Caractérisation des séquences cis-régulatrices dans les régions proximales des gènes chez les plantes

Julien Roziere

université PARIS-SACLAY

Caractérisation des séquences *cis*-régulatrices dans les régions proximales des gènes chez les plantes

Characterization of gene-proximal cis-regulatory sequences in plants

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577 : structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Biologie moléculaire et cellulaire
Graduate School : Sciences de la vie et santé, Référent : Université d'Évry Val d'Essonne

Thèse préparée dans les unités de recherche Institut des Sciences des Plantes Paris-Saclay (Université Paris-Saclay, CNRS, INRAE, Univ Evry) et Institut Jean-Pierre Bourgin (Université Paris-Saclay, INRAE, AgroParisTech), sous la direction de Marie-Laure MARTIN, directrice de recherche INRAE, et la co-direction de Sylvie COURSOL, chargée de recherche INRAE.

Thèse soutenue à Paris-Saclay, le 14 décembre 2022, par

Julien ROZIÈRE

Composition du jury

Claudine Landès Professeur, Université d'Angers (IRHS)	Présidente
Laurent Laplaze Directeur de recherche, IRD (DIADE)	Rapporteur
Morgane Thomas-Chollier Maître de conférences (HDR), ENS PSL (IBENS)	Rapporteur
Maud Fagny Chargée de recherche, INRAE (GQE-Le Moulon)	Examinatrice
Christophe Sallaud Coordinateur de recherche, Limagrain Europe	Examinateur
Sophie Schbath Directrice de recherche, INRAE (MaIAGE)	Examinatrice

THESE DE DOCTORAT
NNT : 2022UPASL087

Remerciements

Ça y est ! C'est une aventure qui s'achève pour moi. Bien que cette thèse ait débuté il y a trois ans, en réalité cette aventure en a bientôt cinq. Depuis mon arrivée en tant que stagiaire en 2018, j'ai eu l'occasion de rencontrer et côtoyer énormément de personnes avec qui j'ai pu tisser des liens. Je vais tâcher, en quelques lignes, de remercier l'ensemble des personnes qui m'auront permis d'avancer sur les travaux que je présente dans ce manuscrit mais également toutes les personnes qui se seront montrées présentes pour m'aider à souffler, me soutenir et me remotiver quand il le fallait.

Marie-Laure et Sylvie, merci infiniment pour votre temps, votre patience, vos encouragements et votre soutien constant. Malgré les difficultés et les confinements qui se sont succédés, vous avez toujours répondu présentes. Je mesure la chance d'avoir été encadré par votre duo aussi complémentaire ! Lorsque je regarde trois ans en arrière, je vois à quel point j'ai pu grandir et en apprendre autant sur les sciences que sur la vie. De piou à piou-piou-piou en passant par piou-piou, je suis officiellement un petit poulet (initialement j'avais écrit "nuggets" mais c'est légèrement glauque).

Je souhaite remercier l'ensemble des membres des deux équipes dont j'ai fait partie au cours de ces dernières années.

À l'ensemble des membres de Qualibiosec, malgré le fait que je ne sois venu que peu de fois à Versailles, je tiens à vous remercier car ce fut toujours un réel plaisir de pouvoir échanger dans la bonne humeur avec vous et d'avoir eu la chance de découvrir la culture au champ le temps d'une après-midi.

À l'ensemble des Ginettes, je vous remercie pour ces presque cinq années passées à vos côtés. Merci pour les rires, les nombreuses discussions et anecdotes, les énigmes qui auront occupé une bonne partie de mon temps (merci Jean-Philippe !) et toutes les pauses café où, finalement, vous aurez réussi à m'en faire boire.

J'ai bien évidemment une pensée toute particulière pour Véro qui, avec Sylvie, m'aura encadré lors de mon arrivée chez les Ginettes. Merci Véro de m'avoir initié à la propagande PLM et de m'avoir motivé à reprendre le volley !

Merci également Kévin, mon n+1, et Perrine, ma jumelle de thèse, pour tous ces fous rires que l'on a pu avoir ("Faut tirer !" ou encore "voychiech") et le soutien que vous avez pu m'apporter durant ces trois années. Au plaisir de se refaire un truc tous ensemble.

Merci Arnaud, Margot, Simon et Yacine pour les moments que nous avons passé aussi bien en formation, en colloque ou au labo ! Merci Arnaud pour les encouragements hebdomadaires sur ce dernier mois de rédaction et bon courage à vous tous pour cette dernière année. C'est bientôt votre tour haha.

Merci Franck pour toutes ces réunions PLMview qui m'auront bien fait marrer avec votre duo du Muppet Show que vous formiez si bien avec Véro !

Je tiens à remercier Olivier Lespinet, David Pot et Hadi Quesneville qui ont composé mon comité de thèse durant ces trois années. Merci pour les nombreux conseils et discussions scientifiques précieuses qui m'auront permis de faire des choix décisifs lorsque cela a été nécessaire.

Je remercie Laurent Laplaze et Morgane Thomas-Chollier d'avoir accepté d'être les rapporteurs de ce travail ainsi que Maud Fagny, Claudine Landès, Christophe Sallaud et Sophie Schbath pour avoir accepté de prendre place dans mon jury.

Je remercie toute l'équipe de la MISS que j'ai eu l'occasion d'intégrer au cours de ma thèse en tant qu'animateur ! Merci de m'avoir laissé cette chance. C'est une expérience que je n'oublierai jamais. Merci Valérie, Elisabeth et Sylvie de faire de la MISS une véritable bouffée d'air frais pour toutes les personnes qui s'y rendent, élèves comme animateurs. Merci justement à tous les animateurs que j'ai eu la chance de côtoyer et avec qui j'ai passé de super moments. Je vous souhaite le meilleur pour la suite.

Merci à toute l'équipe de volley-ball de l'ADAS de Jouy ! Merci pour ces lundis midi de détente et de bons moments à jouer tous ensemble.

Je veux remercier toute ma famille, ma belle-famille et mes amis. En particulier, merci mon Dadidou, ma Lulu et ma Camille. Merci de m'avoir toujours soutenu. Sans vous je n'en serai pas là. Un grand merci Papi d'avoir répondu présent pour assister à ma soutenance et d'avoir été aux commandes de l'organisation du pot de thèse avec papa. Ce moment fut parfait grâce à vous. De même, un grand merci Tata, Tonton et Nico d'être venus m'encourager pour cette dernière épreuve ! Merci Juju et Toto pour les moments que nous avons pu passer ensemble à la maison ces dernières années et d'avoir pris le temps de m'écouter quand je vous racontais mes nombreuses galères. Merci à tous les membres de la TFB ! Bien que nous nous voyons moins souvent ces derniers temps, je vous remercie pour ces sept ans d'amitié et tous ces moments inoubliables que nous avons passé. Merci encore mon Beetle pour avoir pris le temps de m'aider avec la mise en page de ce manuscrit !

Enfin, Magalie, je tiens à te remercier pour tout ton soutien que tu m'apportes au quotidien et d'autant plus depuis ces derniers mois ! Merci d'être à mes côtés dans les bons comme les mauvais moments et de toujours me remotiver quand j'en ai besoin. Merci d'avoir pris le temps de relire ce manuscrit (dont tu ne comprenais rien) juste pour m'aider. Cette fin d'année 2022 marque un nouveau départ et je suis heureux de poursuivre cette aventure à tes côtés (et aux côtés de Kumo et Tsuki aussi, on va pas les oublier quand même).

Table des matières

Table des figures

Liste des tableaux

1 - Introduction

Cette thèse participe à une meilleure compréhension des mécanismes de régulation de l'expression des gènes chez les plantes. Après des généralités sur la vie fixée des plantes et ses contraintes, j'aborderai, dans cette introduction, les mécanismes transcriptionnels qui régulent l'expression des gènes chez les plantes et les acteurs impliqués. J'exposerai ensuite les différentes approches expérimentales et *in silico* mises en place pour caractériser les régions *cis*-régulatrices des génomes de plantes. Enfin, je présenterai les objectifs de mon travail de thèse.

1.1 . Les plantes : des organismes sessiles soumis aux contraintes environnementales

Les plantes présentent la particularité d'être des organismes fixés au sol par leurs racines qui les approvisionnent en eau et en éléments minéraux, leurs feuilles captant l'énergie solaire pour assimiler le carbone inorganique de l'air. Ceci les oblige à devoir s'adapter aux conditions contrastées et fluctuantes de leur environnement. Elles doivent ainsi faire face à des variations de facteurs abiotiques, comme des écarts de température, de luminosité ou d'apport en eau (Hill and Li, 2022; Shelake *et al.*, 2022). Des carences importantes en nutriments minéraux, tels que l'azote, peuvent aussi exister dans les sols, ou à l'inverse des toxicités néfastes dues à l'excès de sel ou de métaux toxiques peuvent survenir. Les plantes sont également soumises à de nombreuses agressions biotiques produites par d'autres êtres vivants, comme des animaux, des insectes, des plantes parasites, des champignons, des bactéries ou des virus. Au cours du milliard d'années nous séparant de l'apparition des premières algues, les plantes ont donc évolué en combinant des caractères fixés dans leur génome et la plasticité de l'expression de ce dernier (Calatayud *et al.*, 2013; Mitra *et al.*, 2021).

La perception des stress par les plantes et les réponses biologiques qui en découlent ont fait l'objet d'intenses recherches ces vingt dernières années. Les plantes sont ainsi capables de percevoir les signaux de l'environnement et de les transmettre aux cellules (Choudhary and Muthamilarasan, 2022). Les mécanismes mis en jeu vont conduire à une re-programmation de l'expression génétique et finalement à la réponse phénotypique de la plante au stress. De manière très schématique, une cascade d'évènements se déroule dans la plante qui va conduire à l'activation d'un certain nombre de facteurs de transcription (TF) (Singh, 2002). Ces TF se fixent alors sur l'ADN pour moduler l'expression des gènes répondant aux stress, ce qui induit des modifications morphologiques, physiologiques et biochimiques adaptées (Figure 1.1). Ainsi, la transcription des gènes constitue un processus clé dans la réponse adaptative des plantes aux stress qu'elles subissent (Hancock *et al.*, 2011; Waters *et al.*, 2017; Alonge *et al.*, 2020; Azodi *et al.*, 2020; Zhou *et al.*, 2022).

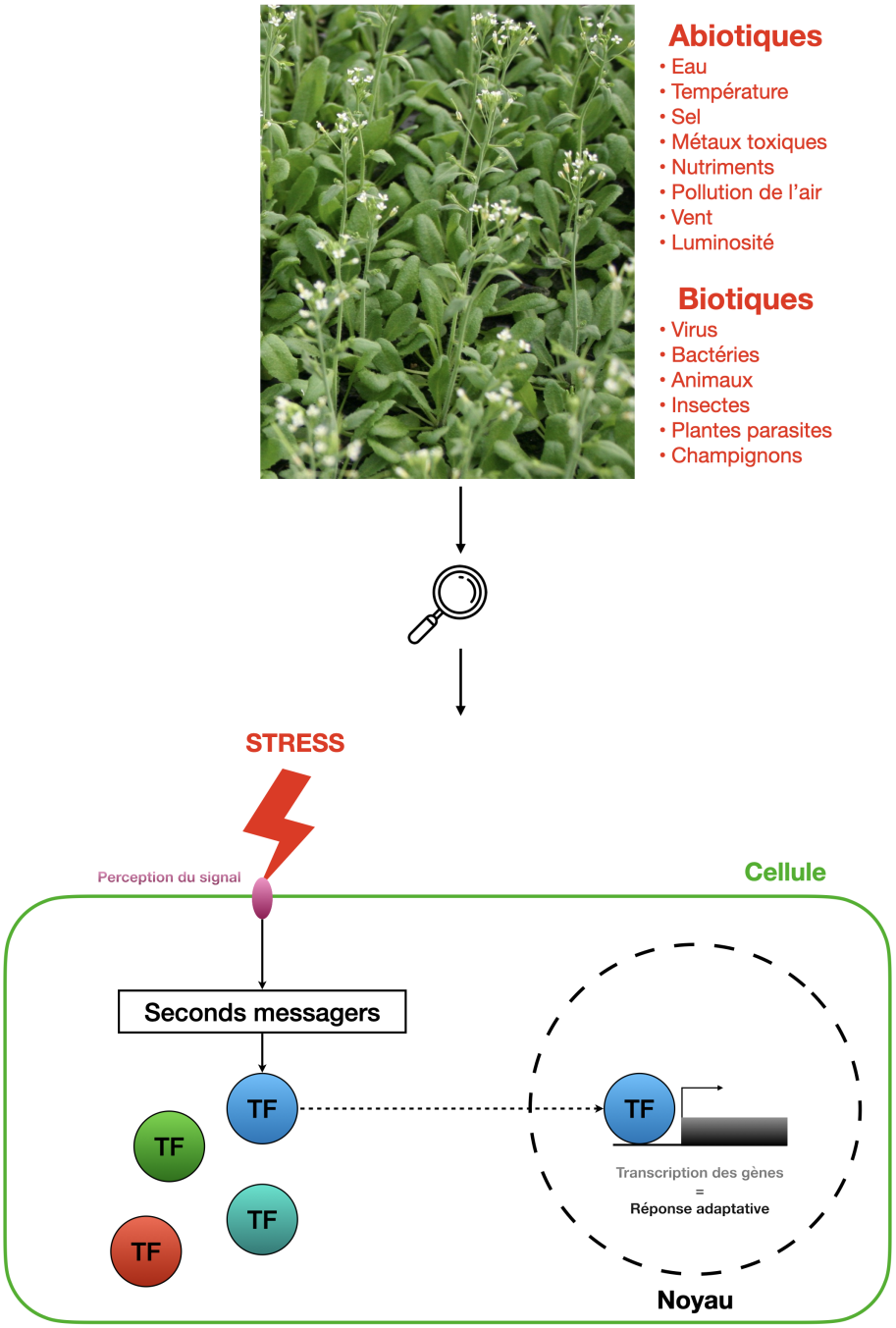


Figure 1.1 – Représentation schématique de la perception des stress par les plantes et de la plasticité de l'expression de leur génome.

1.2 . La régulation de la transcription chez les plantes

1.2.1 . La transcription : première étape dans l'expression d'un gène

L'acide désoxyribonucléique, plus communément appelé ADN, est une macromolécule composée d'un enchaînement ordonné de quatre molécules que l'on nomme des nucléotides. Ces derniers sont l'adénine (A), la cytosine (C), la thymine (T) et la guanine (G). Cet enchaînement précis de nucléotides constitue le support de l'information génétique, autrement dit l'information héréditaire d'un individu. Les molécules d'ADN se compose de deux brins complémentaires (A associé à T et C associé à G) prenant une forme de double hélice. Chez les eucaryotes, l'ADN double brins prend une forme complexe en se compactant autour de protéines appelées histones pour former des octamères connus sous le nom de nucléosomes. Enfin, à une plus grande échelle, ces structures se compactent également pour former des chromosomes qui tous ensemble désignent le génome (Figure 1.2).

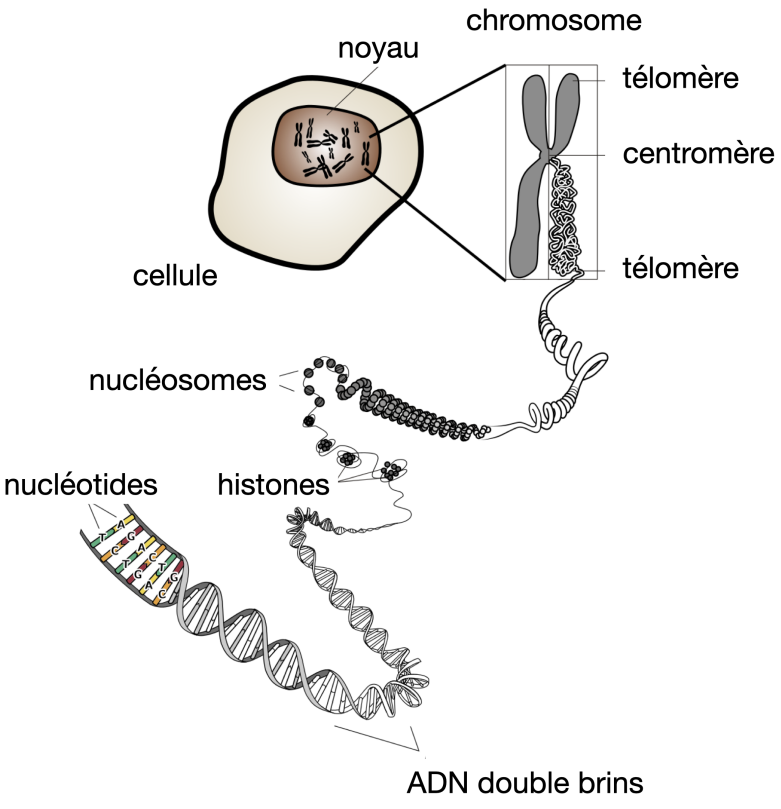


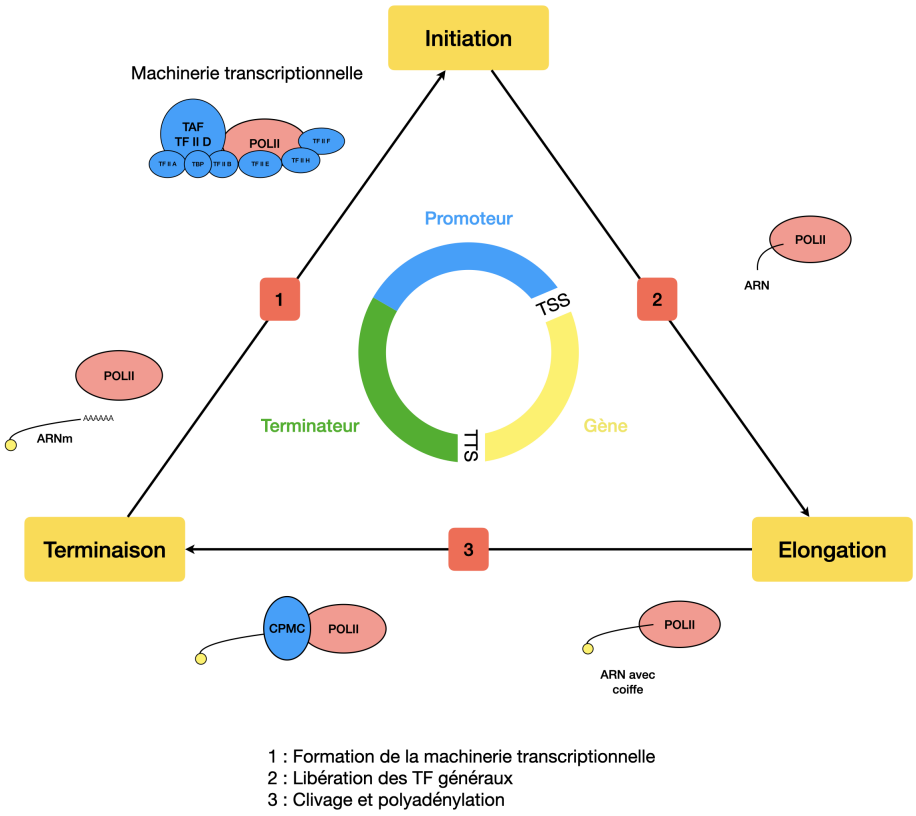
Figure 1.2 – Structure de l'ADN chez les eucaryotes.

Dans cette structure, les gènes font partie intégrantes de l'ADN et se définissent généralement par l'ensemble des portions de l'ADN qui "s'exprime". L'expression d'un gène réfère à l'ensemble des processus mis en place par la cellule pour aboutir à un produit fonctionnel du gène donné. Le premier maillon de cette chaîne de processus est ce que l'on nomme la transcription.

1.2.2 . La machinerie transcriptionnelle

La transcription permet la synthèse d'une molécule d'ARN simple brin en utilisant pour référence la séquence d'ADN du gène. Contrairement à ce qui se passe chez les procaryotes, la transcription des eucaryotes est assurée par plusieurs ARN polymérases, chacune d'entre elles assurant la transcription de différentes classes de gènes (Haag and Pikaard, 2011). Ainsi, l'ARN polymérase I (POLI) assure la transcription des ARN ribosomiques dans le nucléole, alors que l'ARN polymérase II (POLII) assure celle des ARN messagers et l'ARN polymérase III (POLIII) celle des ARN de transfert et autres petits ARN. De plus, deux ARN polymérases supplémentaires, les ARN polymérases IV et V, sont spécifiques des plantes terrestres. Elles sont impliquées dans la synthèse de petits ARN interférents qui jouent un rôle dans la régulation de la transcription de certains gènes (Herr *et al.*, 2005; Landick, 2009).

Contrairement à l'ARN polymérase procaryote, la POLII ne fonctionne pas seule chez les eucaryotes. L'activation de la transcription dépend en effet d'un complexe protéique formé par la POLII et de nombreux co-facteurs qui se recrutent les uns les autres et correspondent aux facteurs généraux de la transcription ("General Transcription Factor" ou GTF) nommés TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH. La fixation du complexe de transcription sur l'ADN provoque l'ouverture et le déroulement des deux brins de l'ADN, tout en indiquant le brin qui sera transcrit. C'est l'initiation de la transcription. La POLII est ensuite libérée du reste des GTF et commence la synthèse de la molécule d'ARN à partir du site d'initiation de la transcription (TSS). Cette étape de synthèse, nommée l'élongation, se poursuit jusqu'à la détection du site de terminaison de la transcription (TTS). La détection de ce site entraîne la troisième et dernière étape de la transcription, nommée terminaison, qui consiste en la libération de la POLII et du transcrit synthétisé (Figure 1.3).



1 : Formation de la machinerie transcriptionnelle
2 : Libération des TF généraux
3 : Clivage et polyadénylation

Figure 1.3 – Cycle de l'ARN polymérase II dans la transcription chez les eucaryotes. *La transcription des ARN messagers comprend trois étapes : (1) l'initiation dans le promoteur, (2) l'élongation le long du gène et (3) la terminaison dans le terminateur. La première étape correspond à la formation de la machinerie transcriptionnelle et à sa fixation au niveau du site d'initiation de la transcription (TSS). La deuxième étape comprend la libération de l'ARN polymérase II (POLII) des facteurs généraux de la transcription, la synthèse de l'ARN messager (ARNm) basé sur le brin sélectionné et l'ajout d'une coiffe à l'ARN. La troisième étape correspond à la détection du site de terminaison de la transcription (TTS), au clivage et à la polyadénylation de l'ARN par le complexe moléculaire de clivage et de polyadénylation (CPMC). TF : Facteurs de transcription.*

Ces différentes étapes peuvent s'enchaîner plusieurs fois pour permettre l'expression rapide d'un gène cible et la synthèse de nombreux transcrits. En effet, le phénomène de "gene looping" qui est induit par l'interaction du facteur TFIIB avec les facteurs responsables de la spécificité/stimulation du clivage et de la polyadénylation, permet une interaction entre la région promotrice (décrite en 1.2.4)

et la région terminatrice (décrite en 1.2.5), et ainsi au gène de former une boucle (Wang *et al.*, 2010). Grâce à cette structure, la POLII peut réitérer la transcription du gène cible.

Pour que la machinerie transcriptionnelle puisse transcrire la séquence d'un gène en ARN, il est nécessaire que l'environnement au niveau du début du gène soit propice : (1) l'ADN doit être accessible et (2) l'ensemble des acteurs moléculaires capables de recruter la machinerie transcriptionnelle doivent être présents.

1.2.3 . Accessibilité de l'ADN

L'expression d'un gène ne peut se faire que lorsque celui-ci est accessible pour la machinerie transcriptionnelle. Pour cela, l'ADN doit être libre, c'est-à-dire non compacté autour d'histones. De ce fait, la structure locale de la chromatine est le premier facteur qui détermine si un gène peut être transcrit. Deux états chromatiniens sont à distinguer : (1) l'hétérochromatine dans lequel l'ADN est condensé et inaccessible, ce qui ne permet pas la transcription des gènes s'y trouvant et (2) l'euchromatine dans lequel l'ADN est décondensé et accessible, permettant ainsi l'expression des gènes. Le passage d'un état de la chromatine à l'autre est influencé par différents niveaux de modifications :

- le relâchement d'histones et la libération de l'ADN peuvent être médiés par des protéines de remodelage de la chromatine. La structure des nucléosomes est ainsi impactée et les régions redeviennent accessibles.
- Des modifications moléculaires des queues N-terminales des histones avec le transfert de groupements acétyles sont assurées par des histones acétyltransférases (HAT). Ces modifications permettent une décondensation de l'ADN et induisent, dès lors, la fixation de protéines impliquées dans l'expression. De la même manière, les histones peuvent subir des méthylations de différents résidus par l'action d'histones méthyltransférases. La nature de l'acide aminé méthylé et le nombre de groupements méthyles ajoutés impactent l'ouverture chromatinienne. A titre d'exemple, la triméthylation de l'histone H3 au niveau de la lysine 27 (H3K27me3) est une marque épigénétique associée à l'hétérochromatine et donc à l'inhibition de l'expression des gènes. Inversement, la triméthylation des histones H3 au niveau des lysines 4 (H3K4me3) est associée à la décondensation de l'ADN et à l'activation de la transcription génique (Schmitz *et al.*, 2021).
- La méthylation de cytosines des dinucléotides CG par l'action de méthyltransférases induit l'inaccessibilité de l'ADN (Schmitz *et al.*, 2021).
- Une restructuration de la chromatine médiée par les facteurs de transcription (TF). Ces facteurs sont des protéines capables d'induire ou de réprimer l'expression des gènes cibles en se fixant à l'ADN. Quatre modèles sont à ce jour proposés pour décrire leur impact sur l'accessibilité chromatinienne (Klemm *et al.*, 2019) :

  1. Le premier modèle mentionne une compétition passive des TF et des

histones pour l'ADN. En effet, lors du remodelage de la chromatine, des histones sont détachées de l'ADN et les TF proches prennent la place de celles-ci. Ce remplacement permet de maintenir l'accessibilité locale de la chromatine pour d'autres TF ou la machinerie transcriptionnelle (Figure 1.4a).

