., D.B., B.L.]. The open access publication charge for this paper has been waived by Oxford University Press - *NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.
Conflict of interest statement. None declared.

REFERENCES

1. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
2. Wingender,E., Schoeps,T., Haubrock,M. and Dönitz,J. (2015) TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.*, **43**, D97–D102.
3. Xie,Z., Hu,S., Qian,J., Blackshaw,S. and Zhu,H. (2011) Systematic characterization of protein-DNA interactions. *Cell. Mol. Life Sci.*, **68**, 1657–1668.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
5. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant. Biol.*, **1**, 115–130.
6. Stormo,G.D. (2015) DNA motif databases and their uses. *Curr. Protoc. Bioinformatics*, **51**, 1–6.
7. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
8. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
9. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
10. Bulyk,M.L., Johnson,P.L.F. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
11. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
12. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
13. Chin,F. and Leung,H.C.M. (2008) DNA motif representation with nucleotide dependency. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 110–119.
14. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
15. Zellers,R.G., Drewell,R.A. and Dresch,J.M. (2015) MARZ: an algorithm to combinatorially analyze gapped n-mer models of transcription factor binding. *BMC Bioinformatics*, **16**, 1–14.
16. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 1–15.
17. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
18. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.
19. Omidi,S., Zavolan,M., Pachkov,M., Breda,J., Berger,S. and van Nimwegen,E. (2017) Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. *PLoS Comput. Biol.*, **13**, e1005176.
20. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
21. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
22. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
23. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D1010.
24. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D1427.
25. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2017) ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seqexperiments. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1092.
26. Eveland,A.L., Goldshmidt,A., Pautler,M., Morohashi,K., Liseron-Monfils,C., Lewis,M.W., Kumari,S., Hiraga,S., Yang,F., Unger-Wallace,E. *et al.* (2014) Regulatory modules controlling maize inflorescence architecture. *Genome Res.*, **24**, 431–443.
27. Verkest,A., Abeel,T., Heyndrickx,K.S., Van Leene,J., Lanz,C., Van De Slijke,E., De Winne,N., Eeckhout,D., Persiau,G., Van Breusegem,F. *et al.* (2014) A generic tool for transcription factor target gene discovery in Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiol.*, **164**, 1122–1133.
28. Li,C., Qiao,Z., Qi,W., Wang,Q., Yuan,Y., Yang,X., Tang,Y., Mei,B., Lv,Y., Zhao,H. *et al.* (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell*, **27**, 532–545.
29. Cui,X., Lu,F., Qiu,Q., Zhou,B., Gu,L., Zhang,S., Kang,Y., Cui,X., Ma,X., Yao,Q. *et al.* (2016) REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in Arabidopsis. *Nat. Genet.*, **48**, 694–699.
30. Birkenbihl,R.P., Kracher,B. and Somssich,I.E. (2017) Induced genome-wide binding of three Arabidopsis WRKY transcription factors during early MAMP-triggered immunity. *Plant Cell*, **29**, 20–38.
31. O'Malley,R.C., Huang,S.-S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.
32. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.
33. Franco-Zorrilla,J.M., López-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
34. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
35. Weirauch,M.T., Yang,A., Albu,M., Cote,A., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

36. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
37. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.
38. Kwon,A.T., Arenillas,D.J., Worsley Hunt,R. and Wasserman,W.W. (2012) oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3*, **2**, 987–1002.
39. Mathelier,A., Lefebvre,C., Zhang,A.W., Arenillas,D.J., Ding,J., Wasserman,W.W. and Shah,S.P. (2015) Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.*, **16**, 1–17.
40. Verfaillie,A., Imrichova,H., Janky,R. and Aerts,S. (2015) iRegulon and i-cistarget: reconstructing regulatory networks using motif and track enrichment. *Curr. Protoc. Bioinformatics*, **52**, 1–39.
41. Arenillas,D.J., Forrest,A.R.R., Kawaji,H., Lassmann,T. and FANTOM ConsortiumFANTOM Consortium, Wasserman,W.W. and Mathelier,A. (2016) CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. *Bioinformatics*, **32**, 2858–2860.
42. Shi,W., Fornes,O., Mathelier,A. and Wasserman,W.W. (2016) Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.*, **44**, 10106–10116.
43. Arner,E., Daub,C.O., Vitting-Seerup,K., Andersson,R., Lilje,B., Drabløs,F., Lennartsson,A., Rönnerblad,M., Hrydziuszko,O., Vitezic,M. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.
44. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
45. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
46. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
47. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