2. Le deuxième modèle suppose que les TF se fixent sur de l'ADN internucléosomal en remplaçant des protéines architecturales qui s'y trouvent de manière transitoire. L'instabilité ainsi créée induit l'activité des protéines de remodelage qui évincent les histones proches et rendent ainsi la chromatine accessible (Figure 1.4b).
3. Le troisième modèle implique une activité en trans des TF qui se fixent sur des sites distaux. Ces TF interviendraient grâce à des repliements chromatiniens sur les histones cibles en permettant leur relâchement et la libération de l'ADN pour des TF secondaires (Figure 1.4c).
4. Le quatrième et dernier modèle mentionne la présence de TF pionniers capables de se fixer à de l'ADN nucléosomal et de le libérer, seul ou conjointement avec des protéines de remodelage, ce qui donne l'accès à des TF secondaires pour initier la transcription (Figure 1.4d).

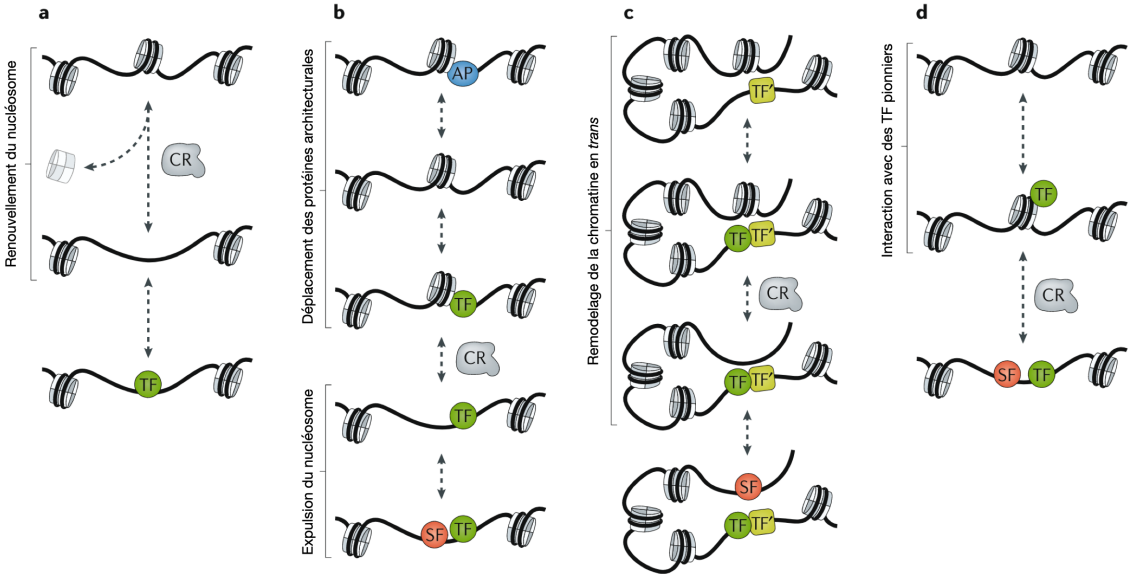


Figure 1.4 – Modèles proposés pour l'implication des facteurs de transcription dans l'accessibilité de la chromatine (adaptée de (Klemm *et al.*, 2019)).
(a) Compétition passive entre protéine de remodelage (CR) et facteur de transcription (TF). (b) Remplacement des protéines architecturales (AP) par des TF induisant le remodelage de la chromatine. (c) Activité distale des TF pour évincer des histones. (d) TF pionnier se fixant à de l'ADN nucléosomal. SF : Facteur de transcription dit secondaire.

En résumé, l'état de condensation de l'ADN autour des histones est un facteur clé dans la régulation de la transcription. La chromatine peut être modulée par l'activité de protéines de remodelage, la présence de marques épigénétiques sur les histones et l'ADN. Seules les régions accessibles de l'ADN peuvent être transcrites. De leur côté, les TF sont indispensables dans la régulation de la transcription. Ces facteurs n'interviennent pas seulement sur la condensation et l'accessibilité de l'ADN, mais peuvent également moduler le recrutement et le fonctionnement de la machinerie transcriptionnelle.

1.2.4 . Fixation de la machinerie transcriptionnelle

Lorsque l'ADN est accessible, des TF supplémentaires peuvent moduler la fixation de la machinerie transcriptionnelle et ainsi réguler la transcription.

Encadré : Focus sur les facteurs de transcription et leurs sites de fixation

Les TF sont caractérisés par la présence d'un domaine de liaison à l'ADN et d'un domaine effecteur permettant des interactions protéine-protéine. La majorité des TF ne possède qu'un seul type de domaine de liaison à l'ADN qui peut être présent en plusieurs copies au sein de la séquence (Charoensawan *et al.*, 2010). Il existe une grande diversité de domaines de liaison à l'ADN et, de ce fait, de TF. Chez les plantes, les TF sont ainsi organisés en 58 familles (Jin *et al.*, 2017), dont quatre leurs sont spécifiques. Il s'agit des familles APETALA 2/Ethylene-Responsive element binding Factor (AP2/ERF), WRKY, NAC et TCP. De manière intéressante, ces familles de TF sont décrites pour intervenir dans la réponse des plantes aux stress biotiques et abiotiques et/ou dans le développement des organes végétatifs (Hong, 2016). Elles sont donc directement associées aux caractéristiques propres aux plantes.

Les domaines de liaison à l'ADN reconnaissent et se fixent à des séquences spécifiques de 4 à 15 nucléotides. Ces séquences sont connues sous le nom de sites de fixation de TF (TFBS). Bien que la reconnaissance d'un TFBS par un TF s'effectue de manière spécifique, il est possible que l'interaction soit effective même si le TFBS présente des variations sur certaines bases. Dès lors, il est dit que les TFBS peuvent être dégénérés. Il est notable que ces éléments en *cis* peuvent être localisés dans différents types de séquences régulatrices, comme le promoteur, le terminateur, les régions non traduites ("Untranslated Transcribed Region" ou UTR), les introns, les exons et des séquences intergéniques (Tu *et al.*, 2020; Burgess *et al.*, 2019; O'Malley *et al.*, 2016; Schmitz *et al.*, 2021).

Les TF fonctionnent en combinaison. Cette dernière induit la présence de modules *cis*-régulateurs (CRM) correspondant à un ensemble de TFBS proches. Les dernières classifications des CRM font état de cinq classes (Schmitz *et al.*, 2021) :

- Les promoteurs centraux correspondent aux séquences minimales permettant l'initiation de la transcription (généralement 50 à 100 bases autour du TSS).
- Les séquences amplificatrices ou enhancers correspondent aux séquences d'ADN qui induisent la transcription du gène cible. Ces séquences peuvent être localisées de manière proximale ou distale (jusqu'à plusieurs centaines de kilobases (kb)) par rapport au gène cible.
- Les séquences inactivatrices ou silencers correspondent aux séquences ADN qui répriment la transcription du gène cible. Comme les enhancers, ces séquences peuvent être localisées de manière proximale ou distale par rapport au gène cible.
- Les éléments multifonctionnels ont à la fois une activité d'enhancer et de silencer selon le stade de développement, les conditions, les cellules ou même les gènes ciblés.
- Les insulators correspondent à des séquences localisées entre les CRM

distaux et les promoteurs centraux et qui, lorsqu'ils sont fixés par des protéines spécifiques, empêchent l'action de l'élément distal sur le gène cible. Aujourd'hui, il n'y a pas de cas rapporté d'insulator chez les plantes.

Il est intéressant de noter que les éléments *cis*-régulateurs sont présents en forte densité dans l'environnement proximal des gènes (Yu *et al.*, 2016; O'Malley *et al.*, 2016; Hammal *et al.*, 2022).

Le promoteur est la région flanquante en 5' du gène, située jusqu'à 1 ou 2 kb en amont du TSS. Cette région est impliquée dans la fixation directe de la machinerie transcriptionnelle, ainsi que dans celle des TF supplémentaires pouvant moduler le recrutement de cette dernière. Elle se compose de deux parties : le promoteur central et le promoteur proximal (Figure 1.5).

Comme déjà décrit précédemment, le promoteur central est la séquence minimale nécessaire pour induire la transcription du gène. Cette séquence comprend l'ensemble des TFBS des facteurs de transcription généraux pour la bonne reconnaissance de la machinerie transcriptionnelle (Hong, 2016)(Figure 1.5) :

- La boite TATA (Lifton *et al.*, 1978) a pour séquence consensus TATAWA. Elle est située entre 25 et 30 bases en amont du TSS et est reconnue par la TATA Binding Protein (TBP).
- La boîte BREu (Lagrange *et al.*, 1998) a pour séquence consensus SSRCGCC. Elle se situe juste en amont de la TATA-box et est fixée par TFIIB.
- La boîte BREd (Deng and Roberts, 2005), de séquence consensus RTDKKKK, se situe juste en aval de la TATA-box et peut être également reconnue par le facteur de transcription général TFIIB.
- La boîte Inr (Smale and Kadonaga, 2003) se caractérise par sa séquence consensus YYANYY située au niveau du TSS. Elle est reconnue par TAF1.
- La boîte DPE (Burke and Kadonaga, 1996) est une séquence ayant pour consensus RGWYV située 30 bases en aval du TSS. Elle est fixée par la protéine TAF6 de la machinerie transcriptionnelle.
- La boîte CCAAT (Bi *et al.*, 1997) a pour séquence GGCCAATCT et se situe entre 60 et 100 en amont du TSS. Elle est fixée par le complexe protéique NF-Y.
- Les motifs TC (Bernard *et al.*, 2010b), de séquence consensus TCTTCTTCT, se situe à environ 39 bases en amont du TSS. Ces séquences peuvent prendrent la place des boîtes BREu et TATA-box dans certains promoteurs. Il n'y a pas de protéines fixatrices de ces motifs répertoriées à ce jour. Sa position peut suggérer que TFIIB et/ou TBP pourraient être capables de reconnaître ces motifs TC.

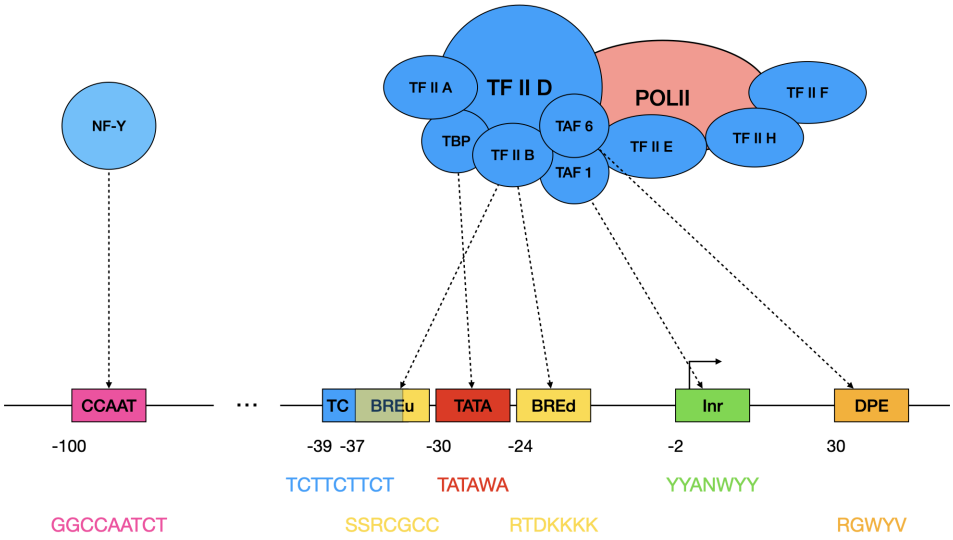


Figure 1.5 – Organisation du promoteur central et de ses interactions avec la machinerie transcriptionnelle chez les plantes.
Chacune des boîtes indiquées correspond à un élément du promoteur central. CCAAT : CCAAT-box ; TC : TC-motifs ; BREu : B Recognition Element upstream ; TATA : TATA-box ; BREd : B Recognition Element downstream ; Inr : motif initiateur ; DPE : Downstream Promoter Element.

En amont du promoteur central, se situe le promoteur proximal qui fixe l'ensemble des autres TF pouvant intervenir dans la régulation du gène cible. Ces TF peuvent influencer l'ouverture chromatinienne au niveau du gène ou réguler (positivement ou négativement) le recrutement de la machinerie transcriptionnelle. Cette régulation est à mettre en relation avec la structure de la région proximale. En effet, pour maintenir une régulation de la transcription, les TFBS présentent des contraintes topologiques les uns par rapport aux autres, ainsi que par rapport au TSS (Grace *et al.*, 2004; Bernard *et al.*, 2010b; Lin *et al.*, 2010; Yu *et al.*, 2016; Chen *et al.*, 2017; Ksouri *et al.*, 2021; Schmitz *et al.*, 2021). Ainsi, à l'échelle d'un génome, les positions des TFBS dans les régions promotrices proximales s'avèrent différentes selon la famille de TF se liant aux sites. Dès lors, une famille de TF donnée aura une affinité plus grande pour certaines portions du promoteur. Cette observation peut s'expliquer par (i) la similarité de séquences des TFBS pour une même famille de TF (Castro-Mondragon *et al.*, 2017) et (ii) la collaboration des TF au sein d'un même module *cis*-régulateur qui nécessite des contraintes topologiques entre les TFBS eux-mêmes (Grace *et al.*, 2004; Schmitz *et al.*, 2021).

Suite à l'initiation, la prochaine étape du processus de transcription est l'élongation. Cette étape correspond à la phase où la POLII se déplace le long de l'ADN

matrice pour synthétiser l'ARN messager. Cette étape peut être régulée par de nombreux facteurs, comme des modificateurs des marques épigénétiques de la chromatine, des remodeleurs de la chromatine ou encore des facteurs d'élongation capables de former un complexe avec la POLII (Godoy Herz and Kornblihtt, 2019; Couvillion *et al.*, 2022). L'intervention de chacun de ces acteurs permet de poursuivre ou mettre en pause l'élongation. Chez les eucaryotes, cette étape se déroule en parallèle de l'épissage des ARN messagers, un mécanisme de régulation post-transcriptionnelle influant sur les isoformes d'ARN messagers obtenus, ainsi que sur les protéines produites. Par conséquent, des modifications de la "vitesse" d'élongation influent sur l'épissage alternatif de l'ARN messager (Godoy Herz and Kornblihtt, 2019) faisant de cette étape un point clé dans la régulation de l'expression génétique. Les travaux que j'ai conduits n'étant pas focalisés sur cette étape du cycle de la transcription, je ne détaillerai pas davantage les régulations qui y sont associées.

1.2.5 . Terminaison de la transcription et libération de la machinerie transcriptionnelle

La partie terminatrice, également appelée région 3' régulatrice, est très fortement impliquée dans la régulation des gènes (Mayr, 2019; Bernardes and Menossi, 2020). Située au niveau du TTS et incluant la partie 3'UTR du gène, cette région possède de nombreux événements de régulation mais reste encore peu étudiée par rapport à la partie promotrice. Elle contient l'ensemble des séquences *cis*-régulatrices reconnues par le complexe moléculaire de clivage et de polyadénylation de l'ARN (CPMC). Ce complexe est constitué d'une vingtaine de protéines nécessaires dans les étapes de clivage et de polyadénylation de l'ARN messager. Il se fixe et agit au niveau de trois éléments *cis*-régulateurs présents dans la région 3' régulatrice et connus sous le nom de "Far Upstream Element" (FUE), de "Near Upstream Element" (NUE) et de "Cleavage Element" (CE) :

- Le FUE est une séquence de 6 à 18 nucléotides, riche majoritairement en T, secondairement en A et minoritairement en G. Cette séquence se situe entre 30 et 140 nucléotides en amont du TTS.
- Le NUE est une séquence riche en A, dont la taille varie entre 6 et 10 nucléotides et qui est située entre 10 et 40 nucléotides en amont du TTS.
- Le CE est une séquence riche majoritairement en T, secondairement en A et minoritairement en C. Cette séquence possède le site de clivage de l'ARN qui sera transcrit.

La reconnaissance de ces trois éléments et le choix du site de clivage et, par extension, de polyadénylation, est une étape clé dans la régulation de l'expression génique puisqu'elle permet d'influer sur la structure du transcrit obtenu. En effet, l'existence de sites alternatifs de polyadénylation (APA) est un des mécanismes régulateurs clés dans la stabilité de l'ARN messager synthétisé, son transport et sa traduction (Mayr, 2019; Bernardes and Menossi, 2020). Il est estimé que plus de la moitié des gènes eucaryotes posséderaient des APA (Shen *et al.*, 2008; Hunt

et al., 2012; Tian and Manley, 2017). Le choix des APA peut aussi être spécifique de lignées cellulaires ou sont impliqués dans des processus biologiques des plantes (Xing and Li, 2011; DeRidder *et al.*, 2012; Tian and Manley, 2017; Ji *et al.*, 2018).

Il est également notable que des TFBS sont aussi présents dans la région 3' régulatrice (O'Malley *et al.*, 2016; Tu *et al.*, 2020; Hammal *et al.*, 2022). Des travaux récents chez l'Homme ont établi que la présence de ces sites est importante pour le choix des APA et a donc un impact sur l'abondance des isoformes d'ARN pre-messager (Kwon *et al.*, 2022). De plus, bien que les interactions entre le promoteur et le terminateur soient encore peu caractérisées, elles semblent nécessaires dans le contrôle de la transcription (Al-Husini *et al.*, 2020). Ainsi, les TFBS présents à la fois dans les régions proximales 5' et 3' des gènes pourraient être impliqués dans les interactions promoteur-terminateur.

En résumé, l'état des connaissances actuelles et les hypothèses qu'elles soulèvent, suggèrent que les régions proximales des gènes sont des acteurs clés de l'expression des gènes et qu'il reste encore de nombreux travaux à poursuivre avant de parvenir à leur caractérisation complète chez les plantes.

1.3 . La caractérisation des régions proximales des gènes

1.3.1 . La quantification de l'efficacité des régions 5'-proximales des gènes : une avancée dans la compréhension des séquences *cis*-régulatrices de cette région

Afin de synthétiser des promoteurs capables d'induire des niveaux d'expression souhaités pour des gènes d'intérêt, une étude a récemment évalué l'efficacité de promoteurs d'*Arabidopsis thaliana*, de *Zea mays* et de *Sorghum bicolor* (Jores *et al.*, 2021). Dans ce but, les auteurs ont mesuré la quantité de transcrits d'un gène rapporteur placé sous le contrôle de différents promoteurs. Cette analyse leur a demandé de mieux caractériser la région 5'-proximale des gènes et de quantifier l'importance des séquences *cis*-régulatrices s'y trouvant.

Cette étude confirme des études précédentes (Kiran *et al.*, 2006; Ponjavic *et al.*, 2006; Mogno *et al.*, 2010) que l'élément *cis*-régulateur le plus critique dans l'efficacité d'un promoteur est la TATA-box dont la position doit être conservée entre 30 et 40 bases en amont du TSS.

Le second élément critique pour l'efficacité d'un promoteur est sa composition nucléotidique. Chez la tomate, il a en effet été observé que les promoteurs ayant un contenu en AT plus fort possèdent une efficacité plus grande (Jores *et al.*, 2021). L'effet inverse a été observé dans les protoplastes de maïs. Ces résultats font écho à la composition nucléotidique des promoteurs du génome de la tomate qui est riches en AT, alors que celui du maïs est riche en GC (Rensink *et al.*, 2005; Singh *et al.*, 2016). La machinerie transcriptionnelle d'une espèce donnée semble donc adaptée à la composition nucléotidique de son génome.

Cette étude a aussi révélé que l'ajout d'une boîte INR et un motif riche en

pyrimidines, respectivement au niveau du TSS et entre la TATA-box et le TSS, peuvent améliorer l'efficacité des promoteurs.

Enfin, l'ajout de TFBS en amont de la TATA-box permet d'accroître l'efficacité des promoteurs. *A contrario*, la présence du même TFBS entre la TATA-box et le TSS réduit considérablement la transcription des rapporteurs. Ces résultats suggèrent une modulation de l'activité de TF (activateur ou répresseur) selon la position à laquelle il se fixe. D'un point de vue mécanistique, ceci peut s'expliquer par un blocage de la fixation de la machinerie transcriptionnelle lorsque le TF est situé dans le promoteur central.

En résumé, cette étude récente a permis de mettre en avant l'importance des séquences *cis*-régulatrices de la région 5'-proximale pour moduler la transcription de gènes cibles.

1.3.2 . La région 3'-proximale des gènes : une région encore peu étudiée à l'échelle des génomes

La région 3'-proximale reste peu étudiée, bien qu'elle soit impliquée dans la régulation de la transcription des gènes et les nombreux processus associés (Mayr, 2019; Bernardes and Menossi, 2020). Ainsi, il n'est pas connu d'approches expérimentales à l'échelle des génomes focalisées sur la caractérisation de la région 3'-proximale comme il en existe pour la région 5'-proximale (Jores *et al.*, 2021). Les données disponibles sur la structure de la région 3'-proximale et les séquences *cis*-régulatrices qui la composent demeurent donc lacunaires comparées à celles disponibles pour la région 5'-proximale. Néanmoins, des études réalisées sur certaines régions 3'-proximales d'*A. thaliana* ou de *Flaveria bidentis* ont permis de mettre en avant la capacité des celles-ci à induire des niveaux d'expression génique plus forts que ceux induits par les terminateurs viraux "nopaline synthase" (NOS) et "octopine synthase" (OCS) classiquement utilisés (Bernardes and Menossi, 2020). Ces données soulignent donc l'intérêt qu'il y a à poursuivre la caractérisation de cette région à l'échelle des génomes des plantes.

1.3.3 . Étude des régions proximales des gènes à l'aide d'approches expérimentales

Aujourd'hui, de nombreuses méthodes permettent de mieux caractériser les régions proximales. On peut diviser les efforts menés en deux types d'approches : les approches expérimentales et les approches *in silico*.

Identification des régions accessibles de la chromatine

L'identification de régions accessibles de la chromatine à l'échelle des génomes participe à la caractérisation des régions proximales des gènes. En effet, cela permet de définir dans quels contextes (environnemental, cellulaire ou développemental) ces régions sont accessibles (Schmitz *et al.*, 2021; Minnoye *et al.*, 2021). Les méthodes présentées ci-après sont des méthodes récentes faisant appel au séquençage haut débit ("Next Generation Sequencing" ou NGS) des fragments accessibles ou

nucléosomaux de la chromatine, selon la technologie utilisée. Après séquençage, les fragments (aussi appelés "reads") sont alignés contre le génome de référence pour obtenir un profil de couverture. Enfin, la dernière étape consiste à détecter des pics dans la couverture des reads sur le génome pour identifier les régions fortement couvertes par les reads et donc correspondant à de l'ADN libre ou nucléosomal.

Les méthodes couramment utilisées se distinguent par les acteurs moléculaires mis à contribution :

— Le MNase-seq ("Micrococcal Nuclease-sequencing") (Schones *et al.*, 2008) est une méthode qui donne accès à la position des nucléosomes. L'ADN est mis en présence de nucléase micrococcale (MNase), une enzyme qui digère l'ADN s'il n'est pas enroulé autour des nucléosomes (ADN non protégé). Par comparaison avec un échantillon contrôle n'ayant pas subit la digestion MNase, il est donc possible de déterminer indirectement les sites d'accessibilité de la chromatine.
— Le DNase-seq ("DNase I hypersensitive sites sequencing") (Boyle *et al.*, 2008) est une technique basée sur le même principe que le MNase-seq, mais qui utilise une autre enzyme, la DNase I, pour couper l'ADN lorsqu'il est accessible. Les extrémités des fragments coupés sont ensuite séquencées. Un fort signal DNase-seq indique donc une région chromatinienne accessible.
— Le FAIRE-seq ("Formaldehyde-Assisted Isolation of Regulatory Elements sequencing") (Giresi *et al.*, 2007) peut être considéré comme une méthode dérivée du DNase-seq. Cette technique fait appel au formaldehyde qui permet de fixer la chromatine et l'ensemble des protéines qui y sont liées. Une fois l'étape de fixation réalisée, l'ADN est fragmenté par sonication, puis purifié afin que seuls les fragments non liés de l'ADN soient conservés. Ils sont ensuite séquencés.
— L'ATAC-seq ("Assay for Transposase-Accessible Chromatin with highthroughput sequencing") (Buenrostro *et al.*, 2013, 2015) fait appel à la transposase Tn5, une transposase mutée hyperactive, capable d'insérer des fragments spécifiques dans les régions non-nucléosomales. L'ADN est ensuite purifié pour récupérer les fragments libres compris entre les insertions qui sont ensuite amplifiés et séquencés. Comme la Tn5 est une version très active, elle donne la position de régions plus compactées que la DNase I. Par ailleurs, il est notable que la DNase I et la Tn5 peuvent agir à une position de l'ADN protégée par un TF. Il est donc possible de déduire la position des TF grâce au DNase-seq et à l'ATAC-seq si la couverture en reads est suffisammment forte.
— Le ChIP-seq ("Chromatin ImmunoPrecipitation and sequencing") (Barski *et al.*, 2007) est une technique qui permet d'analyser les interactions protéines-ADN. Dans le cas de recherche de régions accessibles de la chromatine, le ChIP-seq est utilisé pour cibler les histones ayant subit des modifications

post-traductionnelles particulières, comme des acétylations ou des méthylations. Les fragments d'ADN associés à des histones qui possèdent les marques recherchées, sont récupérés par immunoprécipitation en utilisant des anticorps spécifiques de l'histone.

— le NOMe-seq ("Nucleosome Occupancy and Methylome sequencing") (Lay *et al.*, 2018) est une méthode basée sur l'utilisation de méthyltransférases. Cette méthode permet également de définir les régions accessibles de la chromatine en utilisant l'activité méthyltransférase de M.CviPI, qui ne méthyle que des sites accessibles de la chromatine.

Les approches ci-dessus se révèlent complémentaires et sont extrêmement précieuses pour déterminer les régions proximales accessibles, et donc actives, à l'échelle des génomes. Cependant, seules, elles ne permettent pas de déterminer les acteurs intervenant dans chacune de ces régions. Dans l'optique de caractériser les régions proximales, elles doivent donc être couplées avec d'autres approches.

Identification des sites de fixation de facteurs de transcription

Une grande partie des méthodes actuelles qui contribuent à la caractérisation des régions proximales des gènes, se focalisent sur la recherche de sites de fixation de facteurs de transcription (Lai *et al.*, 2019) :

— Faisant écho à la partie précédente, le ChIP-seq est fréquemment appliqué pour cibler les TF et identifier leurs fixations sur l'ADN. Cette méthode *in vivo* nécessite de pouvoir (1) isoler assez de matériel biologique, ce qui peut-être délicat (dans les premiers stades de méristèmes floraux par exemple) et (2) disposer d'un anticorps spécifique pour le TF étudié. Par ailleurs, il est difficile de savoir si la fixation est directe car le TF ciblé peut être lié directement à l'ADN ou former un complexe avec une protéine qui elle est liée directement à l'ADN.

— PBM ("Protein Binding Microarrays") (Berger and Bulyk, 2006) est une approche pré-NGS qui se sert des technologies de puces à ADN pour déterminer les séquences d'ADN reconnues par les TF. Une étiquette épitope est tout d'abord ajouté aux TF. Ceux-ci sont ensuite confrontés à une banque d'ADN double-brins. La fixation des TF sur l'ADN est révélée par fluorescence après élimination des protéines non fixées par lavage, suivie de l'ajout d'un anticorps spécifique des étiquettes épitopes.

— Le SELEX-seq ("Systematic Evolution of Ligands by Exponential Enrichment sequencing") (Jolma *et al.*, 2010) met en solution des TF et l'ensemble des séquences ADN de 14 bases. Après la mise en solution, une étape de lavage et d'élution permet de ne garder, théoriquement, que les séquences fixées par les TF. Une partie d'entre elles sont ensuite séquencées. Le principe de la méthode est de réitérer ce cycle d'étape en remplaçant le mélange de séquences d'ADN initial par celui gardé après chaque cycle. La réitération permet d'enrichir la solution en fragments d'ADN réellement

fixés et ainsi d'éliminer un maximum de séquences non reconnues par les TF mais ayant passées les étapes de filtres. Cela permet ainsi d'augmenter la qualité du site de fixation identifié.

— Le DAP-seq ("DNA affinity purification sequencing") (O'Malley *et al.*, 2016) est une technologie *in vitro* permettant d'analyser les sites de liaison de TF à l'échelle des génomes. Pour cela, le DAP-seq nécessite la synthèse de TF recombinant auquel est ajouté une étiquette d'affinité fixé sur des billes magnétiques. Ces complexes TF recombinant-billes magnétiques sont ensuite mis en solution avec l'ensemble de l'ADN génomique dépourvu de ses nucléosomes et fragmenté. Après purification et élution, les fragments d'ADN génomique fixés par les TF recombinants élués sont séquencés.
Le DAP-seq est une approche qui a grandement apporté à la compréhension des mécanismes de régulation transcriptionnelle et à la caractérisation des régions proximales des gènes. Par rapport aux autres approches *in vitro* citées ci-avant, elle présente l'avantage de travailler directement sur le matériel génomique. Le DAP-seq présente des avantages par rapport au ChIP-seq également puisqu'il ne nécessite pas de posséder un anticorps spécifique des TF étudiés, ce qui permet de passer à une échelle supérieure et d'analyser un grand nombre de TF. Cependant, le DAP-seq étant une technique *in vitro*, elle ne prend pas en compte entièrement le contexte génomique et d'éventuels co-facteurs nécessaires pour la fixation du TF étudié comme le permet le ChIP-seq.

— Le MOA-seq ("MNase-defined cistrome-Occupancy Analysis") (Savadel *et al.*, 2021) est un dérivé du MNase-seq. Son objectif est d'identifier des marques de présences de TF. Pour cela, une digestion partielle de l'ADN par la MNase est réalisée afin de maintenir également les portions occupées par d'autres protéines que les nucléosomes. Ensuite une étape de sélection de la taille des fragments issus de la digestion est appliquée afin de ne garder que des séquences courtes (< 30 bases) supposées fixées par des TF. L'ensemble des étapes suivantes du traitement des données issues du séquençage sont les mêmes que les autres approches NGS. Ainsi le MOA-seq permet d'identifier *de novo* à l'échelle d'un génome des sites putatifs de fixation de TF avec une grande résolution. Cette méthode est complémentaire de celles citées ci-avant puisqu'elle focalise directement sur l'ensemble des portions d'ADN libres fixées par des TF tandis que les autres se focalisent sur un TF donné et recherchent ses sites de fixation.

1.3.4 . Étude des régions proximales des gènes à l'aide d'approches *in silico*

À l'instar des approches expérimentales citées précédemment, les approches *in silico* ont aussi grandement contribuées à la caractérisation des régions proximales des gènes. Le développement de nouvelles approches *in silico* a suivi le développement des technologies expérimentales. En effet, aujourd'hui, la majorité des efforts

se focalisent sur la modélisation de TFBS à partir des données expérimentales NGS (Lai *et al.*, 2019).

Modélisation des TFBS à partir de données expérimentales

La modélisation la plus fréquemment utilisée pour les TFBS est la matrice poids-position (PWM). Cette matrice est construite à partir d'une matrice de probabilité de position (PPM), qui contient la fréquence de chaque nucléotide à chaque position du TFBS. Chaque fréquence observée contenue dans la matrice est ensuite pondérée par la fréquence attendue du nucléotide convertie dans une échelle logarithmique. Ainsi, les PWM permettent de donner un score à une séquence d'ADN testée.

D'autres modèles ont été développés pour prendre en considération des informations supplémentaires dans la modélisation afin d'améliorer la qualité des prédictions. Le modèle "Transcription Factor Flexible Model" (TFFM) prend ainsi en compte la dépendance entre les nucléotides (Mathelier and Wasserman, 2013). D'autres modèles prennent aussi en compte la conformation de l'ADN (Mathelier *et al.*, 2016) ou les énergies d'interactions entre acide aminé et nucléotide (Ruan *et al.*, 2017). Bien que ces nouvelles modélisations puissent parfois se révéler plus efficaces que les PWMs, ces dernières restent aujourd'hui les plus fréquemment utilisées de par leur simplicité de construction et leur efficacité dans la prédiction de TFBS.

Les efforts réalisés aujourd'hui pour avancer dans ce domaine de recherche ne s'articulent pas uniquement autour de la modélisation des TFBS. En effet, pour valoriser l'ensemble des ressources générées à partir des modèles cités précédemment, le développement de bases de données et d'outils dédiés au stockage et à l'exploitation des TFBS identifiés est une nécessité. De nombreuses bases de données peuvent aujourd'hui être répertoriées et se classer suivant deux catégories. Une partie de ces bases sont dédiées à l'intégration de l'ensemble des pics obtenus à partir de données expérimentales. Celles-ci donnent ainsi accès aux localisations des TFBS à l'échelle génomique (Puig *et al.*, 2021; Hammal *et al.*, 2022). Les autres bases de données mettent à disposition un catalogue de TFBS (majoritairement sous forme de PWM) identifiés à partir des données expérimentales (Weirauch *et al.*, 2014; Castro-Mondragon *et al.*, 2022). Dans l'objectif d'exploiter ces ressources expérimentales, de nombreux outils ont été mis au point pour offrir la possibilité de scanner des séquences ADN d'intérêt et prédire la présence d'un TFBS donné (Turatsinze *et al.*, 2008; Grant *et al.*, 2011; Chow *et al.*, 2019).

Ainsi, l'ensemble du travail réalisé par les approches *in silico* présentées jusqu'ici représente une grande avancée pour l'identification de TFBS dans les régions proximales des gènes. En parallèle de ces approches, d'autres visent à identifier *de novo* des séquences *cis*-régulatrices à l'aide d'une recherche de motifs sur-représentés dans des lots de séquences.

Identification *de novo* de séquences *cis*-régulatrices par la recherche de motifs ADN sur-représentés

À la fin des années 70, dans un contexte où la quantité de séquences ADN séquencées était grandissante, de premiers développements bioinformatiques ont été nécessaires pour extraire de l'information de ces données (Stormo, 2000). En partant du postulat initial que des portions d'ADN fonctionnelles doivent être sélectionnées au cours de l'évolution, la recherche de motifs sur-représentés dans des séquences d'ADN fût au coeur des premiers développements. Les applications de ces premières méthodes étaient réalisées sur des séquences promotrices afin d'identifier des TFBS. Ainsi, de nombreuses méthodes ont vu le jour et restent encore aujourd'hui utilisées. Les méthodes les plus fréquemment utilisées se classent en deux catégories définies selon la représentation du motif utilisée.

En effet, certains algorithmes, tels que l'échantillonnage de Gibbs (Lawrence *et al.*, 1993) ou l'Expectation Maximization (Bailey and Elkan, 1994), utilisent des PWM pour identifier des motifs sur-représentés. Ces deux approches s'initialisent avec des PWM aléatoires. Au cours d'itérations successives, la PWM est ensuite ajustée jusqu'à convergence de l'algorithme pour obtenir un motif sur-représenté stable. Ces algorithmes sont aujourd'hui facilement accessibles au travers de suites d'outils (Thijs *et al.*, 2002; Bailey *et al.*, 2009; Defrance and van Helden, 2009).

A l'inverse des approches basées sur les PWM, certaines méthodes proposent de comparer les occurrences d'un motif représenté sous forme de k-mer, à savoir une courte séquence consensus d'ADN de longueur k, au sein des séquences étudiées par rapport à un comptage attendu. Les méthodes utilisant des k-mers varient quant à l'obtention des comptages attendus. Par exemple, ces derniers peuvent être calculés en déterminant le nombre d'occurrences du k-mer au sein de l'ensemble du génome étudié (van Helden *et al.*, 1998). Ils peuvent aussi être déterminés à l'aide de chaînes de Markov qui modélisent les probabilités de transition entre chaque nucléotide au sein du génome ou certaines portions du génome (Schbath and Hoebeke, 2011; Thomas-Chollier *et al.*, 2012). Ces comptages attendus peuvent aussi être obtenus en dénombrant le nombre d'occurrences du k-mer étudié dans un set de séquences aléatoire ou négatif (Saad *et al.*, 2018). Enfin, d'autres méthodes se servent des k-mers, en ajoutant de la connaissance *a priori* dans la détection des motifs. C'est le cas des méthodes qui se focalisent sur la recherche de séquences *cis*-régulatrices proximales en prenant en compte leurs contraintes de distance avec les gènes. Ainsi, ces méthodes cherchent à identifier des motifs sur-représentés localement dans les séquences étudiées (Helden *et al.*, 2000; Yamamoto *et al.*, 2007; Bernard *et al.*, 2010a).

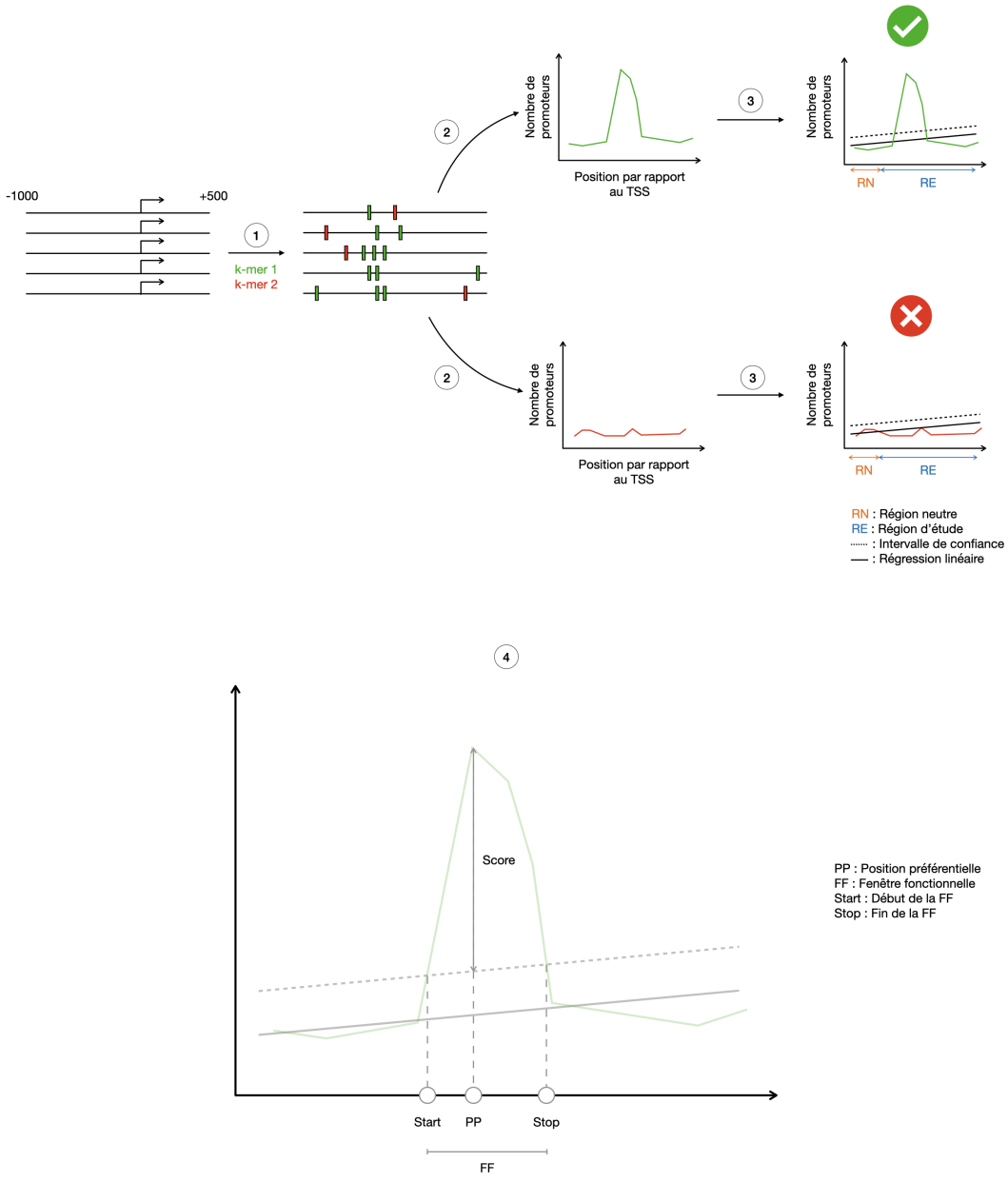
La prise en compte des contraintes topologiques : PLMdetect

La méthode PLMdetect est une méthode *in silico* décrite pour la première fois par (Bernard *et al.*, 2010a). Elle repose sur le constat que les séquences *cis*-

régulatrices dans les promoteurs sont contraintes topologiquement comme le suggérait les motifs du promoteur central, localisés à des distances fixes des TSS. De manière concrète, la méthode PLMdetect vise à identifier, dans un jeu de séquences d'intérêt, des Motifs Préférentiellement Localisés (PLM). Ces derniers correspondent à des motifs sur-représentés à une localisation préférentielle du TSS. Pour se faire, PLMdetect prend deux informations en entrée : les promoteurs de gènes d'intérêt et un ensemble de k-mer à tester. Les promoteurs fournis à PLMdetect sont de même longueur afin que les TSS soient alignés. Pour chaque k-mer, PLMdetect compte ensuite son nombre d'occurrences à chaque position. De ce fait, une distribution de chaque k-mer est obtenue. PLMdetect considère ensuite les premières bases des séquences (généralement 500 bases) comme une région neutre, dans laquelle aucune accumulation de PLM n'est attendue, afin de calculer une régression linéaire. Dans la région étudiée, située juste en aval de la région neutre, les valeurs prédites sont calculées avec un intervalle de confiance à 99%. Enfin, dans cette région d'étude, si la distribution du motif excède l'intervalle de confiance, alors le motif est considéré comme un PLM. À ce PLM peut être alors associé une position préférentielle, définie comme la position à laquelle le maximum d'occurrences du k-mer est atteint dans la portion où le distribution excède l'intervalle de confiance. Une fenêtre fonctionnelle est aussi définie comme les bornes de la portion où la distribution du k-mer est supérieure à l'intervalle de confiance. Enfin, un score est attribué à chaque PLM. Ce score correspond à la différence entre la valeur de la distribution du k-mer à la position préférentielle et l'intervalle de confiance à la même position (Figure 1.6).

La méthode a tout d'abord été implémentée afin d'explorer le contenu des promoteurs d'*A. thaliana* et de *Oryza sativa*. Ce développement a permis de mettre en lumière de nouveaux motifs du promoteur central chez les deux espèces : les TC-motifs et des variants de la TATA-box (Bernard *et al.*, 2010b). La méthode a également été utilisée à des échelles plus réduites afin d'identifier des séquences *cis*-régulatrices sous-jacents la co-régulation de gènes différentiellement exprimés (Bueso *et al.*, 2014, 2016; Frei dit Frey *et al.*, 2014; Cuello *et al.*, 2019; Del Prete *et al.*, 2019).

Depuis son développement, les hypothèses sur lesquelles repose la méthode PLMdetect, à savoir les contraintes topologiques des séquences *cis*-régulatrices proximales, se sont confirmées dans le cadre d'études réalisées chez plusieurs plantes (Yu *et al.*, 2016; Ksouri *et al.*, 2021; Savadel *et al.*, 2021). Dès lors, l'analyse exhaustive des PLM chez plusieurs espèces de plantes semble d'un grand intérêt pour poursuivre la caractérisation des régions proximales des gènes.



1 : Entrée des k-mer 1 et k-mer 2 à rechercher
2 : Comptage et obtention des distributions des k-mer 1 et k-mer 2 le long des séquences
3 : Calcul de la régression linéaire à partir de RN et calcul des valeurs prédites avec un intervalle de confiance à 99% dans RE
4 : Récupération des informations de position et score associées au PLM identifié (ici k-mer 1)

Figure 1.6 – Représentation schématique du fonctionnement de PLMdetect. *Cet exemple présente le cas de k-mer 1, détecté comme PLM, et k-mer 2, non PLM. Pour chaque k-mer, PLMdetect compte son nombre d'occurrences à chaque position de la région étudiée et obtient pour chacun une distribution. En considérant les 500 premières bases comme région neutre, la méthode PLMdetect calcule une régression linéaire sur cette région. La méthode calcule ensuite les valeurs prédites (avec un intervalle de confiance à 99%) dans le reste de la région, considérée comme la région d'étude. La méthode détermine ensuite la présence d'un pic si la distribution dans la région d'étude excède l'intervalle de confiance. Le motif est alors considéré comme un PLM. Pour chaque PLM, la méthode détermine ensuite les indicateurs positionnels et le score.*

1.4 . Objectifs de la thèse

Les régions proximales, en 5' comme en 3' des gènes, sont des acteurs clés dans la régulation de l'expression des gènes. Cependant, les connaissances que nous avons de ces régions et, en particulier, des séquences *cis*-régulatrices qui s'y trouvent, restent encore parcellaires. Par conséquent, cette thèse a eu pour ambition de progresser quant à la compréhension de ces séquences *cis*-régulatrices grâce à la détection de PLM à différentes échelles.

La première contribution porte sur la caractérisation de l'ensemble des PLM de deux espèces de plantes aux génomes contrastés, *A. thaliana* et *Z. mays*. J'ai pour cela étendu PLMdetect pour réaliser une détection *de novo* des PLM dans les régions proximales en 5' et en 3' des gènes de chacune des deux espèces étudiées. L'objectif visé était de déterminer dans quelle mesure les différences architecturales des génomes impactent le contenu en PLM et quels sont les mécanismes sous-jacents la présence des PLM.

La deuxième contribution porte sur le développement d'une interface web permettant une utilisation facile de PLMdetect chez un grand nombre de plantes. L'objectif visé était de permettre une caractérisation des séquences *cis*-régulatrices des régions proximales des gènes chez chacune des espèces étudiées et, à plus long terme, de pouvoir réaliser des comparaisons multi-espèces, et ainsi, de progresser quant à la compréhension des mécanismes sous-jacents l'expression des gènes.

Enfin, la troisième contribution porte sur une étude de cas qui s'inscrit dans le cadre de l'étude de la réponse des plantes aux stress, et en particulier le rôle des éléments *cis*-régulateurs à proximité des gènes dans cette réponse. Dès lors, cette dernière contribution porte sur l'étude du rôle des PLM comme marqueurs moléculaires originaux dans la réponse globale au stress chez *A. thaliana*.

2 - Caractérisation des PLM chez *A. thaliana* et *Z. mays*

2.1 . Résumé du chapitre

Le travail mené visait à identifier et à caractériser l'ensemble des PLM présents chez les plantes. Dans ce but, deux espèces végétales aux génomes contrastés, ont été sélectionnées, ce qui permettait de déterminer si leurs différences génomiques pouvaient impacter le contenu en PLM : (1) *A. thaliana* qui possède un génome de 135 Mb avec environ 28 000 gènes, riche en AT et composé de 20% d'éléments transposables (Berardini *et al.*, 2015) et (2) *Z. mays* qui possède un large génome d'environ 2,3 Gb avec près de 40 000 gènes, riche en GC et composé d'au moins 80% d'éléments transposables (Portwood *et al.*, 2019).

Pour identifier l'ensemble des PLM présents chez *A. thaliana* et *Z. mays,* il a été nécessaire d'adapter la méthode PLMdetect (Bernard *et al.*, 2010a) pour réaliser une détection *de novo* dans les deux régions proximales (5' et 3') des gènes des deux espèces sélectionnées. Pour *A. thaliana*, la détection a abouti à l'identification de 6 998 PLM dans la région 5'-proximale (5'-PLM) et 7 447 PLM dans la région 3'-proximale (3'-PLM). Pour *Z. mays*, l'analyse a permis d'identifier 9 768 5'-PLM et 6 639 3'-PLM.

D'un point de vue structural, les analyses ont montré que les PLM ne sont pas distribués uniformément dans chacune des régions proximales de chaque espèce. Trois groupes de PLM ont ainsi pu être identifiés pour chaque région chez chacune des espèces étudiées. Les comparaisons réalisées ont aussi révélé que 7% des 5'-PLM et 14% des 3'-PLM sont partagés entre les deux espèces. Il est notable que près de 90% des PLM conservés se situent dans les 200 bases qui entourent le TSS ou le TTS. Cette plus grande proximité du contenu en 3'-PLM entre les deux espèces est également visible au niveau des profils de distribution. L'ensemble de ces observations souligne l'importance de la région 3'-proximale dans la structure génomique.

D'un point de vue fonctionnel, les PLM *de novo* identifiés ont été comparés à deux types ressources distinctes : (1) des TFBS déterminés à partir de données expérimentales (Fornes *et al.*, 2020; Tu *et al.*, 2020) et (2) des séquences de microARN (Dai *et al.*, 2018). Ces comparaisons ont permis d'assigner 21% des PLM *de novo* identifiés à une de ces ressources dans les deux régions proximales étudiées. De manière surprenante, 79% des PLM identifiés ne peuvent pas être assignés aux ressources interrogées et ont donc été qualifiés de "unassigned PLM" (uPLM). Afin de les caractériser plus en avant, des enrichissements fonctionnels ont été réalisés pour les groupes de gènes possédant des PLM. Ces enrichissements montrent que les uPLM apportent des prédictions fonctionnelles distinctes des autres types de

PLM identifiés. En parallèle, une partie des uPLM a pu être appuyée par des données expérimentales de type MOA-seq qui décrivent les portions accessibles de la chromatine avec une grande résolution (Savadel *et al.*, 2021). Ces résultats suggèrent donc qu'une partie, au moins, des uPLM correspond à des séquences *cis*-régulatrices.

Pour conclure, l'ensemble des travaux menés a fourni une cartographie exhaustive des PLM dans les régions 5' et 3'-proximales des gènes d'*A. thaliana* et *Z. mays*. Ces travaux ont aussi permis de montrer que la recherche de PLM peut être utilisée pour progresser quant à la fonction de gènes peu annotés.

2.2 . A comprehensive map of preferentially located motifs reveals distinct proximal *cis*-regulatory sequences in plants

Les travaux réalisés dans ce chapitre ont été valorisés sous la forme d'un préprint (Rozière *et al.*, 2022a) et d'un article validé par les pairs (Rozière *et al.*, 2022b) dont je suis le 1er auteur et co-auteur en charge de la correspondance.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 October 2022
DOI 10.3389/fpls.2022.976371

OPEN ACCESS

EDITED BY
Victoria Mironova,
Radboud University, Netherlands

REVIEWED BY
Zongliang Chen,
Rutgers, The State University of New Jersey, United States
Elena Zemlyanskaya,
Institute of Cytology and Genetics (RAS), Russia

*CORRESPONDENCE
Julien Rozière
julien.roziere@inrae.fr
Marie-Laure Martin
Mlmartin@Agroparistech.Fr
Sylvie Coursol
sylvie.coursol@inrae.fr

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 23 June 2022
ACCEPTED 21 September 2022
PUBLISHED 12 October 2022

CITATION
Rozière J, Guichard C, Brunaud V, Martin M-L and Coursol S (2022) A comprehensive map of preferentially located motifs reveals distinct proximal *cis*-regulatory sequences in plants.
Front. Plant Sci. 13:976371.
doi: 10.3389/fpls.2022.976371

A comprehensive map of preferentially located motifs reveals distinct proximal *cis*-regulatory sequences in plants

Julien Rozière[1,2,3]*, Cécile Guichard[1,2], Véronique Brunaud[1,2], Marie-Laure Martin[1,2,4]* and Sylvie Coursol[3]*

[1]Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), Gif sur Yvette, France, [2]Université de Paris Cité, Institute of Plant Sciences Paris-Saclay (IPS2), Gif sur Yvette, France, [3]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), Versailles, France, [4]Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris-Saclay, Palaiseau, France

Identification of *cis*-regulatory sequences controlling gene expression is an arduous challenge that is being actively explored to discover key genetic factors responsible for traits of agronomic interest. Here, we used a genome-wide *de novo* approach to investigate preferentially located motifs (PLMs) in the proximal *cis*-regulatory landscape of *Arabidopsis thaliana* and *Zea mays*. We report three groups of PLMs in both the 5′- and 3′-gene-proximal regions and emphasize conserved PLMs in both species, particularly in the 3′-gene-proximal region. Comparison with resources from transcription factor and microRNA binding sites shows that 79% of the identified PLMs are unassigned, although some are supported by MNase-defined cistrome occupancy analysis. Enrichment analyses further reveal that unassigned PLMs provide functional predictions that differ from those derived from transcription factor and microRNA binding sites. Our study provides a comprehensive map of PLMs and demonstrates their potential utility for future characterization of orphan genes in plants.