Résumé

Chez les angiospermes, la floraison est un processus qui prend part en plusieurs étapes. Le méristème caulinaire, un réservoir de cellule souche d'où émergent la totalité des organes aériens de la plante, va d'abord se différencier en méristème d'inflorescence. Des méristèmes floraux vont alors émerger des flancs du méristème d'inflorescence pour donner naissance aux différents organes qui composent la fleur : les pétales, les sépales, les étamines et le carpelle. Chacune de ces phases est régulée avec finesse par des facteurs de transcription, une famille de protéines se liant à l'ADN pour induire l'activation ou la répression des gènes. Si cette thèse nous a permis de contribuer à la mise à jour de JASPAR, une base de données qui recense des profils liaison de facteurs de transcription, elle a avant tout pour but d'apporter un regard nouveau sur la compréhension des phénomènes qui contrôlent le développement des fleurs à travers l'étude d'une poignée de facteurs de transcription clé dans ce processus. Nous essayerons au mieux d'expliquer les paramètres qui influent sur la liaison de ces facteurs de transcription en utilisant des modèles bioinformatiques associés à des expériences de génomique.

Nous nous pencherons d'abord sur les facteurs de réponse à l'auxine à travers l'étude de deux représentants de cette famille de 23 protéines : ARF2 et ARF5. Si les facteurs de transcription de cette famille sont connus pour se lier en dimères, nous avons montré que ARF2 et ARF5 préféraient des espacements différents entre les sites de liaison monomériques sur l'ADN. Nous avons également montré que certaines configurations semblent favoriser l'activation des gènes liés.

Ensuite, nous avons étudié LFY, un facteur de transcription maître du développement floral. Nous avons amélioré un modèle de liaison existant et nous avons pu voir que l'intégration de données génomiques de natures diverse permettait de mieux comprendre la liaison du facteur de transcription in vivo.

Enfin, nous avons analysé les préférences des facteurs de transcription à boîte MADS, connus pour lier les mêmes séquences d'ADN et dont le rôle est de déterminer l'identité des organes floraux. À travers l'étude du complexe SEP3/AG, qui contrôle la formation du carpelle, nous avons montré que le domaine de tétramérisation de ces facteurs confère une spécificité de liaison expliquant potentiellement que des groupes de facteurs de transcription à boîte MADS régulent la formation d'organes floraux différents en activant des gènes distincts.

Abstract

In angiosperms, the development of flowers takes place in several stages. The meristem, a stem cell reservoir from which all the plant's aerial organs emerge, first differentiate into an inflorescence meristem. Floral meristems then emerge from the flanks of the inflorescence meristem to give birth to the different organs that compose the flower : the petals, sepals, stamens and carpel. Each of these phases is finely regulated by transcription factors, a family of proteins that bind to DNA to induce gene activation or repression. If this thesis allowed us to contribute to the JASPAR database, which gather transcription factor binding profiles, its main goal is to provide a new perspective on the understanding of the phenomena that control flower development through the study of a handful of key transcription factors in the regulation of floral development. We have tried to explain the parameters that influence the binding of these transcription factors using bioinformatics models associated with genomics experiments.

We have analysed the auxin response factors (ARF) through the study of two representatives of this family of 23 proteins : ARF2 and ARF5. The transcription factors of this family are known to bind in dimers and we have shown that ARF2 and ARF5 prefer different spacings between monomeric binding sites on DNA. We have also shown that some configurations seem to favour the activation of bound genes.

Then, we have studied LFY, a master transcription factor of floral development. We have improved an existing binding model and have seen that the integration of genomic data of various kinds provides a better understanding of the binding of the transcription factor in vivo.

Finally, we have analyzed the preferences of MADS box transcription factors, known to bind the same DNA sequences and whose role is to determine the identity of floral organs. Through the study of the SEP3/AG complex, which controls the formation of the carpel, we have found that the tetramerization domain of these factors confers binding specificity, potentially explaining that groups of MADS box transcription factors regulate the formation of different floral organs by activating distinct genes.