KEYWORDS

gene expression, *cis*-regulatory elements, preferentially located motifs, gene-proximal regions, gene regulatory network, plant

Introduction

As sessile organisms, plants, must adapt to local constraints such as bacteria, fungi, and pests, as well as to environmental changes. One of the fundamental drivers of their

adaptation is the activation or repression of gene transcription (Waters et al., 2017; Alonge et al., 2020; Azodi et al., 2020; Liu et al., 2020; Zhou et al., 2022). These processes are tuned by numerous *cis*-regulatory DNA sequences, the characterization of which is a central question for the complete understanding of transcriptional response mechanisms [for a recent review, see (Schmitz et al., 2021)]. Numerous experimental and predictive efforts (Lai et al., 2019; Savadel et al., 2021; Schmitz et al., 2021) have been made to characterize them, highlighting several *cis*-regulatory regions. In addition to the distal *cis*-regulatory DNA sequences, which include enhancers (Fagny et al., 2021) and can be more than 1 Mbp away from their target gene, there are the 5'- and 3'-gene-proximal regions that are rich in *cis*-regulatory DNA sequences (Li et al., 2012; Wallace et al., 2014; Zemlyanskaya et al., 2021).

The 5'-gene-proximal region is located in the bases framing the transcription start site (TSS) and includes the core promoter and the promoter. The core promoter is directly involved in the binding of the transcription initiation complex and is by definition essential for gene expression. The promoter is a region upstream of the core promoter that is involved in binding of many additional transcription factors (TFs) that can modulate basal gene expression (Schmitz et al., 2021). The 3'-gene-proximal region, also called terminator, is a *cis*-regulatory region that is strongly involved in regulating gene expression, but unlike the 5'-gene-proximal region, it has been little studied (Mayr, 2019; Bernardes and Menossi, 2020). It frames the transcription termination site (TTS) and is known to influence gene transcription termination by allowing the binding of the cleavage and polyadenylation complex (CPMC). This region is also rich in TF binding sites (TFBSs) and may interact with the 5'-gene-proximal region through the phenomenon of gene looping (Wang et al., 2010). Despite our current knowledge of these regions fundamental to gene function, our understanding remains incomplete, and much effort is still required to achieve their complete characterization at the genome level.

Interestingly, the *cis*-regulatory DNA sequences in these two gene-related regions appear to be associated with fixed topological constraints. This observation holds for all core promoter sequences with motifs such as the TATA-box, which is located about 30 bases upstream of the TSS (Yamamoto et al., 2007; Bernard et al., 2010; Jores et al., 2021). This phenomenon is also observed in the promoters of several plant species for TFBSs, which occupy preferential position depending on the associated TF family (Yu et al., 2016; Ksouri et al., 2021). Finally, the sites involved in CPMC binding in the 3'-gene-proximal region also show topological constraints with respect to the TTS (Bernardes and Menossi, 2020). Based on this biological context and in order to contribute to a better characterization of these gene-proximal regions in plants, we propose to use and extend an *in silico* method called PLMdetect (Preferentially Located Motif detection) (Bernard et al., 2010). Originally, this method aimed to identify DNA motifs in *Arabidopsis thaliana* that are overrepresented at a specific position compared with TSS and are therefore referred to as preferentially located motifs (PLMs) (Bernard et al., 2010; Bueso et al., 2014; Frei dit Frey et al., 2014; Martínez et al., 2015).

Here, we performed a genome-wide and *de novo* PLMdetect-based study of the 5' and 3'-proximal regions of genes from *A. thaliana* and *Zea mays*. We aimed to determine the extent of which their differences in genome content and genome architecture were reflected in the characteristics of their PLMs in both gene-proximal regions. Our results revealed the organizing principle of the plant PLM landscape and provide a valuable resource for the characterization of unannotated genes in plants.

Results

Implementation of large-scale and *de novo* PLM detection

To define the PLM profile associated with the 5'- and 3'-gene-proximal regions, we extended the PLMdetect method (Bernard et al., 2010). Given the 5'-gene-proximal regions and a motif, this method first calculates the number of motif occurrences at each position in the sequence to obtain a motif distribution. Second, a linear regression is calculated for a neutral region defined as the first 500 bp of the 5'-gene-proximal region where no accumulation of PLMs is expected. Third, in the region under study, the predicted values are calculated with a confidence interval of 99%. If the observed occurrence distribution exceeds the confidence interval, the motif is considered as a PLM. Thus, a PLM is visually defined by a motif distribution that has a peak in the region under study, indicating that it is statistically overrepresented at a preferential distance from the TSS. The PLM is characterized by (i) its preferential position, defined as the position of the peak's top, (ii) a functional window, defined as the portion located between the peak boundaries, and (iii) a score defined as the difference between the peak's top value and the upper bound of the confidence interval at the preferential position. To implement large-scale and *de novo* PLM detection, we also investigated the 3'-gene-proximal regions by computing the motif distribution according to the TTS and considered all non-polymorphic DNA 4-mers to 8-mers (Figure 1).

Genome-wide PLM identification in gene-proximal regions of *A. thaliana* and *Z. mays*

Distribution by score values revealed two populations of PLMs with a score less than or greater than 2 in each gene-

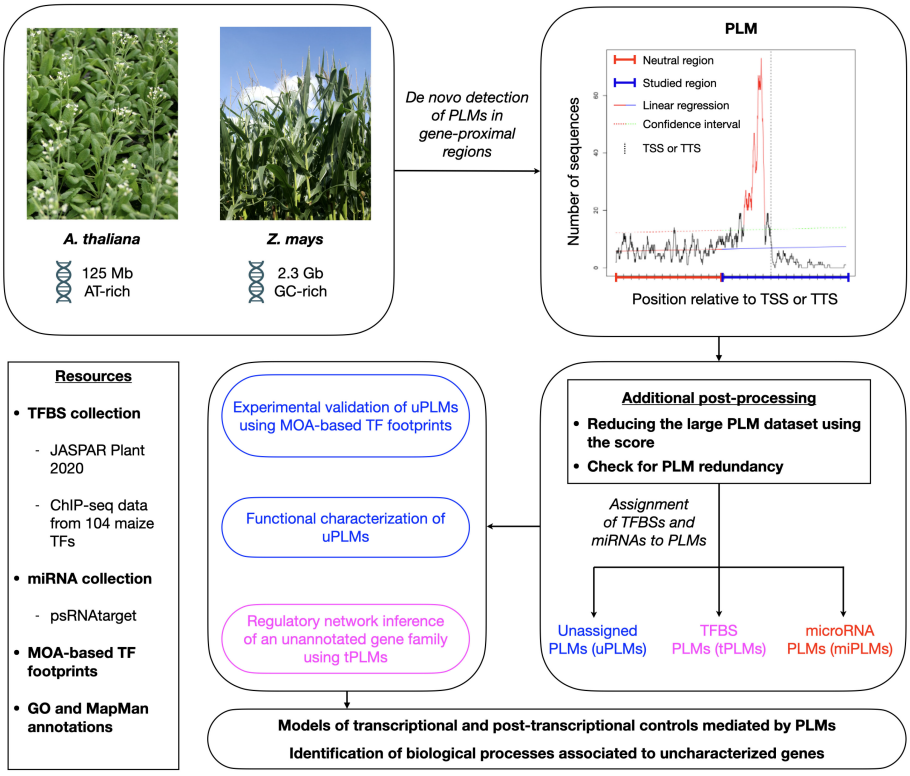


FIGURE 1
Schematic overview of the workflow used in this study for de novo and genome-wide quantification, characterization, and exploitation of PLMs in gene-proximal regions of *A. thaliana* and *Z. mays*. We first detected PLMs in the 5'- and 3'-proximal regions of genes from both plant species. Next, we selected the PLMs according to their score, checked for PLM redundancy and determined whether some of the detected PLMs might be TFBSs (referred to as tPLMs) or targeted by miRNAs (referred to as miPLMs) using distinct resources of TFBSs [JASPAR Plant 2020 (Fornes et al., 2019) and top 1% k-mers from ChIP-seq data of 104 maize TFs (Tu et al., 2020)] and miRNA binding sites [psRNATarget (Dai et al., 2018)]. Unassigned PLMs (referred to as uPLMs)-containing gene sets were functionally characterized with GO and MapMan enrichments relative to the genome. Moreover, we showed that some 5'-uPLMs were supported by MOA-based TF footprints data (Savadel et al., 2021) of Z. mays. Using 5'-tPLMs, we finally inferred the regulatory network of a poorly characterized maize-specific gene family. Collectively, our results provide functional inside into the regulatory role of PLMs and highlight biological processes associated to uncharacterized genes.

proximal region of each species (Supplementary Figure 1). A score greater than 2 indicates a position where the occurrence of the motif is very high compared with the value calculated from the neutral region. In addition, the PLM subpopulation described by a score above 2 was smaller than the one with a score below 2. Both arguments led us to consider only the PLM population with a score above 2 to characterize the 5'- and 3'-gene-proximal regions. We identified 6,998 and 9,768 (7,447 and 6,639) PLMs in the 5' (3')-gene-proximal regions (referred to as 5' (3')-PLMs) of *A. thaliana* and *Z. mays*, respectively (Figure 2A and Supplementary Table 1). To verify that detected PLMs were not redundant, we tested the inclusion relationship between two PLMs (a k-mer included in a larger k-mer) if they shared 50% of their functional windows and occurred in almost the same gene sets (Jaccard index >= 0.9) (Supplementary Table 2). Only 84 PLM pairs corresponding to 159 5'-PLMs of *Z. mays* had the same PLM-containing gene sets. This meant that a maximum of 79 PLMs could be filtered, i.e. 0.8% of the 5'-PLMs detected in *Z. mays*. Therefore, we considered the redundancy of PLMs to be negligible and retained our original number of PLMs based on the score for all subsequent analyses.

Comparison of the PLM content of the two species revealed that *A. thaliana* and *Z. mays* shared 1,063 5'-PLMs and 1,677 3'-PLMs (Figure 2A). It is worth noting that 98% of these PLMs were located in the 200 bases around the TSS or the TTS. Examination of the preferred position of the PLMs also revealed three visually distinguishable groups within each target region of each species with similar distribution patterns (Figure 3). In the 5'-gene-proximal region, groups 1 (*A. thaliana*:]-450;- 175]; *Z. mays*:]-450;-225]) and 2 (*A. thaliana*:]-60;-25]; *Z. mays*:]-75;-30]) were localized upstream of the TTS, while group 3 (*A. thaliana*:]-25;+10]; *Z. mays*:]-30;+10 bp]) was

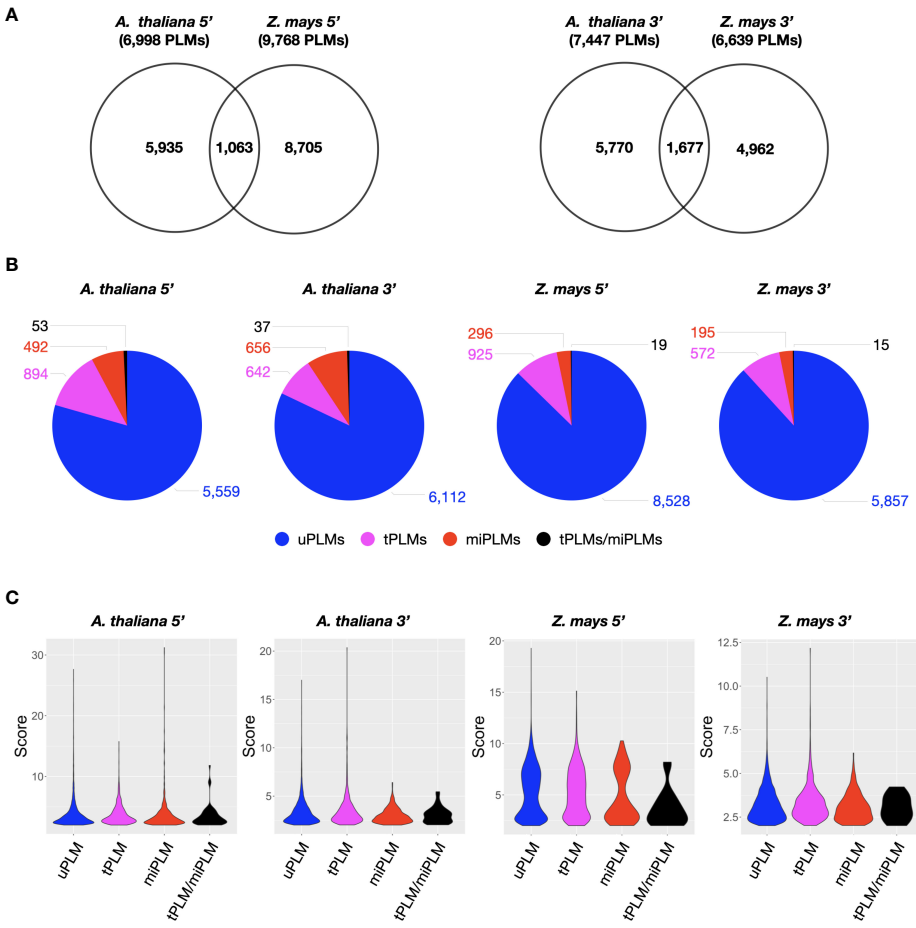


FIGURE 2
Characterization of PLM content in gene-proximal regions of *A. thaliana* and *Z. mays*. **(A)** Venn diagram showing the extent of overlap between 5'- or 3'-PLMs of *A. thaliana* and *Z. mays*. **(B)** Dissection of PLM types identified in the 5'- or 3'-gene-proximal region of *A. thaliana* and Z. mays. **(C)** Violin plot of PLM scores according to the PLM types in the 5'- or 3'-gene-proximal region of *A. thaliana* and *Z. mays*.

localized on the TTS. We also found that 72% of 5'-PLMs in *Z. mays* were localized upstream and downstream of the identified groups, whereas in *A. thaliana* 72% of the 5'-PLMs were localized in groups.

Additionally, each group of 5'-PLMs had specific nucleotide content. Group 1 was composed of A, T, C and G nucleotides in equal proportions in both species. In contrast, group 2 was composed predominantly of A/T (74% and 64% in *A. thaliana* and *Z. mays*, respectively) in agreement with previous observations reporting TATA and TATA-like boxes in this region (Joshi, 1987). In the case of group 3, we found that the GC content of the 5'-PLMs differed between the two species (37% of GC in *A. thaliana vs* 55% in *Z. mays*), in agreement with the report of GC-rich genes in monocot species (Clément et al., 2014; Sundararajan et al., 2016) and recent promoter comparisons using *A. thaliana* and the three cereal species brachypodium, wheat and barley (Peng et al., 2016).

For the 3'-PLMs of both species, we found that groups 2 and 3 consisted predominantly of A/T nucleotides (>70%), whereas group 1 in both species consisted mainly of C/G nucleotides (>60%). We did also observe that 3,877 and 2,130 3'-PLMs detected in the [-40;+10] bp interval relative to the TTS (which corresponds to the end of group 2 and the whole group 3) in *A. thaliana* and *Z. mays*, respectively, showed similarities to the *cis*-elements that guide the CPMC essential for mRNA biogenesis (Bernardes and Menossi, 2020). These 3'-PLMs were A/T-rich (over 78%) consistently with the far upstream element (FUE) and near upstream element (NUE). Furthermore, those localized 10 bases upstream and downstream of the TTS were composed of sequences rich in T (42% in both species) >A (38% in *A. thaliana* and 34% in *Z. mays*) >C (11% in *A. thaliana* and 14% in *Z. mays*) >G (9% in *A. thaliana* and 11% in *Z. mays*), in agreement with the known proportions of nucleotides in the cleavage element (CE) (Bernardes and Menossi, 2020).

10.3389/fpls.2022.976371

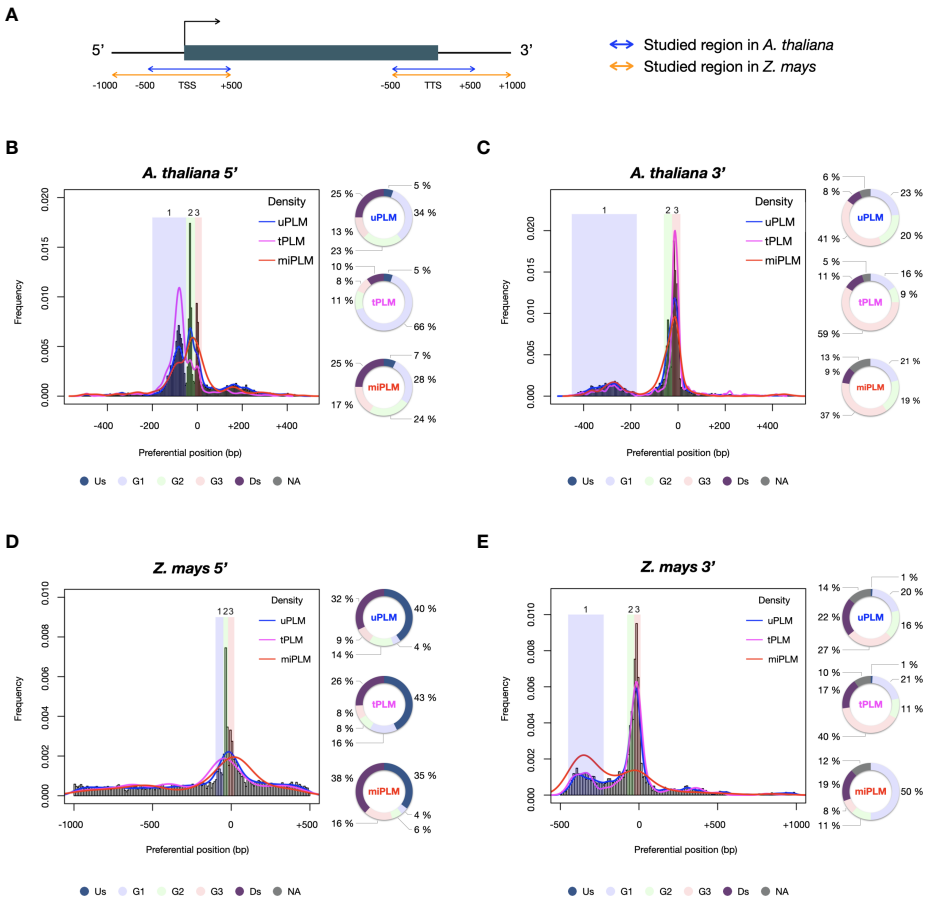


FIGURE 3
PLM frequency and characterization according to their preferential position. **(A)** Schema of the studied regions with respect to the gene in *A. thaliana* and Z. mays. **(B)** 5'-PLMs in *A. thaliana*. Group 1: [-200;-50[; Group 2: [-50;-10[; Group 3: [-10;+20[. **(C)** 3'-PLMs in *A. thaliana*. Group 1]:-450;-175]; Group 2]:-60;-25]; Group 3]:-25;+10]. **(D)** 5'-PLMs in *Z. mays*. Group 1: [-100;-50[; Group 2: [-50;-20[; Group 3: [-20;+20[. **(E)** 3'-PLMs in *Z. mays*. Group 1]:-450;-225]; Group 2]:-75;-30]; Group 3]:-30;+10 bp]. Us: region located upstream the three groups; G1, G2 and G3: groups 1, 2 and 3; Ds: region located downstream the three groups; NA: region located between the groups when they are not juxtaposed.

Identification and positional distribution of TFBS-like PLMs in gene-proximal regions

We anticipated that some of the detected PLMs might be TFBSs. Although the genomes of *A. thaliana* and *Z. mays* are distant, the TFBSs of orthologous TFs are found to be similar in sequence (Tu et al., 2020). Therefore, we decided to use a common resource of plant TFBSs from experimental data to assign our PLMs. This consisted of the JASPAR Plant 2020 database (Fornes et al., 2019) in combination with the top 1% k-mers from ChIP-seq data of 104 maize TFs (Tu et al., 2020). It led to the discovery that 13.5% of the 5'-PLMs (9.1% of the 3'-PLMs) of *A. thaliana* were indeed similar in sequence to known TFBS and were therefore referred to as tPLMs (Figure 1A, Figure 2B and Supplementary Tables 3A, B). In *Z. mays*, 9.6% of the 5'-PLMs and 8.8% of the 3'-PLMs corresponded to tPLMs (Figure 2B and Supplementary Tables 3C, D). To evaluate these tPLM predictions, we used the experimental *A. thaliana* ChIP-seq and DAP-seq data integrated into the ReMap database (Hammal et al., 2022). We found that 61% of 5'-tPLMs and 55% of 3'-tPLMs of *A. thaliana* were covered by experimental peaks for the corresponding TFs, supporting our TFBS assignment of PLMs (Supplementary Tables 3A, B). It is worth noting that in *A. thaliana* 66% of the 5'-tPLMs were localized in group 1, whereas 59% of the 3'-tPLMs were localized in group 3 (Figures 3B, C). In Z. mays, 26% of the 5'-tPLMs were localized upstream and 35% downstream of the identified groups, whereas the 3'-tPLMs followed the same behavior as in *A. thaliana*, with greater localization in group 3. Overall, these results show strong localization of the 5'-tPLMs in the interval between 200 and 50 bp upstream of TSS in agreement with previous observations in *A. thaliana* and *Prunus Persica*

(Yu et al., 2016; Ksouri et al., 2021). In contrast, the 3'-tPLMs mainly localized in the TTS region in both species.

We next investigated how the different TF families were distributed in each proximal region. Among the 47 TF families listed in our reference, 39 and 40 (35 and 37) were susceptible to bind to 5' (3')-tPLMs in *A. thaliana* and *Z. mays*, respectively (Figure 4). We observed that all TF families associated to 3'-PLMs also targeted 5'-PLMs. We also noted that some TF families were detected only with the 5'- or 3'-tPLMs of *A. thaliana* or *Z. mays* (Figure 4A). Using the ReMap data, we determined whether the lack of detection of these TF families was also observed in experimental data from *A. thaliana*. In contrast to predictions, we found that these families indeed bind experimentally to these regions, indicating that their TFBSs have fewer topological constraints. Additionally, we found that all TF families were not similarly distributed in each gene-proximal region. For example, the MYB TFs had tPLMs in all three groups of each region and species studied (Figure 4). In contrast, the Trihelix TFs was only present in group 1 in the 5'-proximal region of *A. thaliana*. Other TF families, such as the G2-like TFs, were likely to target different number of PLM groups according to the region and species considered.

PLMs occur at microRNA binding sites in gene-proximal regions

Previous studies showed that microRNAs (miRNAs) can target transcripts with sequence complementarity (Bartel, 2009), thus inducing their degradation. It was also described in *Brassica* that miRNA methylates the promoter region of *SP11* gene to silence it (Tarutani et al., 2010). Hence, we predicted that 5'- and 3'-PLMs could be associated to miRNA binding sites. Using the plant small RNA target analysis server psRNATarget (Dai et al., 2018)". we found that 7.8% and 3.2% (9.3% and 3.2%) of the 5' (3')-PLMs can be targeted by miRNAs (referred to as 5' (3')-miPLMs) in *A.*

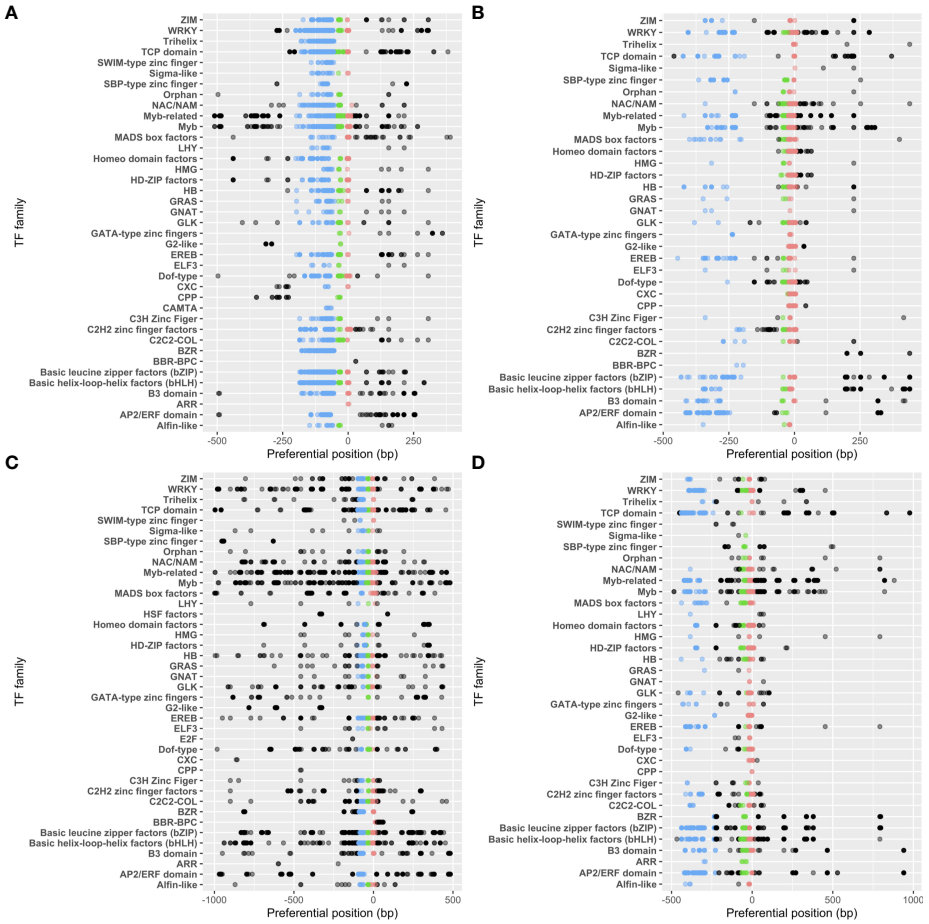


FIGURE 4
tPLM preferential positions per TF family. tPLMs in 5'- **(A)** and 3'- **(B)** gene-proximal regions of *A. thaliana*. tPLMs in 5'- **(C)** and 3'- **(D)** gene proximal regions of *Z. mays*. Blue, green and red points correspond to the tPLMs in groups 1, 2 and 3, respectively. The black points correspond the tPLMs that do not belong to any group. The opacity of the points is relative to the number of tPLMs at that position.

thaliana and *Z. mays*, respectively (Figure 1, Figure 2B and Supplementary Table 4). To assess the quality of the predictions, we performed full-length miRNA alignment at miPLM sites, allowing 0-5 mismatches as described previously for miRNA-target interactions (Ossowski et al., 2008; Axtell, 2013). We found that 51% and 29% of 5'-miPLMs (55% and 41% of 3'-miPLMs) in *A. thaliana* and *Z. mays*, respectively, corresponded to full-length miRNAs (Supplementary Tables 4). We further noticed that 5'-miPLMs had a maximum density downstream of the TSS, which is consistent with the main mode of action of miRNAs, and supports our approach and findings (Figures 3B, D). Surprisingly, more than half of the 5'-miPLMs of *A. thaliana* were located in groups 1 and 2 (Figure 3B), while those of *Z. mays* were overwhelmingly found outside the groups (Figure 3D). We also noticed that 3'-miPLMs were more localized in group 3 than in the other two groups in *A. thaliana* (Figure 3C), while half of them were found in group 1 in *Z. mays* (Figure 3E).

We then investigated which sequence of miPLMs was homologous to that of miRNAs, since the latter are composed of different parts that do not all have the same function (Ossowski et al., 2008). It is known that the 5'-seed region (positions 2-8) of miRNAs is involved in target recognition, and the cleavage site (positions 10-11) is also critical for post-transcriptional regulation. Therefore, we characterized the coverage of PLMs-miRNA homologies (Figure 5). We found that 5'-miPLMs from *A. thaliana* had more frequent homologies with the 5'-seed region, whereas those from *Z. mays* had more frequent homologies with the compensatory 3'-end (Figure 5). For 3'-miPLMs, homologies were more frequent in the center of the miRNA, surrounding the cleavage site in both species, e frequent homologies in the 5' seed region, whereas those belonging to any of the three groups had higher homology frequencies in the bases surrounding the cleavage site in both species, suggesting that miPLMs have distinct functions depending on the region and species considered.

Interestingly, we found that 53 and 19 5'-miPLMs (37 and 15 3'-miPLMs) corresponded to tPLMs in *A. thaliana* and *Z. mays*, respectively (Figure 2B and Supplementary Tables 1A-D). In *A. thaliana*, we noticed that WRKY, Basic helix-loop-helix factors (bHLH) and Basic leucine zipper factors (bZIP) represented the three major TF families identified in the 5'-gene-proximal region, while C2H2 zinc finger factors, Myb-related and HD-ZIP factors were the three major TF families identified in the 3'-gene-proximal region (Supplementary Tables 4E, F). In *Z. mays*, bHLH, bZIP and BZR represented the three major TF families identified in the 5'-gene-proximal region, while bHLH, bZIP and TCP domain were the major TF families in the 3'-gene-proximal region (Supplementary Tables 4G, H).

Unassigned PLMs are putative *cis*-regulatory players

Comparison with resources of TF and miRNA binding sites revealed that more than 79% of the identified PLMs were unassigned PLMs (referred to as uPLMs) (Figures 1, 2B, C and Supplementary Table 1). To determine whether uPLMs with the strongest topological constraints (score > 10) were only core promoter motifs (i.e., motifs bound by the transcription machinery), we evaluated the score of uPLMs and the content of RNA polymerase II binding site. We found that 39.2% and 83.1% of the major 5'-uPLMs detected in *A. thaliana* and *Z. mays*, respectively, were distinct from RNA polymerase II binding sites (11.5% and 6.9% of the 5'-uPLMs detected in *A. thaliana* and *Z. mays*, respectively) (Fornes et al., 2019), suggesting that uPLMs may contain *cis*-regulatory players (Figure 2C and Supplementary Table 5).

In *A. thaliana*, one-third of the 5'-uPLMs were localized in group 1, while more than one-third of the 3'-uPLMs were localized in group 3. In *Z. mays*, the 5'-uPLMs were preferentially localized upstream and downstream of the three groups detected, showing a greater dispersion than that observed in *A. thaliana*. Furthermore, the 3'-uPLMs of *Z. mays* had a more balanced distribution among the different groups and downstream part than those of *A. thaliana*. Interestingly, the density of the 5'-uPLMs in both species was higher in the core promoter region corresponding to group 2 and known to be the locus of many regulatory events (Grosschedl and Birnstiel, 1980; Molina and Grotewold, 2005; Yamamoto et al., 2007; Bernard et al., 2010), confirming the relevance of our hypothesis (Figure 3).

Recently, MNase-defined cistrome-occupancy analysis (MOA-seq) to identify chromatin-accessible regions in developing maize ears led to the identification of 215 small (<30 bp) TF footprints distributed in total across 100,000 non-overlapping binding sites in the genome (Savadel et al., 2021). Given the relatively small size of these footprints and their remarkable clustering within 100 bp proximal to the promoters, we examined them for sequence and position (Figure 1). We found that 85 of these 215 TF footprints significantly matched 203 of our motifs. Considering the position of these motifs (plus or minus 30 bases upstream and downstream of the corresponding PLM functional window), 30% of them covered 79 PLMs (Supplementary Table 9), including 19 tPLMs, 13 miPLMs and 50 uPLMs. Overall, these results support our hypothesis that uPLMs comprise putative *cis*-regulatory players.

10.3389/fpls.2022.976371

uPLMs provide specific functional predictions

To characterize further the uPLMs, we used GO-term and MapMan functional category enrichment analysis to classify them according to the genes in which they occur (Figure 1). In both species, the 5'- and 3'-uPLMs-containing gene sets constituted two highly differentiated populations in terms of their biological processes or MapMan categories relative to the other identified PLM classes, further confirming that they include *cis*-regulatory players (Figure 6 and Supplementary Tables 6, 7).

Comparing 5'- and 3'-uPLMs-containing gene sets revealed specific terms associated with each of the two sets (Supplementary Figure 2 and Table 7). Notably, we observed that "cellular response to ethylene stimulus" was one of the five most enriched GO terms in the 3'-uPLMs-containing gene set of both *A. thaliana* and *Z. mays*. Some of the genes considered are characterized by uPLMs signals in the -450 to -200 bases relative to the TTS. These signals are further supported by the fact that the uPLM sequences are very conserved between the two species (CGTCG and its reverse-complementary CGACG for *A. thaliana*; ACGCCCAC/GGGCGTCC and its reverse-complementary GGACGCCC for *Z. mays*). Terms related to "Cell wall organisation" were also present in the five most enriched MapMan terms for the *A. thaliana*-uPLMs-containing gene set in both regions, although the protein classes identified were different. For example, we found that part of the genes encoding alpha-expansin are characterized by 5'-uPLMs signals localized after the TSS, while part of the genes encoding acyl omega-hydroxylases are characterized by 3'-uPLMs signals localized in group 1 and 2.

As expected, each species had also specific enriched terms (Supplementary Figure 2 and Table 7). For the 5'-uPLMs-gene

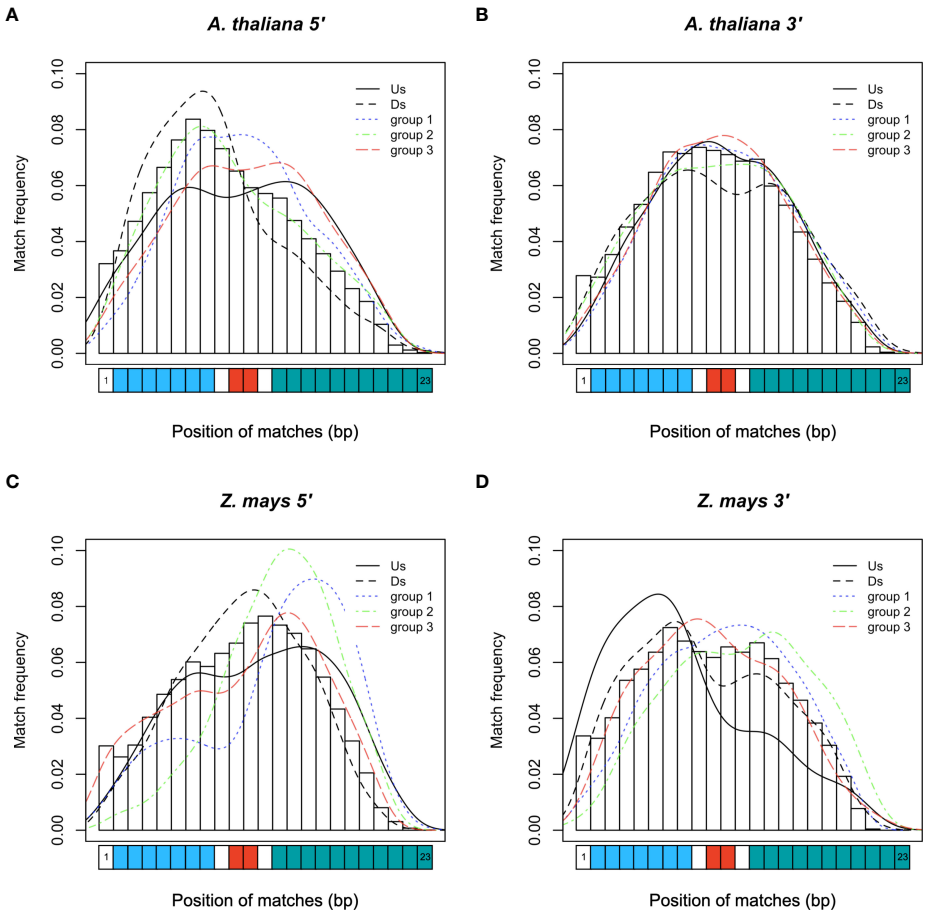


FIGURE 5
Frequency of miRNA bases covered by miPLMs. **(A)** Frequencies for the 5'-miPLMs and the **(B)** 3'-miPLM of *A. thaliana*. **(C)** Frequencies for the 5'-miPLMs and the **(D)** 3'-miPLM of *Z. mays*. The color curves indicate the densities of matched-ribonucleotide positions depending on whether the miPLM that matches belongs to one of the PLM groups (groups 1, 2 or 3) or is located upstream (Us) or downstream (Ds) of these groups. The blue, red and green rectangles on the abscissa represent the bases of the 5'- seed region, the cleavage site and the 3'-compensatory end of the miRNA, respectively.

10.3389/fpls.2022.976371

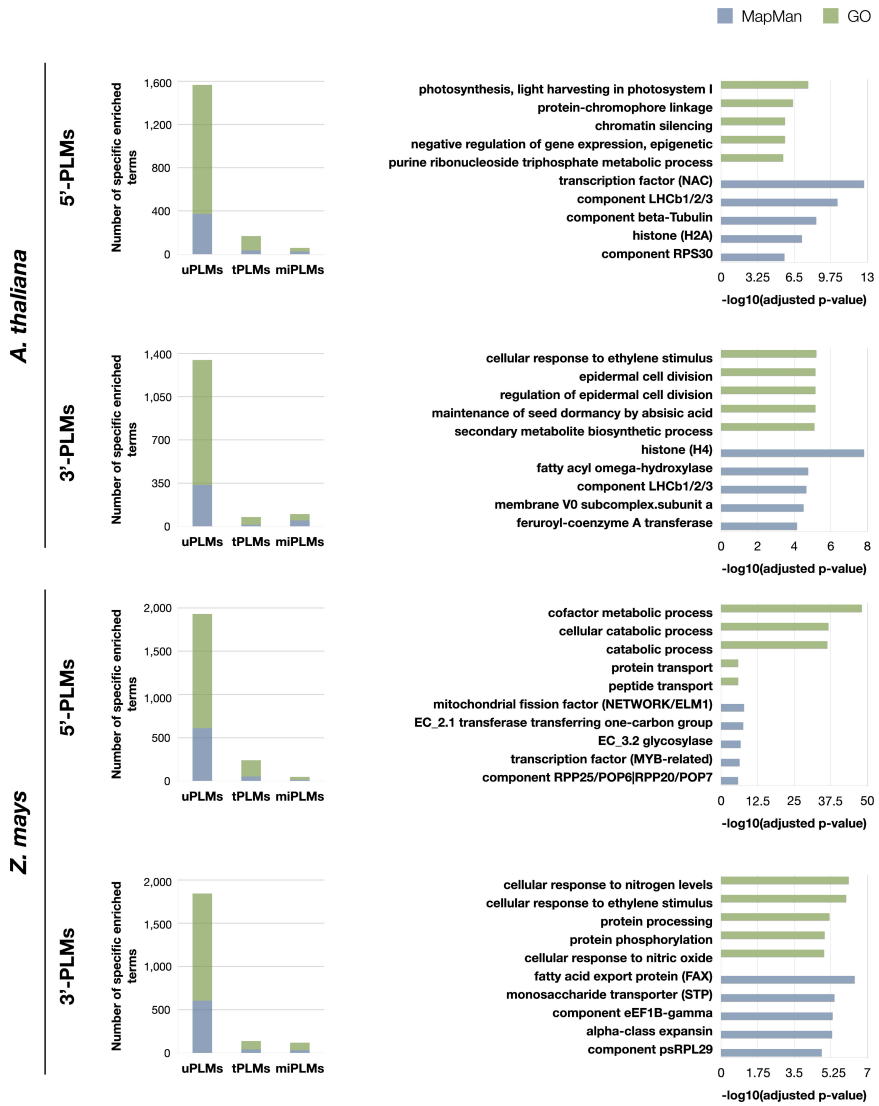


FIGURE 6
GO and MapMan terms enriched specifically for each type of PLMs-containing gene sets. Values for GO and Mapman terms are shown in green and blue, respectively. On the left: histograms of the number of GO and MapMan terms enriched specifically for each type of PLMs-containing gene sets in the two species studied. On the right: the 5 most enriched terms specifically for uPLMs. The bar values of the histograms indicate the -log(adjusted p-values) for each term. Map-Man terms have been truncated at the maximum precision level. The corresponding integer MapMan terms are as follows: transcription factor (NAC): RNA biosynthesis. transcriptional regulation.transcription factor (NAC); component LHCb1/2/3: Photosynthesis.photophosphorylation.photosystem II.LHC-II complex.component LHCb1/2/3; component beta-Tubulin: Cytoskeleton organisation.microtubular network.alpha-beta-Tubulin heterodimer.component beta-Tubulin; histone (H2A): Chromatin organization. histones.histone (H2A); component RPS30: Protein biosynthesis.ribosome biogenesis.small ribosomal subunit (SSU).SSU proteome.component RPS30; histone (H4): Chromatin organisation.histones.histone (H4); fatty acyl omega-hydroxylase: Cell wall organisation.cutin and suberin.cuticular lipid formation.fatty acyl omega-hydroxylase; membrane V0 subcomplex.subunit a: Solute transport.primary active transport.V-type ATPase complex.membrane V0 subcomplex.subunit a; feruroyl-coenzyme A transferase: Cell wall organisation.cutin and suberin.alkyl-hydrocinnamate biosynthesis.feruroyl-coenzyme A transferase; mitochondrial fission factor (NETWORK/ELM1): Cell cycle organisation.organelle division.mitochondrion and peroxisome division.mitochondrial fission factor (NETWORK/ELM1); EC_2.1 transferase transferring one-carbon group: Enzyme classification.EC_2 transferases.EC_2.1 transferase transferring one-carbon group; EC_3.2 glycosylase: Enzyme classification.EC_3 hydrolases.EC_3.2 glycosylase; transcription factor (MYB-related): RNA biosynthesis.transcriptional regulation.MYB transcription factor superfamily.transcription factor (MYB-related); component RPP25/POP6|RPP20/POP7: RNA processing.ribonuclease activities.RNA-dependent RNase P complex.component RPP25/POP6|RPP20/POP7; fatty acid export protein (FAX): Lipid metabolism.lipid trafficking.fatty acid export protein (FAX); monosaccharide transporter (STP): Solute transport.carrier-mediated transport.MFS superfamily. SP family.monosaccharide transporter (STP); component eEF1B-gamma: Protein biosynthesis.translation elongation.eEF1 aminoacyl-tRNA binding factor activity.eEF1B eEF1A-GDP-recycling complex.component eEF1B-gamma; alpha-class expansin: Cell wall organisation.cell wall proteins.expansin activities.alpha-class expansin; component psRPL29 Protein biosynthesis.organelle machinery.plastidial ribosome.large ribosomal subunit proteome.component psRPL29.

set, these terms were mainly related to "cell killing" in *A. thaliana*, while they were associated with "transposition" and antibiotic metabolic/catabolic processes in *Z. mays* (Supplementary Tables 8G, H). For the 3'-uPLMs-gene set, specific enriched terms were once again related to "cell killing" in *A. thaliana*, while they were mainly related to "translation" in Z. mays. Taken together, these findings reveal that uPLMs provide functional predictions that differ from those derived from tPLMs and miPLMs (Figure 6 and Supplementary Tables 6-8).

Biological processes associated to uncharacterized genes through integration of PLM information

Taking into account the contribution of PLMs, we thought to use them to infer gene regulatory networks and go further and deeper into the characterization of some gene families (Figure 1). We focused on a poorly characterized, *Z. mays*-specific gene family (referred to as HOM04M002476 by PLAZA) defined only by the GO term "transposition" (Supplementary Table 8G). This gene family consists of 65 genes, 64 of which were considered in the detection of 5'-PLMs (Supplementary Table 11A). Using the 5'-tPLMs detected for all these 64 genes, we investigated the TF-target gene relationships (Figure 7A). A total of 545 tPLMs were associated with 416 TFs belonging to 37 distinct TF families. Among them, AP2/ERF domain, Myb-related and WRKY were the three most abundant TF families (Supplementary Table 11).

Clustering based on latent block model (LBM) revealed three modules of target genes referred to as G1, G2 and G3 with 7, 3 and 54 members, respectively (Figure 7A and Supplementary Table 11). It also revealed three modules of TFs referred to as TF1, TF2 and TF3 with 6, 380 and 30 TFs belonging to 2, 29 and 6 TF families, respectively (Figures 7A, B and Supplementary Table 11). We found that genes belonging to module G1 were regulated by TFs from modules TF1 (2/2 families), TF2 (16/29 families) and TF3 (1/6 family). It is worth noting that the SWIM-type zinc finger TF family of module TF1 was specific to genes from G1 module. Genes belonging to module G2 were also regulated by TFs belonging to all three modules, including the BZR TF family of module TF1, all TF families of module TF2, and the HSF factors and G2-like TF families of module TF3. In contrast, genes belonging to module G3 were only regulated by TFs from modules TF2 and TF3. Furthermore, the CXC, CPP and SBP-type zinc finger TF families of module TF3 covered specifically genes from G3 module.

To elucidate the potential involvement of this gene regulatory network in biological processes, we conducted GO and MapMan enrichment analysis of the TF modules. As expected, terms enriched in a common way in all three modules were related to the regulation of transcription (Supplementary Table 12). Additionally, in module TF1, the most specifically enriched terms were related to brassinosteroid responses (Figure 7A and Supplementary Tables 12A, B). In module TF2, except for terms related to transcriptional regulation, the most enriched terms was related to "response to stimulus" (Figure 7A; Supplementary Tables 12C, D). Finally, in module TF3 the most enriched terms were mainly related to methylation, response to nutrient levels and primary root development (Figure 7A and Supplementary Tables 12E, F). Our computational approach therefore paves the way for pinpointing the function of this gene family in these distinct processes in future follow-up studies.

Discussion

Understanding gene transcriptional regulation requires understanding where regulatory factors bind genomics DNA. Although several efforts have recently been undertaken to characterize TFBSs, the identification of high resolution *cis*-regulatory sequences at the genome-wide scale remains an arduous challenge. Hence, we attempted to reveal the whole PLM landscape by using genome-wide *de novo* PLM detection to systematically profile proximal putative *cis*-regulatory sequences. The three PLM group structure revealed in *A. thaliana* and *Z. mays* with distantly related genomes echoes and enriches established knowledge of the 5'-gene-proximal region (Figure 8). Omitting potential annotation errors, we found that the localization of these motifs, including the core promoter, was less constrained *in Z. mays* compared to *A. thaliana*, as recently reported for TATA-boxes at varying distances from TSS in *Z. mays* (Jores et al., 2021). Similarly, we observed that the dispersion of tPLMs remained more important in *Z. mays* than in *A. thaliana*, indicating that putative TFBSs have also a less constrained preferential localization in *Z. mays* than in *A. thaliana*. Overall, these data suggest that the 5'-proximal genomic context may be less constrained in *Z. mays* than in *A. thaliana*. This could be related to the richness of the *Z. mays* genome in transposable elements (TEs) compared with that of *A. thaliana* (Stitzer et al., 2021). TEs are known to be involved in regulating gene expression by introducing TFBSs into gene-proximal regions (Quesneville, 2020). Therefore, the study of TE-derived PLMs is an important perspective to obtain more information about PLMs and to better characterize the associated TEs, and deserves to be the subject of future studies.

Our finding of conserved PLMs between *A. thaliana* and *Z. mays* suggests that the closer we get to the genes, the more the context, including *cis*-regulatory elements (here given by tPLMs), are conserved between species. Notably, we have shown that this context appears to be more conserved in the 3'-gene-proximal region than in the 5'-gene proximal region: 14% of 3'-PLMs shared between the two species compared to 7% of 5'-PLMs (Figure 2A). This emphasize the importance of the

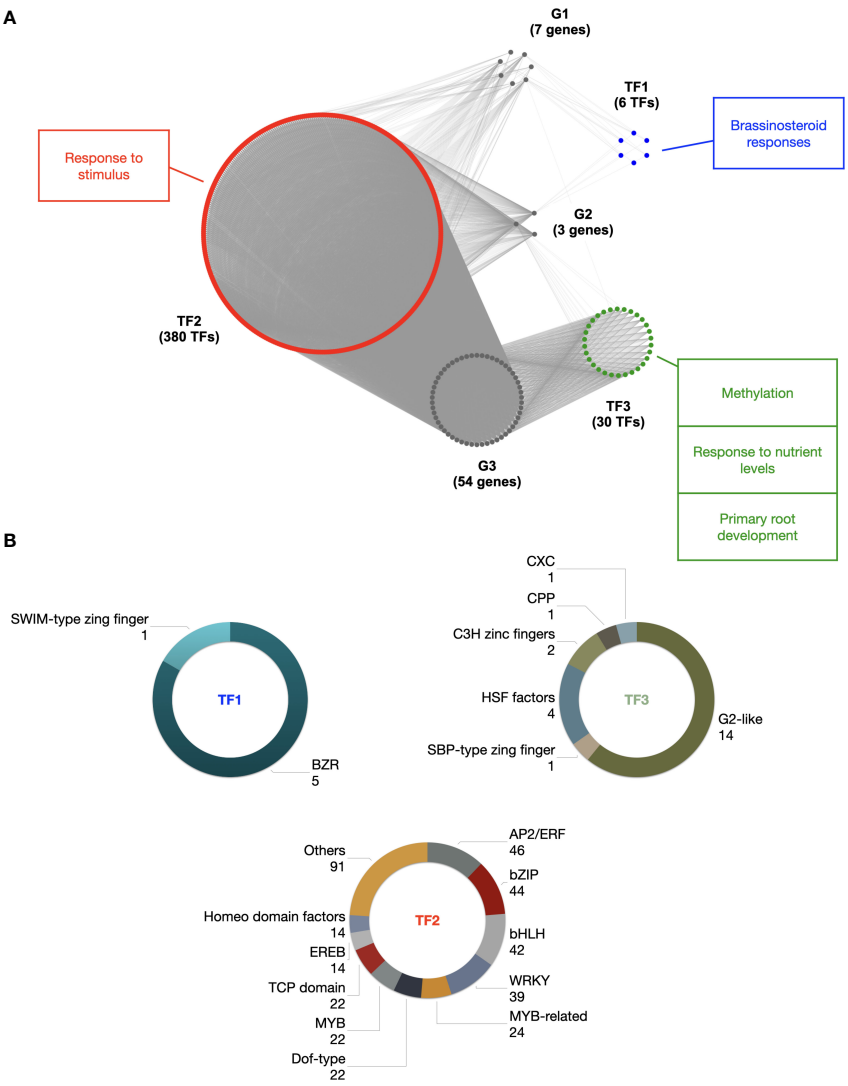


FIGURE 7
Gene regulatory network of the HOM04M002476 gene family of *Z. mays*. **(A)** Topological structure of the gene regulatory network of the HOM04M002476 gene family. G1, G2 and G3 are the three HOM04M002476 gene modules (grey) identified by the LBM procedure. TF1 (red), TF2 (green) and TF3 (blue) are the three TF modules identified by the LBM procedure potentially involved in the regulation of HOM04M002476 genes. GO enrichments of TF modules are indicated with corresponding colors. **(B)** TF families associated to the three TF modules.

3'-gene-proximal region in genomic structure. Despite its key role in gene expression, the 3'-gene-proximal region remains poorly studied in plants (Srivastava et al., 2018; Mayr, 2019; Bernardes and Menossi, 2020). Because the density maxima observed for tPLMs and uPLMs was reached in the *cis*-elements that guide the CPMC and overlapped with groups 2 and 3, it is quite possible that the 3'-PLMs detected in these portions of DNA sequence constitute a catalog of NUE, FUE and CE (Figure 8). In support of this hypothesis, we observed that the AATAAA motif (and its complementary reverse), which is the key site involved in polyadenylation and is extremely conserved in mammals and somewhat less in plants, was located between 10 and 20 bases upstream of the TTS. Furthermore, the nucleotide percentages of 3'- PLMs detected in these regulatory regions are consistent with known proportions of nucleotides in these *cis*-elements. Together, these data support the idea that 3'-PLMs may constitute an accurate catalog of CPMC-guiding *cis*-elements. In this respect, the presence of 3'-tPLMs in this catalog (around 11%) and more generally in the whole region, opens interesting mechanistic perspectives on the role of TFs. First, they may act as activators or repressors of the transcriptional machinery (Figure 8). Second, by binding to tPLMs located in the FUE/NUE and CE regions, they could impact pre-mRNAs length and thus mRNA stability by

influencing the choice of an alternative polyadenylation site at the end of transcripts (Srivastava et al., 2018; Mayr, 2019; Bernardes and Menossi, 2020).

The retained proportion observed between uPLMs (up to 25%) and tPLMs (up to 30%), also suggests that uPLMs could constitute a context that needs to be conserved. In this regard, the genomic significance of uPLMs is already supported by experimental maize ear MOA-seq data, suggesting that some of the identified uPLMs are potential TF footprints. It will be important to validate these uPLMs by functional assays to determine what proportion of them are indeed proximal *cis*-regulatory players. In addition, our finding raises the question of what mechanisms underlie the presence of uPLMs. First, as mentioned earlier, some uPLMs may be players in the core promoter or polyadenylation process (Figure 8). Second, some uPLMs may be non-annotated binding sites. Indeed, we showed that the 104 maize TF ChIP-seq data (Tu et al., 2020) contributed 16% and 33% more tPLMs for *A. thaliana* and *Z. mays*, respectively, compared to the JASPAR Plant 2020 database that was updated prior to the release of these ChIP-seq data. Similar to the post-transcriptional regulation by miRNAs, RNA-binding proteins (Lee and Kang, 2016; Cho et al., 2019) are major players that can potentially bind PLMs at the transcriptional level. Consequently, there is no doubt that future resources will supplement the assignment of uPLMs. Finally, uPLMs may be motifs that are not directly bound by TFs but that play a crucial role in the correct binding of these regulators to neighboring TFBSs (Figure 8) (Stringham et al., 2013; Crocker et al., 2015; Stampfel et al., 2015). This concept of "flanking sequence context" appears extremely relevant because of the nature of PLMs, which are constrained motifs at a distance from genes. This idea also raises many questions about the existence and role of tPLMs/uPLMs associations, with tPLMs bound by TFs and uPLMs serving as essential context sequences for the formation of the DNA-TF complex. Additional analyses and integration with other *in vivo* information will be key to advance functional tests needed to ascertain the relative importance of tPLMs and uPLMs as *cis*-regulatory elements controlling gene expression. Meanwhile, our results have broader implications for future characterization of unannotated genes in plants.

In summary, the implementation of the genome-wide and *de novo* PLMdetect method has demonstrated the richness of the gene-proximal regions and the interest in their further characterization. In particular, this work has highlighted the importance of the 3'-gene-proximal region as a major source of new knowledge and great interest for future studies.

Methods

Genomic datasets

TAIR10 (Lamesch et al., 2012) and B73v4.39 (Jiao et al., 2017) genomes and their annotations were considered to extract the 5'- and 3'-gene-proximal sequences of *A. thaliana* and *Z. mays* genes, respectively.

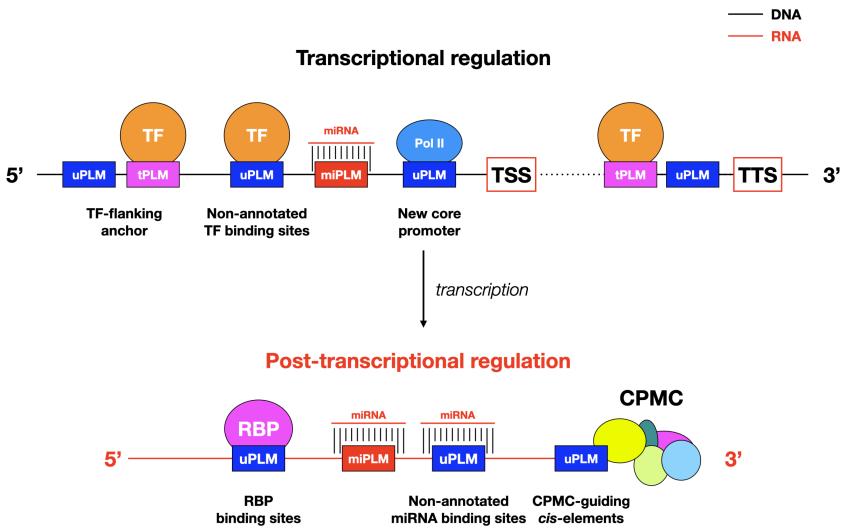


FIGURE 8
Models of PLM-mediated transcriptional and post-translational controls in plants. The models integrate only the results presented here. Possible interactions between the illustrated components and the diverse array of other intermediates such as enhancers and long-distance chromatin interaction sites, remain to be investigated. CPMC, cleavage and polyadenylation molecular complex; Pol II, RNA polymerase II; RPB: RNA binding proteins.

Preparation of the gene-proximal sequence files

For the 5'-gene-proximal region, annotation of the TSS was ensured by filtering genes without a 5'-UTR region (in GFF3/GTF file). Genes on reverse strand were reverse-complemented to analyze all sequences in the same orientation. Extracted sequences corresponded to the intervals [-1000;+500] and [-1500;+500] bp relative to the TSS for *A. thaliana* and *Z. mays*, respectively. In total, 19,736 and 25,848 genes were analyzed for *A. thaliana* and *Z. mays*, respectively. For the 3'-gene-proximal region, similarly to what has been done for the 5'-gene-proximal region sequences, annotation of the TTS was ensured by filtering genes without a 3'-UTR region annotated. To standardize the PLM detection step, genes on forward strand were reverse-complemented. Extracted sequences were [-500; +1000] and [-500;+1500] bp with respect to the TTS for *A. thaliana* and *Z. mays*, respectively. Taking in consideration only annotated 3'-UTR, 20,573 and 25,199 genes were processed for *A. thaliana* and *Z. mays*, respectively.

Preparation of the motif file

Every non-polymorphic DNA 4-mers to 8-mers was generated representing 87,296 motifs. Among these motifs, 256, 1,024, 4,096, 16,384, and 65,536 had a length of 4, 5, 6, 7 and 8 bp, respectively.

Processing of potential PLM redundancy

To check PLM redundancy, we calculated the Jaccard index of each pair of PLMs for each PLM containing-gene set with an inclusion link. This index was obtained by dividing the intersection of the two lists by their union. A Jaccard index of 0.9 indicates an almost perfect match between the gene lists. This index was also calculated on the functional window of each pair of PLMs to quantify their overlap. We set the threshold for the functional window Jaccard index at 0.5.

TFBS and microRNA resources and assignment

TFBSs (676 total) were extracted from JASPAR Plant 2020 (Fornes et al., 2019) and ChIP-seq of 104 maize leaf TFs (Tu et al., 2020). MicroRNAs (miRNAs) were obtained from psRNATarget (Dai et al., 2018) and only those from *A. thaliana* (427 total) and *Z. mays* (321 total) were kept.

We first assigned TFBSs to PLMs for both species in each region separately using the TOMTOM web tool (Gupta et al., 2007). Euclidean distance was next used as a comparison function with a q-value threshold at 0.05 and the complete scoring option deselected. PLMs were also compared to the top 1% of k-mers of the 104 maize TFs (Tu et al., 2020) by considering only exact matches. Because miRNAs regulate genes by sequence complementarity (Bartel, 2009) and are molecules from 19 to 22 nt, we only considered our 8 bp PLMs for this comparison (the 20 bp size by which miRNAs regulate gene expression was not tested due to computational time constraints). If a PLM was exactly found in a miRNA, it was assigned as a miPLM.

Functional annotation

Functional annotation of genes from both species was based on MapMan X4 (Thimm et al., 2004) and Gene Ontology (GO) from PLAZA 4.5 (Van Bel et al., 2018). Functional enrichment analysis of genes containing an identified PLM (Supplementary Datasets) was performed by comparing the relative occurrence of each term to its relative occurrence in a reference list for each region and species using a hypergeometric test with the R function phyper. These reference lists consisted of all genes considered for PLM detection in each gene-proximal region in both species as described in the 'Preparation of the gene-proximal sequence files' of the Methods section. *P*-values were adjusted by the Benjamini-Hochberg (BH) procedure to control the False Discovery Rate (FDR). An enriched term had its adjusted *P*-value lower than 0.05.

Comparative analysis of PLMs and MOA-seq motifs

Z. mays 5'-PLMs were compared to published MOA-seq motifs (Savadel et al., 2021) using TOMTOM with the same parameters as described for TFBSs assignment and according to the following two criteria: (1) both sequence types had to be identical (TOMTOM q-value <0.05) and (2) the position of the MOA-seq motif had to be within the functional window of the PLM extended by 30 bases upstream and downstream.

Inference and topology analysis of the HOM04M002476 gene family regulatory network

A matrix of dimension 37x64 was generated with TF families in rows and HOM04M002476 genes in columns. Links between TF families and target genes were established when tPLMs, and thus associated TFs, were identified for a given target gene. These links were indicated by ones in the matrix. A zero indicated the absence of a tPLM associated with the TF family in the target

gene. Gene and TF modules were obtained using LBM with the R package blockmodels (Leger, 2016).

Functional enrichment of each TF module was performed by comparing the relative occurrence of each term to its relative occurrence in the list of genes encoding TFs in *A. thaliana* (2,208 genes) and *Z. mays* (2,164 genes) using a hypergeometric test with the R phyper function. *P*-values were adjusted by the BH procedure to control the FDR. An enriched term had its adjusted *P*-value lower than 0.05.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding authors. Code used to generate the data are available at https://forgemia.inra.fr/GNet/plmdetect/plmdetect_tool.

Author contributions

VB, M-LM and SC conceived the project. M-LM and SC designed and supervised the study. JR generated, analyzed and interpreted the data. CG formatted the functional annotation files. JR, M-LM and SC wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Plant2Pro® Carnot Institute in the frame of the PLMViewer program. Plant2Pro® is supported by ANR (agreement #19 CARN 0024 01). The IPS2 and IJPB laboratories benefit from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007).

Acknowledgments

We thank all members of the 'Genomic Networks' (GNet) and 'Biomass Quality and Interactions with Drought' (QUALIBIOSEC) teams past and present. We also thank Hank W. Bass for sharing some works prior to publication, Hervé Vaucheret for very helpful discussions, and two reviewers for comments that improved the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.976371/full#supplementary-material

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182, 145–161.e23. doi: 10.1016/j.cell.2020.05.021

Axtell, M. J. (2013). Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.* 64, 137–159. doi: 10.1146/annurev-arplant-050312-120043

Azodi, C. B., Lloyd, J. P., and Shiu, S.-H. (2020). The cis-regulatory codes of response to combined heat and drought stress in arabidopsis thaliana. *NAR Genomics Bioinf.* 2:lqaa049. doi: 10.1093/nargab/lqaa049

Bartel, D. P. (2009). MicroRNA target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

Bernard, V., Brunaud, V., and Lecharny, A. (2010). TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics* 11, 166. doi: 10.1186/1471-2164-11-166

Bernardes, W. S., and Menossi, M. (2020). Plant 3' regulatory regions from mRNA-encoding genes and their uses to modulate expression. *Front. Plant Sci.* 11, 1252. doi: 10.3389/fpls.2020.01252

Bernard, V., Lecharny, A., and Brunaud, V. (2010). Improved detection of motifs with preferential location in promoters. *Genome* 53, 739–752. doi: 10.1139/G10-042

Bueso, E., Muñoz-Bertomeu, J., Campos, F., Brunaud, V., Martínez, L., Sayas, E, et al. (2014). ARABIDOPSIS THALIANA HOMEOBOX25 uncovers a role for gibberellins in seed Longevity1[C][W]. *Plant Physiol.* 164, 999–1010. doi: 10.1104/pp.113.232223

Cho, H., Cho, H. S., and Hwang, I. (2019). Emerging roles of RNA-binding proteins in plant development. *Curr. Opin. Plant Biol.* 51, 51–57. doi: 10.1016/j.pbi.2019.03.016

Clément, Y., Fustier, M.-A., Nabholz, B., and Glémin, S. (2014). The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biol. Evol.* 7, 336–348. doi: 10.1093/gbe/evu278

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S, et al. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* 160, 191–203. doi: 10.1016/j.cell.2014.11.041

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 46, W49–W54. doi: 10.1093/nar/gky316

Fagny, M., Kuijjer, M. L., Stam, M., Joets, J., Turc, O., Rozière, J., et al. (2021). Identification of key tissue-specific, biological processes by integrating enhancer information in maize gene regulatory networks. *Front. Genet.* 11, 606285. doi: 10.3389/fgene.2020.606285

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., et al. (2019). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48, D87–D92. doi: 10.1093/nar/gkz1001

Frei dit Frey, N., et al. (2014). Functional analysis of arabidopsis immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences. *Genome Biol.* 15, R87. doi: 10.1186/gb-2014-15-6-r87

Grosschedl, R., and Birnstiel, M. L. (1980). Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants *in vivo*. *PNAS* 77, 1432–1436. doi: 10.1073/pnas.77.3.1432

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol.* 8, R24. doi: 10.1186/gb-2007-8-2-r24

Hammal, F., de Langen, P., Bergon, A., Lopez, F., and Ballester, B. (2022). ReMap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* 50, D316–D325. doi: 10.1093/nar/gkab996

Jiao, Y., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546, 524–527. doi: 10.1038/nature22971

Jores, T., et al. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* 7, 842–855. doi: 10.1038/s41477-021-00932-y

Joshi, C. P. (1987). An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucleic Acids Res.* 15, 6643–6653. doi: 10.1093/nar/15.16.6643

Ksouri, N., Castro-Mondragón, J. A., Montardit-Tarda, F., van Helden, J., Contreras-Moreira, B., and Gogorcena, Y. (2021). Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. *Plant Physiol.* 185, 1242–1258. doi: 10.1093/plphys/kiaa091

Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., et al. (2019). Building transcription factor binding site models to understand gene regulation in plants. *Mol. Plant* 12, 743–763. doi: 10.1016/j.molp.2018.10.010

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210. doi: 10.1093/nar/gkr1090

Lee, K., and Kang, H. (2016). Emerging roles of RNA-binding proteins in plant growth, development, and stress responses. *Mol. Cells* 39, 179–185. doi: 10.14348/molcells.2016.2359

Leger, J.-B. (2016). Blockmodels: A r-package for estimating in latent block model anpd stochastic block model, with various probability functions, with or without covariates. *arXiv:1602.07587 [stat]*. doi: 10.48550/arXiv.1602.07587

Li, X., Zhu, C., Yeh, C.-T., Wu, W., Takacs, E. M., Petsch, K. A., et al. (2012). Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.* 22, 2436–2444. doi: 10.1101/gr.140277.112

Liu, S., Li, C., Wang, H., Wang, S., Yang, S., Liu, X., et al. (2020). Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biol.* 21, 163. doi: 10.1186/s13059-020-02069-1

Martínez, F., Arif, A., Nebauer, S. G., Bueso, E., Ali, R., Montesinos, C., et al. (2015). A fungal transcription factor gene is expressed in plants from its own promoter and improves drought tolerance. *Planta* 242, 39–52. doi: 10.1007/s00425-015-2285-5

Mayr, C. (2019). What are 3′ UTRs doing? *Cold Spring Harb. Perspect. Biol.* 11, a034728. doi: 10.1101/cshperspect.a034728

Molina, C., and Grotewold, E. (2005). Genome wide analysis of arabidopsis core promoters. *BMC Genomics* 6, 25. doi: 10.1186/1471-2164-6-25

Ossowski, S., Schwab, R., and Weigel, D. (2008). Gene silencing in plants using artificial microRNAs and other small RNAs: Engineering small RNA-mediated gene silencing. *Plant J.* 53, 674–690. doi: 10.1111/j.1365-313X.2007.03328.x

Peng, F. Y., Hu, Z., and Yang, R.-C. (2016). Bioinformatic prediction of transcription factor binding sites at promoter regions of genes for photoperiod and vernalization responses in model and temperate cereal plants. *BMC Genomics* 17, 573. doi: 10.1186/s12864-016-2916-7

Quesneville, H. (2020). Twenty years of transposable element analysis in the arabidopsis thaliana genome. *Mobile DNA* 11, 28. doi: 10.1186/s13100-020-00223-x

Savadel, S. D., Hartwig, T., Turpin, Z. M., Vera, D. L., Lung, P.-Y., Sui, X., et al. (2021). The native cistrome and sequence motif families of the maize ear. *PloS Genet.* 17, e1009689. doi: 10.1371/journal.pgen.1009689

Schmitz, R. J., Grotewold, E., and Stam, M. (2021). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *Plant Cell* 34, 718–741. doi: 10.1093/plcell/koab281

Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z., and Zhu, J.-K. (2018). UTR dependent control of gene expression in plants. *Trends Plant Sci.* 23, 248–259. doi: 10.1016/j.tplants.2017.11.003

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* 528, 147–151. doi: 10.1038/nature15545

Stitzer, M. C., Anderson, S. N., Springer, N. M., and Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PloS Genet.* 17, e1009768. doi: 10.1371/journal.pgen.1009768

Stringham, J. L., Brown, A. S., Drewell, R. A., and Dresch, J. M. (2013). Flanking sequence context-dependent transcription factor binding in early drosophila development. *BMC Bioinf.* 14, 298. doi: 10.1186/1471-2105-14-298

Sundararajan, A., Dukowic-Schulze, S., Kwicklis, M., Engstrom, K., Garcia, N., Oviedo, O. J., et al. (2016). Gene evolutionary trajectories and GC patterns driven by recombination in zea mays. *Front. Plant Sci.* 7, 1433. doi: 10.3389/fpls.2016.01433

Tarutani, Y., Shiba, H., Iwano, M., Kakizaki, T., Suzuki, G., Watanabe, M., et al. (2010). Trans-acting small RNA determines dominance relationships in brassica self-incompatibility. *Nature* 466, 983–986. doi: 10.1038/nature09308

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., et al. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939. doi: 10.1111/j.1365-313X.2004.02016.x

Tu, X., Mejía-Guerra, M. K., Franco, J. A.V., Tzeng, D., Chu, P.-Y., Shen, W., et al. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat. Commun.* 11, 1–13. doi: 10.1038/s41467-020-18832-8

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., et al. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* 46, D1190–D1196. doi: 10.1093/nar/gkx1002

Wallace, J. G., Bradbury, P. J., Zhang, N., Gibon, Y., Stitt, M., and Buckler, E. S. (2014). Association mapping across numerous traits reveals patterns of functional variation in maize. *PloS Genet.* 10, e1004845. doi: 10.1371/journal.pgen.1004845

Wang, Y., Fairley, J. A., and Roberts, S. G. E. (2010). Phosphorylation of TFIIB links transcription initiation and termination. *Curr. Biol.* 20, 548–553. doi: 10.1016/j.cub.2010.01.052

Waters, A. J., Makarevitch, I., Noshay, J., Burghardt, L. T., Hirsch, C. N., Hirsch, C. D., et al. (2017). Natural variation for gene expression responses to abiotic stress in maize. *Plant J.* 89, 706–717. doi: 10.1111/tpj.13414

Yamamoto, Y. Y., Ichida, H., Abe, T., Suzuki, Y., Sugano, S., and Obokata, J. (2007). Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res.* 35, 6219–6226. doi: 10.1093/nar/gkm685

Yu, C.-P., Lin, J.-J., and Li, W.-H. (2016). Positional distribution of transcription factor binding sites in arabidopsis thaliana. *Sci. Rep.* 6, 25164. doi: 10.1038/srep25164

Zemlyanskaya, E. V., Dolgikh, V. A., Levitsky, V. G., and Mironova, V. (2021). Transcriptional regulation in plants: Using omics data to crack the cis-regulatory code. *Curr. Opin. Plant Biol.* 63, 102058. doi: 10.1016/j.pbi.2021.102058

Zhou, P., Enders, T. A., Myers, Z. A., Magnusson, E., Crisp, P. A., Noshay, J. M., et al. (2022). Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *Plant Cell* 34, 514–534. doi: 10.1093/plcell/koab267

3 - Recherche de PLM chez 20 espèces de plantes à fleurs

3.1 . Extension de la méthode PLMdetect à 18 autres espèces de plantes à fleurs et développement de la base de données Plant-PLMview

Le travail mené au cours de ce chapitre visait à étendre les analyses réalisées chez *A. thaliana* et *Z. mays* (Chapitre 2, Rozière *et al.* (2022b)) à 18 autres espèces de plantes à fleurs afin de déterminer dans quelle mesure les PLM sont conservés au sein de cette famille. Pour se faire, il a été nécessaire de (1) développer un pipeline automatique permettant d'extraire l'ensemble des régions proximales des 20 espèces d'angiospermes sélectionnées et (2) de réaliser la détection des PLM *de novo* pour chacune d'entre elles.

La développement du pipeline automatique a fait l'objet du stage de L2 d'Anne Duveau (rapport en annexe A) que j'ai co-encadré avec Marie-Laure Martin. Le travail d'Anne Duveau a tout d'abord permis de développer un pipeline pour extraire les régions proximales des gènes de chaque espèce. Ce pipeline permet de moduler les bornes d'extraction et conserve uniquement les régions proximales des gènes pour lesquelles les régions transcrites non traduites (UTR) de la région considérée sont annotées. Ce dernier critère permet d'assurer la position du TSS et du TTS et ainsi une détection correcte des PLM dans chaque région proximale. Une fois l'extraction des régions proximales réalisée, le pipeline permet de réaliser la détection *de novo* des PLM.

Ce travail a permis de mettre en évidence que dès lors que l'on passe à une échelle multi-espèces, la quantité de données à traiter et d'informations générées est démultipliée. Pour valoriser l'ensemble de ces données, j'ai développé la base de données Plant-PLMview en collaboration étroite avec Franck Samson (LaMME), Cécile Guichard (IPS2), Margot Correa (IPS2) et Véronique Brunaud (IPS2). Les deux mots d'ordre qui ont guidé ce développement étaient (1) d'offrir la possibilité d'utiliser simplement la méthode PLMdetect et (2) d'interpréter les résultats aisément en visualisant des modules de PLM impliqués dans la co-régulation d'un groupe de gènes d'intérêt. Au final, cette base met à la disposition de l'ensemble de la communauté scientifique la méthode PLMdetect pour les 20 espèces sélectionnées. Elle constitue donc un outil pour explorer et caractériser les régions à proxmités des gènes chez les plantes.

Le développement de Plant-PLMview est présenté ci-après sous la forme d'un article dont je suis le premier auteur et qui sera prochainement soumis pour évaluation par les pairs.

Plant-PLMview: a database for identifying *cis*-regulatory sequences with preferential positions in gene-proximal regions of plants

Julien Rozière[1,2,3], Franck Samson[4], Cécile Guichard[1,2], Margot Correa[1,2,4], Sylvie Coursol[3], Marie-Laure Martin[1,2,5], and Véronique Brunaud[1,2,✉]

[1]Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
[2]Université de Paris Cité, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
[3]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France
[4]Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry-Courcouronnes, France
[5]Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris-Saclay, 91120, Palaiseau, France

Abstract

Background. **Establishing relationships between transcription factors and target genes is essential for understanding the mechanisms regulating gene expression, which play a fundamental role in plant adaptation to the local environment. Despite the importance of this research area and the tremendous progress in sequencing methods such as ChIP-seq and DAP-seq, we are still far from a complete reconstruction of the *cis*-regulatory landscape. Only a small number of transcription factors can be evaluated by experimental data to identify their *cis*-regulatory binding sites. This highlights the role that *in silico* approaches can play to complement experimental data.**

Results. **We developed Plant-PLMview, a web-accessible database for detecting preferentially located *cis*-regulatory sequences in the gene-proximal regions of 20 plant species. Users of Plant-PLMview can (i) access the proximal regions of controlled genes from 20 plant species, (ii) query their own DNA motifs or access 840 *cis*-regulatory sequences from various plant resources, and (iii) use a tool called PLMdetect to search for preferentially located DNA motifs in the gene-proximal regions of a list of genes. Results are displayed via a web interface with a list of DNA motifs preferentially located in a region near the start or end of genes, the distribution of these motifs and associated annotations. In addition, a graphical map of the preferential locations of the motifs in the 5' and 3'-proximal regions of the genes makes it possible to have an overview of all motifs and genes examined.**

Conclusion. **Plant-PLMview provides the opportunity to explore the proximal landscape of gene regulation in 20 plant genomes. The originality of the database lies in its ease of use thanks to the curated data (*cis*-regulatory sequences and proximal regions of genes) and the possibility to search through PLMdetect in the 5'-gene-proximal region, but also in the 3'-gene-proximal region, which is rarely explored. The web interface provides numerous graphical views that allow users to get an overview and to interpret the results more easily.**

Correspondence: ***veronique.brunaud@inrae.fr***

Background

Transcriptional activity of plant genes is a fundamental process in adaptation of these sessile organisms to the environment (1–6). In this context, many combined experimental (7) and *in silico* (8) efforts have been made to fully understand the mechanisms underlying the regulation of gene expression. Among the regulatory elements, transcription factors (TFs) are playing a major role in determining the tissue-specificity and developmental-stage-specificity of gene expression (9). These proteins regulate the expression of one or more target genes by binding a very short DNA sequence, typically between 5 and 25 bases, in their enhancer or promoter regions.

Predicting and characterizing these DNA motifs or transcription factor binding sites (TFBS) in a genome of several mega- or giga-bases poses an algorithmic and statistical challenge that can be partially solved by experimental approaches such as ChIP-seq (10) or DAP-seq (11) to name the best known. These sequencing-based technologies have allowed an extraordinary progress in the experimental determination of TFBSs, the results of which are stored in various databases such as JASPAR (12), REMAP (13) and Cis-BP (14). These databases analyze the sequencing profile results of all these experiments to focus on regions fixed by TF and propose TFBS candidates, often using a Position Weight Matrix (PWM) and logos. In this way, more than 700 TFBSs have been defined for the regulation of plant gene expression. However, these approaches have some limitations, such as the requirement of specific antibodies for each TF or TF family, the artificial accessibility of DNA in the case of *in vitro* methods, and the large DNA regions to be studied, ranging from 100 to 10,000 bases (10). To illustrate this limitation, only 400 of 2000 TFs encoded by the plant model *Arabidopsis thaliana* were examined using ChIP-seq and DAP-seq to identify their binding sites (source: REMAP database). Fortunately, inference of TFBS using *in silico* methods has become a relevant and complementary alternative. These methods have several advantages: they are able to rapidly analyze a large number of potential TFBSs, identify potential *cis*-regulatory sequences de novo without prior knowledge, and generate DNA binding sites with the highest-resolution. Interestingly, *cis*-regulatory sequences in the proximal regions located in the bases framing the transcription start site (TSS) or the transcription termination site (TTS) appear

to be associated with fixed topological constraints (15–18). Consequently, these proximal regions are essentially rich in *cis*-regulatory sequences (19–21). In this context, the *in silico* PLMdetect method was developed to identify short DNA sequences in *A. thaliana* and maize that are overrepresented when distance is constrained with respect to the TSS or the TTS and are therefore referred to as preferentially located motifs (PLMs) (21, 22). Numerous applications are already demonstrating interest in using PLMdetect to advance the characterization of the gene-proximal environment in targeted transcriptomic datasets (23–27) and at the genome level (21). To make PLMdetect accessible to the entire community, we introduce Plant-PLMview, a web database that allows all users to explore the proximal environment of genes by performing PLM searches for 20 plant species. Its use has been simplified to limit the amount of input data and interpretation of results is facilitated by visualization of PLMs. Plant-PLMview is original because it (i) offers the possibility to work on a panel of 20 plant species selected for their representativeness to the Angiosperm family and their presence in the orthogroup coding gene SyntenyViewer database (https://urgi.versailles.inra.fr/synteny/synteny/viewer.do); (ii) provides the opportunity to explore PLMs in the 3'-proximal region of the genes; (iii) provides access to an extensive catalog of 840 *cis*-regulatory sequences integrated from four complementary resources (12, 16, 28, 29), and (iv) provides a clear and interactive visualization of the results.

Construction and content

General structure. Plant-PLMview is an analysis database accessible directly through the internet (http://plmview.ips2.universite-paris-saclay.fr/). It was developed to identify *cis*-regulatory sequences in a number of genes of interest using PLMdetect, as described below. Plant-PLMview consists of two modules (Figure 1). The first module consists of two resources: (1) all 5'- and 3'-gene proximal sequences from 20 plant species and (2) the 840 *cis*-regulatory sequences integrated from experimental data. The second module is the query and tool part. It allows the user to run PLMdetect through the web interface by specifying a list of genes and a list of motifs. Once the analysis is completed, the results are returned by visualizing the PLMs on a map and as tables that can be downloaded.

Description of PLMdetect implemented in Plant–PLMview. PLMdetect attempts to identify DNA motifs that are overrepresented at a given distance from the TSS or the TTS. Given a set of sequences and a motif, PLMdetect counts the number of occurrences of the motif at each position in the sequence set to obtain a motif distribution. Then, a linear regression is estimated for the neutral region corresponding to the first 500 bases of the sequences, and the predicted values with an associated 99% confidence interval are calculated for the studied region corresponding to the remaining sequence downstream of the neutral region. Finally, PLMdetect declares the motif as PLM if a peak in the examined region exceeds the confidence interval. Each PLM is characterized by three indicators: (i) a preferential position, defined as the position of the maximum distribution peak, (ii) a functional window corresponding to the boundaries of the peak, and (iii) a score determined by the height of the peak (Figure 2).

Gene-proximal sequence resource. Plant-PLMview contains gene-proximal region sequences of 20 plant species covering the diversity of the Angiosperm family with representatives of the Brassicaceae, Rosaceae, Fabaceae, Cucurbitaceae, Solanaceae, and the Poaceae. These plant genomes and annotations were mainly extracted from the Phytozome database (Table 1) (30–35), and only gene sequences with an annotated 5' (3')-UTR were retained to ensure the TSS (TTS) annotations and thus the quality of the detected PLM. The sequences provided in the database consist of the interval [-1000;+500] relative to TSS for the 5'-gene-proximal region and the interval [-500;+1000] relative to TTS for the 3'-gene-proximal region. The sequences of antisense genes were added in reverse to standardize the application of PLMdetect.

***Cis*-regulatory sequence resource.** Plant-PLMview integrates experimentally characterized *cis*-regulatory sequences from four complementary plant-related resources: (i) 469 sequences from PLACE, containing all sequences described up to 2007, (ii) 99 sequences from AGRIS for *A. thaliana*, (iii) 15 sequences from work of (16) corresponding to TATA-like sequences, and (iv) 656 sequences from JASPAR Plant 2022 listing TFBSs identified mainly from PBM, ChIP-seq, DAP-seq or SELEX-seq data. Since JASPAR stores the information as Position Weight Matrices, they have to be converted into consensus sequences. This was done using the convert-matrix tool of the RSAT suite (36) with default parameters. We eliminated redundancies between resources and removed sequences larger than 16 bp or with too many IUPAC indeterminacies ($\geq$ 2 N, $\geq$ 3 DHVB, $\geq$ 4 RYSWKM). Following this curation, the unique 840 *cis*-regulatory sequences were organized into a motif-oriented resource. It is possible to explore this catalog of sequences from the query page using the "access to motif catalog" button. We have chosen a hierarchical graph to highlight the overlaps and inclusion links between the different *cis*-regulatory sequences. Figure 3 is the screenshot of the motif resource when a query was made to TGACG. The information provided is the description of the sequence, its other names, any associated bibliographic references, the list of species in which the sequence was described, and the source database (PLACE, AGRIS, JASPAR or (16)).

Database and web implementation. The database is built on the relational system PostgreSQL (version 13). Plant-PLMview allows downloading the list of 840 motifs and the studied 5' and 3' proximal regions as Fasta files. The interfaces of Plant-PLMview and plmdbview are written in Javascript Python3 and use the modules Flask, psycopg2, easyui, vis, and d3js.

Utility and discussion

The interface is typically used to identify *cis*-regulatory sequences among differentially expressed genes or co-expressed genes.

Use Case. We illustrate this by analyzing a group of genes from (23), the example accessible via the "DEMO" button on the query page. First, the user must select a species, enter the gene identifiers and the gene-proximal region (5' or 3') to be analyzed. In this use case, the selected species is *A. thaliana*, the queried genes are specified in the "paste a list of gene IDs" and we want to examine the 5'-gene proximal region (Figure 4). After entering his email address, the user then clicks on the "Run Demo" button to get a link with all the results. For further analysis, the user clicks the "Run PLM" button after entering the inputs.

For advanced users, the interface also provides optional parameters for filtering PLMs by score, setting the size of the sliding window for counting *cis*-regulatory sequences (22), and completing the search in a directed manner by disabling the "Two strands" option. In the use case, we leave the optional parameters as default.

After some time, Plant-PLMview returns a map of all identified PLMs. It allows the user to see at a glance the *cis*-regulatory motif candidates and their position in the gene-proximal region. This map is interactive: it is possible to zoom in and out, filter by genes or PLMs, and take screenshots. In our use case, Plant-PLMview identified 40 PLMs distributed across 125 genes (Figure 5). The map shows a strong presence of TATA-box motifs and their TATA-box derivatives upstream of TSS of the tested genes. It also shows strong conservation of some PLMs downstream of TSS, such as TGACG and GANTTNC, which are known to be bound by bZIP and MYB TFs, respectively. The visualization proposed by Plant-PLMview also highlights combinations of PLMs that may be involved in this co-regulation of genes. Using the same example of TGACG and GANTTNC PLMs, it can be sen t that they are co-present tin 31 genes, suggesting a potential combination actions of bZIP and MYB TFs in these 5'-gene proximal regions.

The results are also reproduced in a table showing all the information on each PLM (Figure 6). It includes: the DNA motif, the distribution of its occurence, its score, its positional window, its preferential position, its function, and the known information n about this *cis*-regulatory sequence. This table is downloaded using the "Download Results" button. The results are then organized in different directories: the "PLM_lists" directory contains for each PLM the list of genes in which it occurs , the "PLM_graphs" directory contains the distribution of occurrence and a file with all the information available on these PLMs.

To locate the PLM with respect to the *cis*-regulatory motifs stored in Plant-PLMview, the user simply clicks on the name of the motif in the result table and a an image like Figure 3 appears.

Future implementations. Plant-PLMview allows users to identify candidate *cis*-regulatory sequences underlying potential co-regulation of gene groups belonging to any of the 20 available species among the 840 *cis*-regulatory sequences in the database. For users who wish to examine species or *cis*-regulatory sequences not included in Plant-PLMview, we have provided the ability to provide 5' or 3'-gene-proximal regions and user-defined *cis*-regulatory sequences. The user-provided gene-proximal region file should list sequences defined by the range [-1000;+500] bp relative to TSS and [-500;+1000] bp relative to TTS. The sequence orientation guidelines provided in the "Genomes and gene-proximal sequences" section should also be followed. The custom motif file must contain DNA motifs that follow the IUPAC codes and are separated by a line break. The motifs provided must follow the filter guidelines (if motif size and indeterminacies are not respected, the motif will be removed) in the "*Cis*-regulatory sequence data sources and processing" section. In the current version of Plant-PLMview, only one species can be queried at a time. Our near term goal is to provide access to a cross-species PLM search to investigate co-evolution of *cis*-regulatory motifs by establishing the ability to select multiple species simultaneously on the query page and enter gene identifiers.

Conclusions

The search for *cis*-regulatory sequences involved in the regulation of gene clusters of interest is a widely studied topic. Access to *in silico* approaches remains limited for laypersons, indicating the need for user-friendly tools to accelerate research in this area. Here, we have introduced Plant-PLMview, which provides an easy-to-use web interface and easy-to-interpret results. This is made possible by the numerous pre-processing steps on the data and the PLM visualization as an interactive map.

References

1. Christina B Azodi, John P Lloyd, and Shin-Han Shiu. The cis-regulatory codes of response to combined heat and drought stress in Arabidopsis thaliana. NAR Genomics and Bioinformatics, 2(3), September 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa049.
2. Peng Zhou, Tara A Enders, Zachary A Myers, Erika Magnusson, Peter A Crisp, Jaclyn M Noshay, Fabio Gomez-Cano, Zhikai Liang, Erich Grotewold, Kathleen Greenham, and Nathan M Springer. Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. The Plant Cell, 34(1):514–534, January 2022. ISSN 1040-4651. doi: 10.1093/plcell/koab267.
3. Michael Alonge, Xingang Wang, Matthias Benoit, Sebastian Soyk, Lara Pereira, Lei Zhang, Hamsini Suresh, Srividya Ramakrishnan, Florian Maumus, Danielle Ciren, Yuval Levy, Tom Hai Harel, Gili Shalev-Schlosser, Ziva Amsellem, Hamid Razifard, Ana L. Caicedo, Denise M. Tieman, Harry Klee, Melanie Kirsche, Sergey Aganezov, T. Rhyker Ranallo-Benavidez, Zachary H. Lemmon, Jennifer Kim, Gina Robitaille, Melissa Kramer, Sara Goodwin, W. Richard McCombie, Samuel Hutton, Joyce Van Eck, Jesse Gillis, Yuval Eshed, Fritz J. Sedlazeck, Esther van der Knaap, Michael C. Schatz, and Zachary B. Lippman. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell, 182(1):145–161.e23, July 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.05.021.
4. Amanda J. Waters, Irina Makarevitch, Jaclyn Noshay, Liana T. Burghardt, Candice N. Hirsch, Cory D. Hirsch, and Nathan M. Springer. Natural variation for gene expression responses to abiotic stress in maize. The Plant Journal, 89(4):706–717, 2017. ISSN 1365-313X. doi: 10.1111/tpj.13414. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13414.
5. Shengxue Liu, Cuiping Li, Hongwei Wang, Shuhui Wang, Shiping Yang, Xiaohu Liu, Jianbing Yan, Bailin Li, Mary Beatty, Gina Zastrow-Hayes, Shuhui Song, and Feng Qin. Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. Genome Biology, 21(1):163, July 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02069-1.

6. Alan E. Yocca and Patrick P. Edger. Current status and future perspectives on the evolution of cis-regulatory elements in plants. Current Opinion in Plant Biology, 65:102139, February 2022. ISSN 1879-0356. doi: 10.1016/j.pbi.2021.102139.
7. Xuelei Lai, Arnaud Stigliani, Gilles Vachon, Cristel Carles, Cezary Smaczniak, Chloe Zubieta, Kerstin Kaufmann, and François Parcy. Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. Molecular Plant, 12(6):743–763, June 2019. ISSN 16742052. doi: 10.1016/j.molp.2018.10.010.
8. Valentina Boeva. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Frontiers in Genetics, 7:24, 2016. ISSN 1664-8021. doi: 10.3389/fgene.2016.00024.
9. Liang Song, Shao-Shan Carol Huang, Aaron Wise, Rosa Castanon, Joseph R. Nery, Huaming Chen, Marina Watanabe, Jerushah Thomas, Ziv Bar-Joseph, and Joseph R. Ecker. A transcription factor hierarchy defines an environmental stress response network. Science (New York, N.Y.), 354(6312):aag1550, November 2016. ISSN 1095-9203. doi: 10.1126/science.aag1550.
10. Narayan Jayaram, Daniel Usvyat, and Andrew C. R Martin. Evaluating tools for transcription factor binding site prediction. BMC bioinformatics, 17(1):547, November 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1298-9.
11. Anna Bartlett, Ronan C. O'Malley, Shao-Shan Carol Huang, Mary Galli, Joseph R. Nery, Andrea Gallavotti, and Joseph R. Ecker. Mapping genome-wide transcription-factor binding sites using DAP-seq. Nature Protocols, 12(8):1659–1672, August 2017. ISSN 1750-2799. doi: 10.1038/nprot.2017.055.
12. Jaime A. Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y. Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 50(D1):D165–D173, January 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab1113.
13. Fayrouz Hammal, Pierre de Langen, Aurélie Bergon, Fabrice Lopez, and Benoit Ballester. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Research, 50(D1):D316–D325, January 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab996.
14. Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J.M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell, 158(6):1431–1443, September 2014. ISSN 0092-8674. doi: 10.1016/j.cell.2014.08.009.
15. Yoshiharu Y. Yamamoto, Hiroyuki Ichida, Minami Matsui, Junichi Obokata, Tetsuya Sakurai, Masakazu Satou, Motoaki Seki, Kazuo Shinozaki, and Tomoko Abe. Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC genomics, 8:67, March 2007. ISSN 1471-2164. doi: 10.1186/1471-2164-8-67.
16. Virginie Bernard, Véronique Brunaud, and Alain Lecharny. TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. BMC Genomics, 11(1):166, March 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-166.
17. Willian Souza Bernardes and Marcelo Menossi. Plant 3' Regulatory Regions From mRNA-Encoding Genes and Their Uses to Modulate Expression. Frontiers in Plant Science, 11: 1252, 2020. ISSN 1664-462X. doi: 10.3389/fpls.2020.01252.
18. Tobias Jores, Jackson Tonnies, Travis Wrightsman, Edward S. Buckler, Josh T. Cuperus, Stanley Fields, and Christine Queitsch. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. Nature Plants, 7(6):842–855, June 2021. ISSN 2055-0278. doi: 10.1038/s41477-021-00932-y.
19. Chun-Ping Yu, Jinn-Jy Lin, and Wen-Hsiung Li. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Scientific Reports, 6(1):25164, April 2016. ISSN 2045-2322. doi: 10.1038/srep25164. Number: 1 Publisher: Nature Publishing Group.
20. Najla Ksouri, Jaime A Castro-Mondragón, Francesc Montardit-Tarda, Jacques van Helden, Bruno Contreras-Moreira, and Yolanda Gogorcena. Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. Plant Physiology, 185 (3):1242–1258, January 2021. ISSN 0032-0889. doi: 10.1093/plphys/kiaa091.
21. Julien Rozière, Cécile Guichard, Véronique Brunaud, Marie-Laure Martin, and Sylvie Coursol. A comprehensive map of preferentially located motifs reveals distinct proximal cis-regulatory sequences in plants. Frontiers in Plant Science, 13, 2022. ISSN 1664-462X.
22. Virginie Bernard, Alain Lecharny, and Véronique Brunaud. Improved detection of motifs with preferential location in promoters. Genome, 53(9):739–752, September 2010. ISSN 1480-3321. doi: 10.1139/g10-042.
23. Eduardo Bueso, Jesús Muñoz-Bertomeu, Francisco Campos, Veronique Brunaud, Liliam Martínez, Enric Sayas, Patricia Ballester, Lynne Yenush, and Ramón Serrano. ARABIDOPSIS THALIANA HOMEOBOX25 Uncovers a Role for Gibberellins in Seed Longevity. Plant Physiology, 164(2):999–1010, February 2014. ISSN 0032-0889. doi: 10.1104/pp.113.232223.
24. Eduardo Bueso, Jesús Muñoz-Bertomeu, Francisco Campos, Cándido Martínez, Carlos Tello, Irene Martínez-Almonacid, Patricia Ballester, Miguel Simón-Moya, Veronique Brunaud, Lynne Yenush, Cristina Ferrándiz, and Ramón Serrano. Arabidopsis COGWHEEL1 links light perception and gibberellins with seed tolerance to deterioration. The Plant Journal, 87(6):583–596, 2016. ISSN 1365-313X. doi: 10.1111/tpj.13220. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13220.
25. Nicolas Frei dit Frey, Ana Victoria Garcia, Jean Bigeard, Rim Zaag, Eduardo Bueso, Marie Garmier, Stéphanie Pateyron, Marie-Ludivine de Tauzia-Moreau, Véronique Brunaud, Sandrine Balzergue, Jean Colcombet, Sébastien Aubourg, Marie-Laure Martin-Magniette, and Heribert Hirt. Functional analysis of Arabidopsis immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences. Genome Biology, 15(6):R87, June 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-6-r87.
26. Clément Cuello, Aurélie Baldy, Véronique Brunaud, Johann Joets, Etienne Delannoy, Marie-Pierre Jacquemot, Lucy Botran, Yves Griveau, Cécile Guichard, Ludivine Soubigou-Taconnat, Marie-Laure Martin-Magniette, Philippe Leroy, Valérie Méchin, Matthieu Reymond, and Sylvie Coursol. A systems biology approach uncovers a gene co-expression network associated with cell wall degradability in maize. PloS One, 14(12):e0227011, 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0227011.
27. Stefania Del Prete, Anne Molitor, Delphine Charif, Nadia Bessoltane, Ludivine Soubigou-Taconnat, Cécile Guichard, Véronique Brunaud, Fabienne Granier, Paul Fransz, and Valérie Gaudin. Extensive nuclear reprogramming and endoreduplication in mature leaf during floral induction. BMC Plant Biology, 19(1):135, April 2019. ISSN 1471-2229. doi: 10.1186/s12870-019-1738-6.
28. Kenichi Higo, Yoshihiro Ugawa, Masao Iwamoto, and Tomoko Korenaga. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Research, 27(1):297–300, January 1999. ISSN 0305-1048. doi: 10.1093/nar/27.1.297.
29. Alper Yilmaz, Maria Katherine Mejia-Guerra, Kyle Kurz, Xiaoyu Liang, Lonnie Welch, and Erich Grotewold. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Research, 39(Database issue):D1118–1122, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1120.
30. David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, and Daniel S. Rokhsar. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research, 40(Database issue):D1178–D1186, January 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr944.
31. Tanya Z. Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait, and Eva Huala. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome: Tair: Making and Mining the "Gold Standard" Plant Genome. genesis, 53(8):474–485, August 2015. ISSN 1526954X. doi: 10.1002/dvg.22877.
32. Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, Andrew Berry, Jyothish Bhai, Alexandra Bignell, Konstantinos Billis, Sanjay Boddu, Lucy Brooks, Mehrnaz Charkhchi, Carla Cummins, Luca Da Rin Fioretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Jose Gonzalez Martinez, Cristina Guijarro-Clarke, Arthur Gymer, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Thomas Juettemann, Vinay Kaikala, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, José Carlos Marugán, Shamika Mohanan, Aleena Mushtaq, Marc Naven, Denye N Ogeh, Anne Parker, Andrew Parton, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, José G Pérez-Silva, William Stark, Emily Steed, Kyösti Sutinen, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Michal Szpak, Anja Thormann, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Natalie Willhoft, Andrea Winterbottom, Elizabeth Wass, Marc Chakiachvili, Bethany Flint, Adam Frankish, Stefano Giorgetti, Leanne Haggerty, Sarah E Hunt, Garth R Ilsley, Jane E Loveland, Fergal J Martin, Benjamin Moore, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, John Tate, David Thybert, Stephen J Trevanion, Sarah Dyer, Peter W Harrison, Kevin L Howe, Andrew D Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2022. Nucleic Acids Research, 50(D1):D988–D995, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1049.
33. Yi Zheng, Shan Wu, Yang Bai, Honghe Sun, Chen Jiao, Shaogui Guo, Kun Zhao, Jose Blanca, Zhonghua Zhang, Sanwen Huang, Yong Xu, Yiqun Weng, Michael Mazourek, Umesh K Reddy, Kaori Ando, James D. McCreight, Arthur A. Schaffer, Joseph Burger, Yaakov Tadmor, Nurit Katzir, Xuemei Tang, Yang Liu, James J. Giovannoni, Kai-Shu Ling, W. Patrick Wechter, Amnon Levi, Jordi Garcia-Mas, Rebecca Grumet, and Zhangjun Fei. Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. Nucleic Acids Research, 47(D1):D1128–D1136, January 2019. ISSN 1362-4962. doi: 10.1093/nar/gky944.
34. Bárbara Hufnagel, André Marques, Alexandre Soriano, Laurence Marquès, Fanchon Divol, Patrick Doumas, Erika Sallet, Davide Mancinotti, Sébastien Carrere, William Marande, Sandrine Arribat, Jean Keller, Cécile Huneau, Thomas Blein, Delphine Aimé, Malika Laguerre, Jemma Taylor, Veit Schubert, Matthew Nelson, Fernando Geu-Flores, Martin Crespi, Karine Gallardo, Pierre-Marc Delaux, Jérôme Salse, Hélène Bergès, Romain Guyot, Jérôme Gouzy, and Benjamin Péret. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. Nature Communications, 11(1):492, January 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14197-9. Number: 1 Publisher: Nature Publishing Group.
35. John L Portwood, II, Margaret R Woodhouse, Ethalinda K Cannon, Jack M Gardiner, Lisa C Harper, Mary L Schaeffer, Jesse R Walsh, Taner Z Sen, Kyoung Tak Cho, David A Schott, Bremen L Braun, Miranda Dietze, Brittney Dunfee, Christine G Elsik, Nancy Manchanda, Ed Coe, Marty Sachs, Philip Stinard, Josh Tolbert, Shane Zimmerman, and Carson M Andorf. MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Research, 47(D1):D1146–D1154, January 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1046.
36. Nga Thi Thuy Nguyen, Bruno Contreras-Moreira, Jaime A. Castro-Mondragon, Walter Santana-Garcia, Raul Ossio, Carla Daniela Robles-Espinoza, Mathieu Bahin, Samuel Collombet, Pierre Vincens, Denis Thieffry, Jacques van Helden, Alejandra Medina-Rivera, and Morgane Thomas-Chollier. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Research, 46(W1):W209–W214, July 2018. ISSN 1362-4962. doi: 10.1093/nar/gky317.

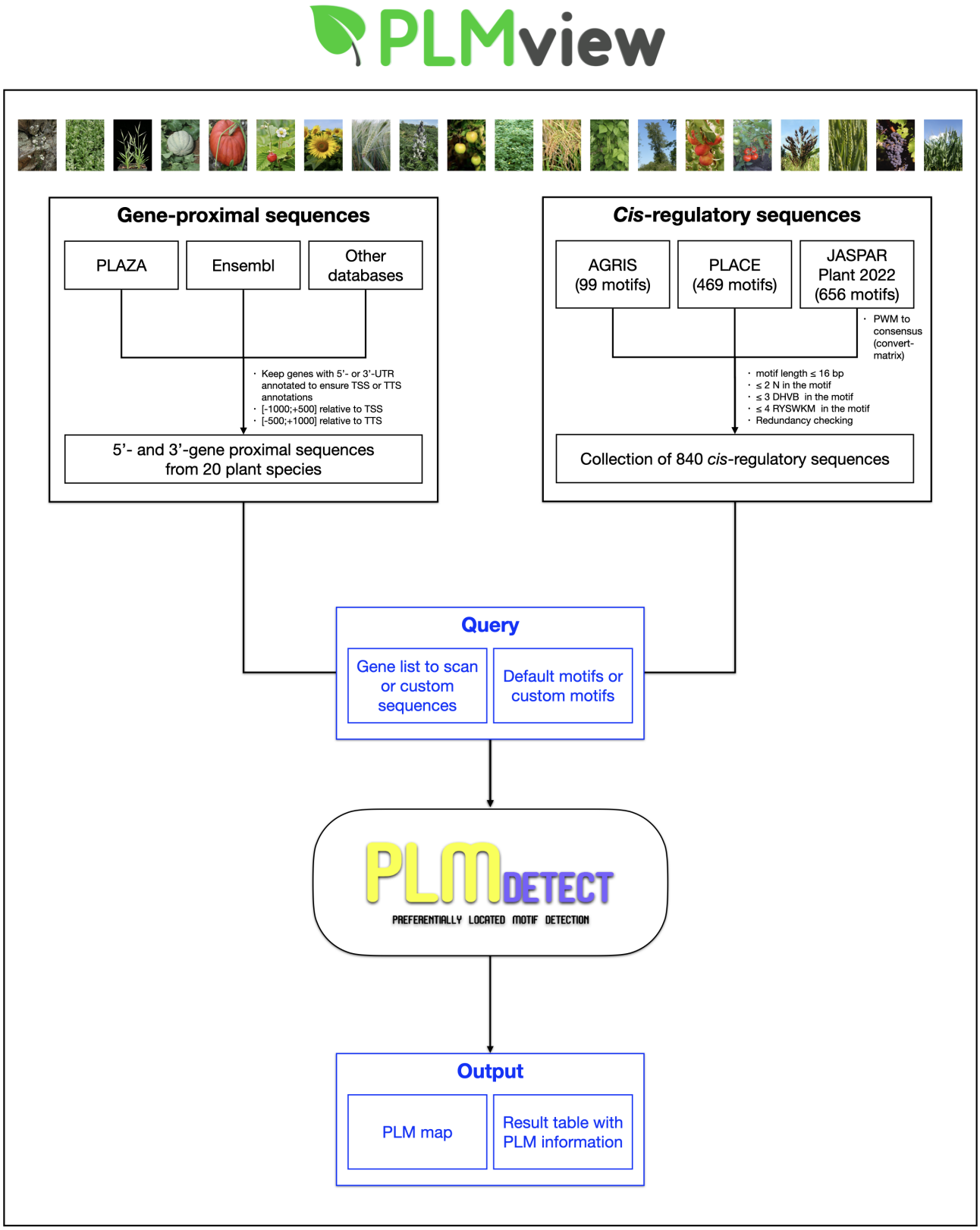


Fig. 1. General structure of Plant-PLMview. The resource module is shown in black and the query module in blue.

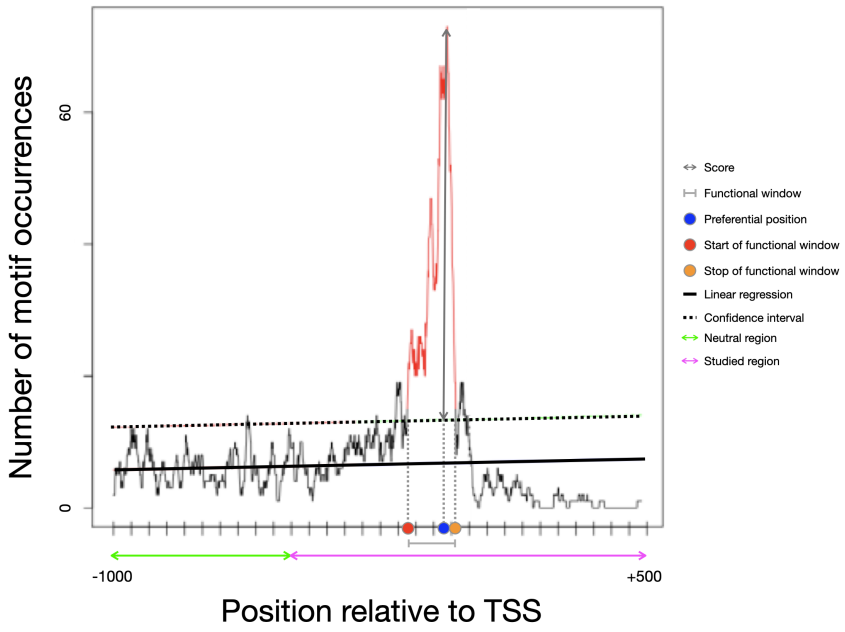


Fig. 2. Example of a PLM distribution illustrating the PLM characteristics.

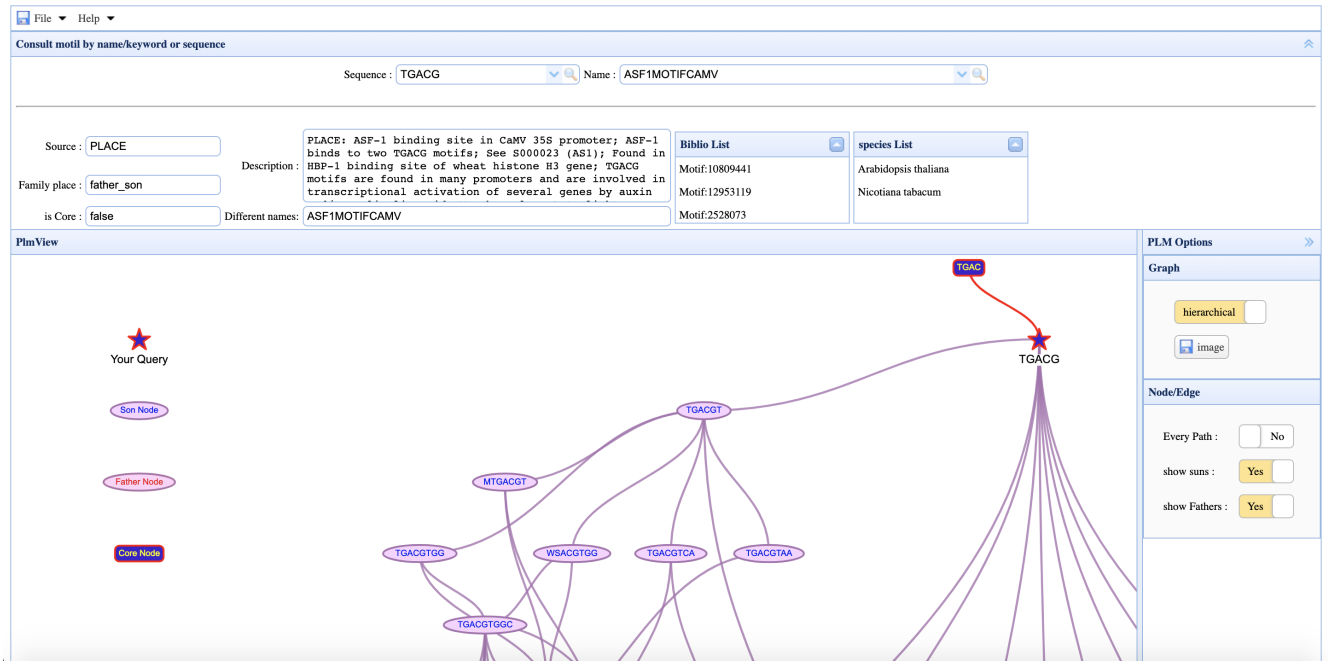


Fig. 3. Visualization of the TGACG family and information. The motif resource interface shows the details of each cis-regulatory sequence in our catalog. It allows access to the description of the motif, its source database, its place in the family, other names, associated bibliographic references, and the species in which this sequence has been described. A sequence family is defined by a set of sequences grouped by inclusion links. In a given family, if Sequence 1 is included in Sequence 2, Sequence 1 is considered as the father of Sequence 2. By reciprocal, Sequence 2 is the son of Sequence 1. In this example, TGACG is the father of TGACGT and the son of TGAC. The graphic part allows the user to visualize the sequence's family and to access each member by clicking on it.

Table 1. Description of the 20 plant species available in Plant-PLMview. Indicated are the genome versions, annotation versions, sources, number of genes described in the species, and genes included in the analyses of the 5' and 3'-gene proximal regions.

Species	Genome version	Genome annotation version	Total number of genes in reference	Number of 5'-gene-proximal regions considered	Number of 3'-gene-proximal regions considered	Sources
Arabidopsis lyrata	v1	v2.1	31,132	29,792 (96%)	29,726 (95%)	Phytozome (Goodstein *et al.*, 2012)
Arabidopsis thaliana (Thale cress)	TAIR10	TAIR10	28,775	19,736 (69%)	20,573 (71%)	TAIR (Berardini *et al.*, 2015)
Brachypodium distachyon (Purple false brome)	v3	v3.2	32,439	26,787 (83%)	27,315 (84%)	Phytozome (Goodstein *et al.*, 2012)
Cucumis melo (Melon)	v4	v2	27,427	18,886 (69%)	19,731 (72%)	Ensembl (Cunningham *et al.*, 2022)
Cucurbita maxima (Squash)	v1.1	v1.1	32,076	14,780 (46%)	16,054 (50%)	CuGenDB (Zheng *et al.*, 2019)
Fragaria vesca (Wild strawberry)	v4.0	v4.0	34,006	19,710 (58%)	20,117 (59%)	Phytozome (Goodstein *et al.*, 2012)
Helianthus annuus (Sunflower)	r1.0	r1.2	52,243	32,581 (62%)	35,930 (69%)	Phytozome (Goodstein *et al.*, 2012)
Hordeum vulgare (Barley)	r1	r1	39,734	31,486 (79%)	33,407 (84%)	Phytozome (Goodstein *et al.*, 2012)
Lupinus albus (White lupin)	20171117r1	20171117r1	38,258	32,122 (84%)	31,914 (83%)	White Lupin Genome (Hufnagel *et al.*, 2020)
Malus domestica (Apple)	v1.1	v1.1	45,116	28,834 (64%)	30,583 (68%)	Phytozome (Goodstein *et al.*, 2012)
Medicago truncatula (Barrelclover)	v4.0	v4.0	50,894	21,022 (41%)	21,849 (43%)	Phytozome (Goodstein *et al.*, 2012)
Oryza sativa (Rice)	v7.0	v7.0	42,189	22,122 (52%)	23,182 (55%)	Phytozome (Goodstein *et al.*, 2012)
Phaseolus vulgaris (Common bean)	v2.0	v2.1	27,433	22,367 (82%)	22,281 (81%)	Phytozome (Goodstein *et al.*, 2012)
Populus trichocarpa (Black cottonwood)	v4.0	v4.1	34,699	30,661 (88%)	30,846 (89%)	Phytozome (Goodstein *et al.*, 2012)
Prunus persica (Peach)	v2.0	v2.1	26,873	21,129 (79%)	21,817 (81%)	Phytozome (Goodstein *et al.*, 2012)
Solanum lycopersicum (Tomato)	v3.2	v3.0	35,768	17,683 (49%)	18,408 (51%)	Phytozome (Goodstein *et al.*, 2012)
Sorghum bicolor (Sorghum)	v3.0.1	v3.1.1	34,211	23,539 (69%)	24,636 (72%)	Phytozome (Goodstein *et al.*, 2012)
Triticum aestivum (Common wheat)	v2.2	v2	99,386	37,640 (38%)	48,429 (49%)	Phytozome (Goodstein *et al.*, 2012)
Vitis vinifera (Common grape vine)	v2.1	v2.1	31,845	14,412 (45%)	16,155 (51%)	Phytozome (Goodstein *et al.*, 2012)
Zea mays (Maize)	v4.39	v4.39	39,498	25,847 (65%)	25,199 (64%)	MaizeGDB (Portwood *et al.*, 2018)

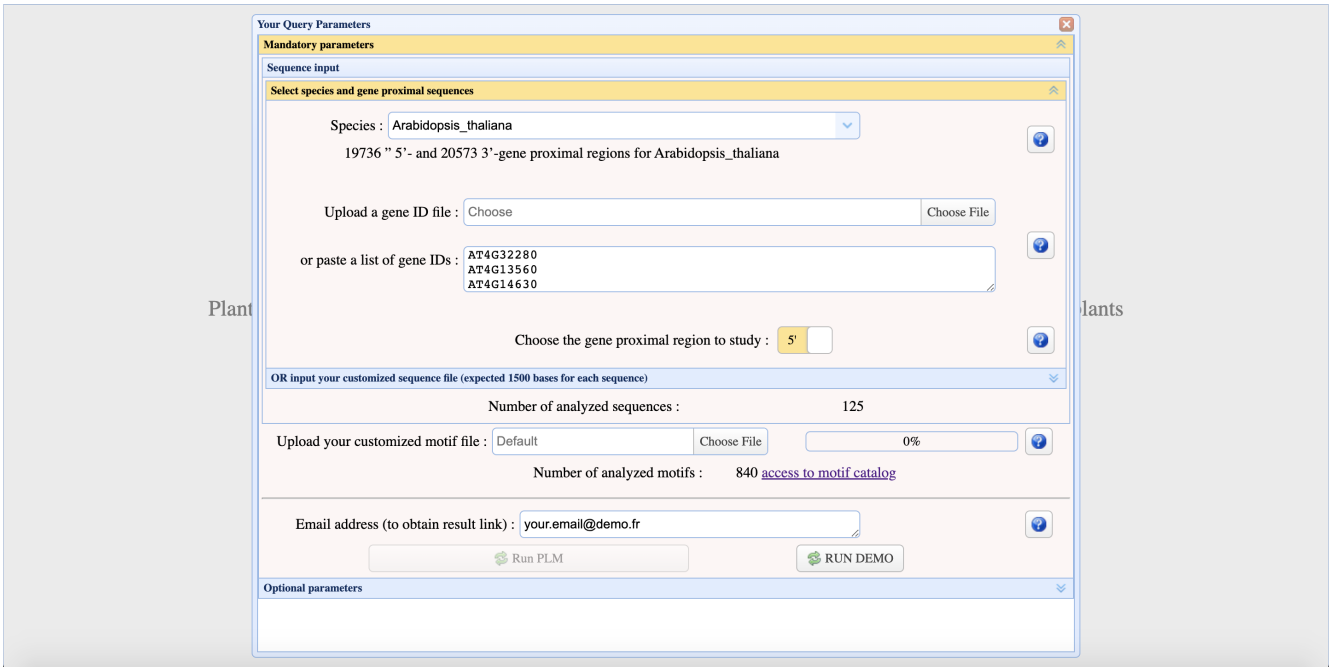


Fig. 4. Plant-PLMview query interface. To start a query, the user must select a species, specify the gene IDs, and select the gene-proximal region to be examined.

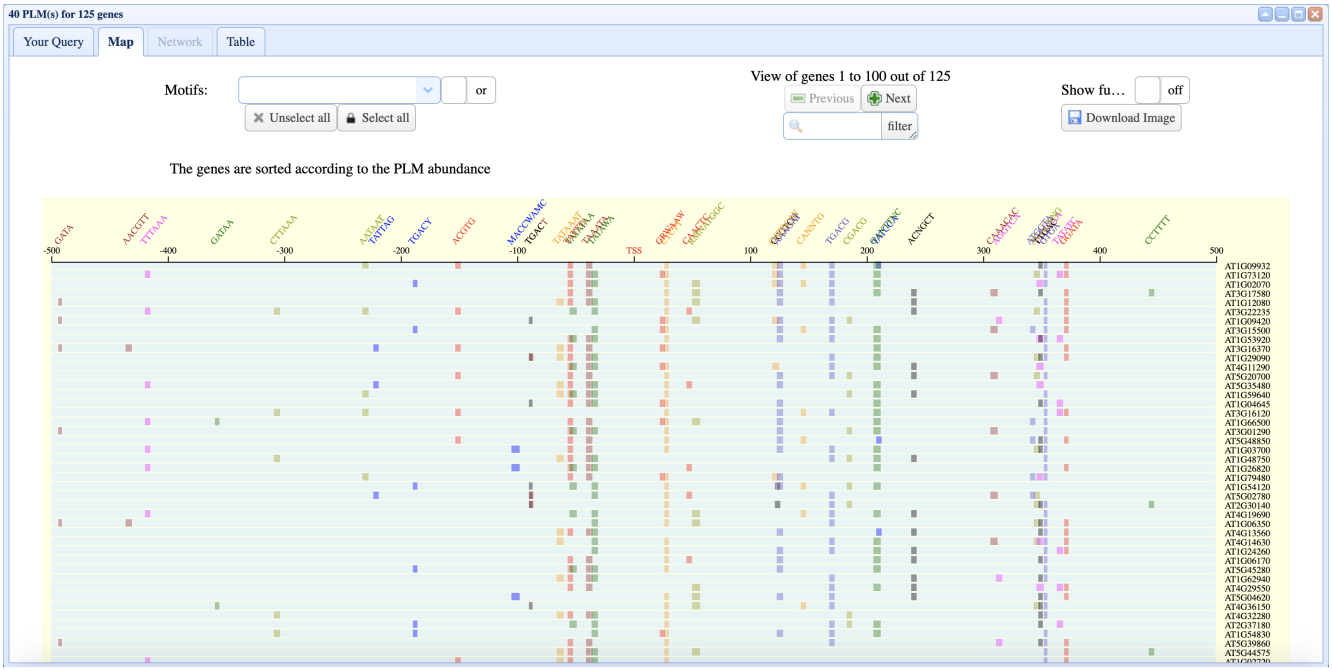


Fig. 5. Example of a PLM map. The map shows the PLMs identified by genes. The map is interactive and can be modified by filtering genes or PLMs. The visualization can be downloaded by clicking the "Download Image" button.

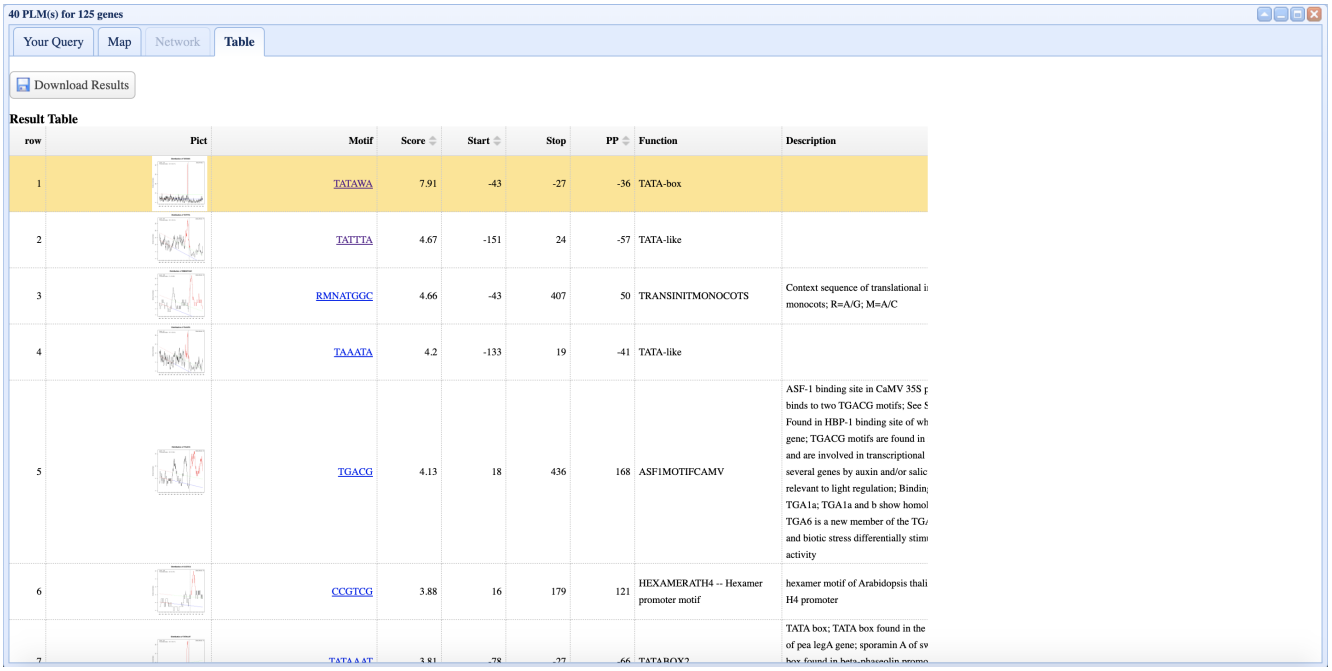


Fig. 6. Example of a PLM result table. This table lists the distribution of motifs for each PLM, the positional information, and the functions known in the bibliography. It is possible to click on the motif of each PLM to see it in the motif resource.

3.2 . Etude préliminaire de la conservation des PLM chez 18 espèces de plantes à fleurs

Parallèlement au développement de Plant-PLMview, j'ai initié des analyses pour étudier la conservation des PLM au sein des 20 espèces sélectionnées. Ce travail a fait l'objet du stage de M1 de Camille Lemercier que j'ai co-encadré avec Marie-Laure Martin (rapport en annexe B).

Dans un premier temps, Camille Lemercier a mis en place un contrôle qualité des PLM identifiés. Ce contrôle a révélé que les détections automatiques de PLM réalisées pour *Z. mays* et *Sorghum bicolor* n'étaient pas correctes. En effet, les séquences utilisées étaient issues d'une ancienne extraction et n'étaient pas de bonne longueur pour la détection automatique des PLM. Ces deux espèces ont donc été mises de côté et Camille Lemercier a poursuivi son travail avec les 18 autres espèces sélectionnées. Les données qu'elle a obtenues ont, à nouveau, permis d'identifier deux populations de PLM selon que leur score était inférieur ou supérieur à 2 (Figure 3.1). Ce résultat est similaire à celui observé chez *A. thaliana* et *Z. mays* (Rozière *et al.*, 2022b). Par ailleurs, chaque région proximale présente un nombre équivalent de PLM détectés, ce qui suggère, à nouveau, que la région 3'-proximale a un rôle important dans la structure génomique (Table 3.1).

Table 3.1 – Nombre de PLM détectés chez les 18 espèces.

Espèce	Nombre de 5'-PLM	Nombre de 3'-PLM	Nombre de séquences en 5'-proximales analysées	Nombre de séquences en 3'-proximales analysées
A.lyrata	7598	9986	29792	29726
A.thaliana	6387	6424	19736	20573
B.distachyon	6789	5732	26787	27315
C.maxima	3202	2770	14780	16054
C.melo	3024	3558	18463	19270
F.vesca	3244	1964	19710	20117
H.annuus	5605	3056	32581	35930
H.vulgare	3851	3845	31486	33407
L.albus	5288	4134	32122	31914
M.domestica	4944	4323	28834	30583
M.truncatula	3263	2594	21022	21849
O.sativa	7053	5948	22122	23182
P.persica	2725	2328	21129	21817
P.trichocarpa	5532	4962	30661	30846
P.vulgaris	2778	3125	22367	22281
S.lycopersicum	3244	2711	17683	18408
T.aestivum	6041	8118	37640	48429
V.vinifera	3715	4662	14412	16155

Enfin, l'étude de la distribution des PLM au sein de chacune des régions proximales, a révélé trois groupes de PLM dans la région 5'-proximale comme dans la région 3'-proximale des gènes (Figure 3.2), comme observé précédemment chez *A. thaliana* et *Z. mays* (Rozière *et al.*, 2022b). Cette structure est donc très conservée chez les plantes à fleurs.

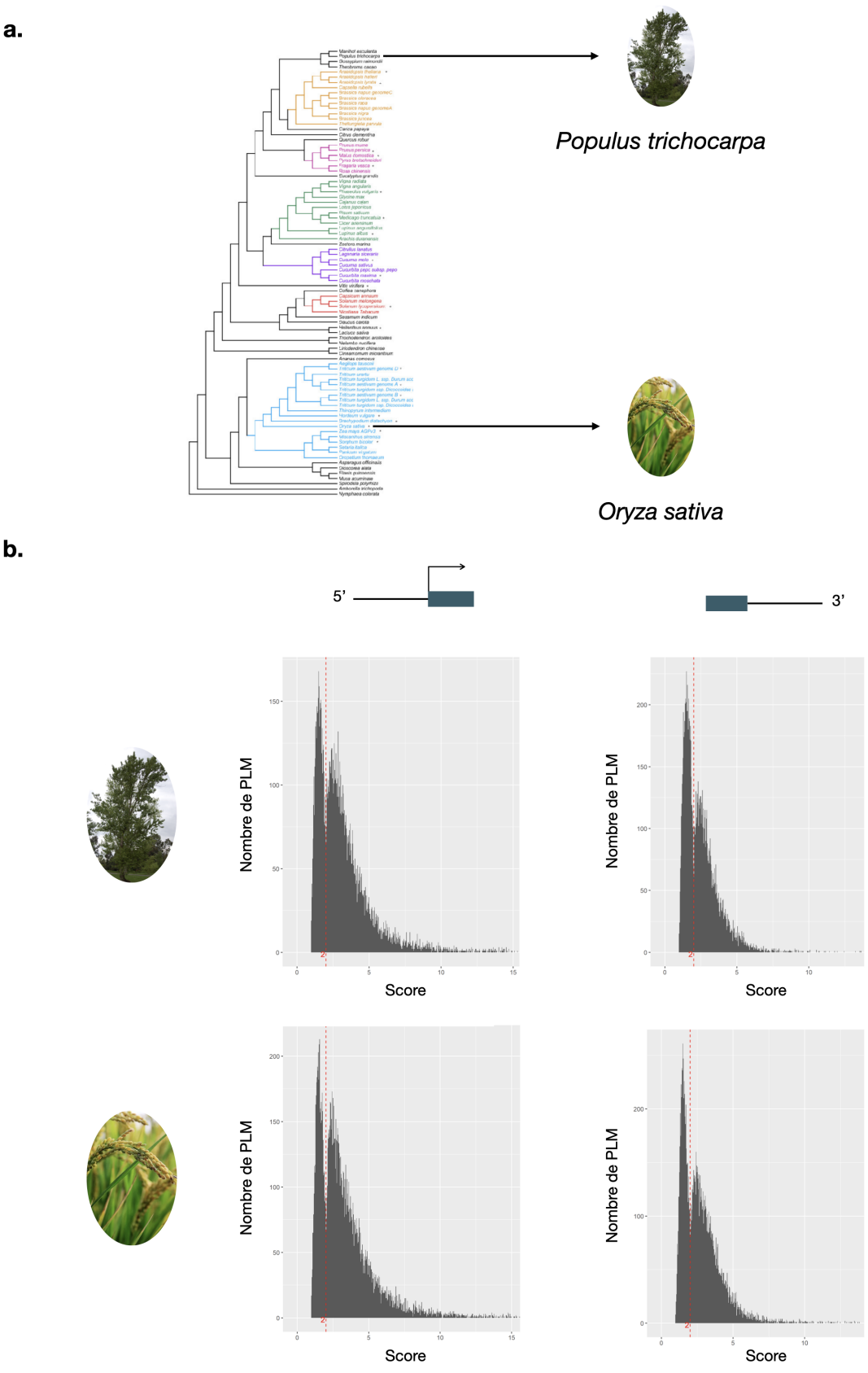


Figure 3.1 – Distribution des PLM selon leurs scores chez *Populus trichocarpa* et *Oryza sativa*. *(a) Position des deux espèces au sein de l'arbre des angiospermes. (b) Distribution des scores des PLM des deux espèces dans les régions 5' et 3'-proximales.*

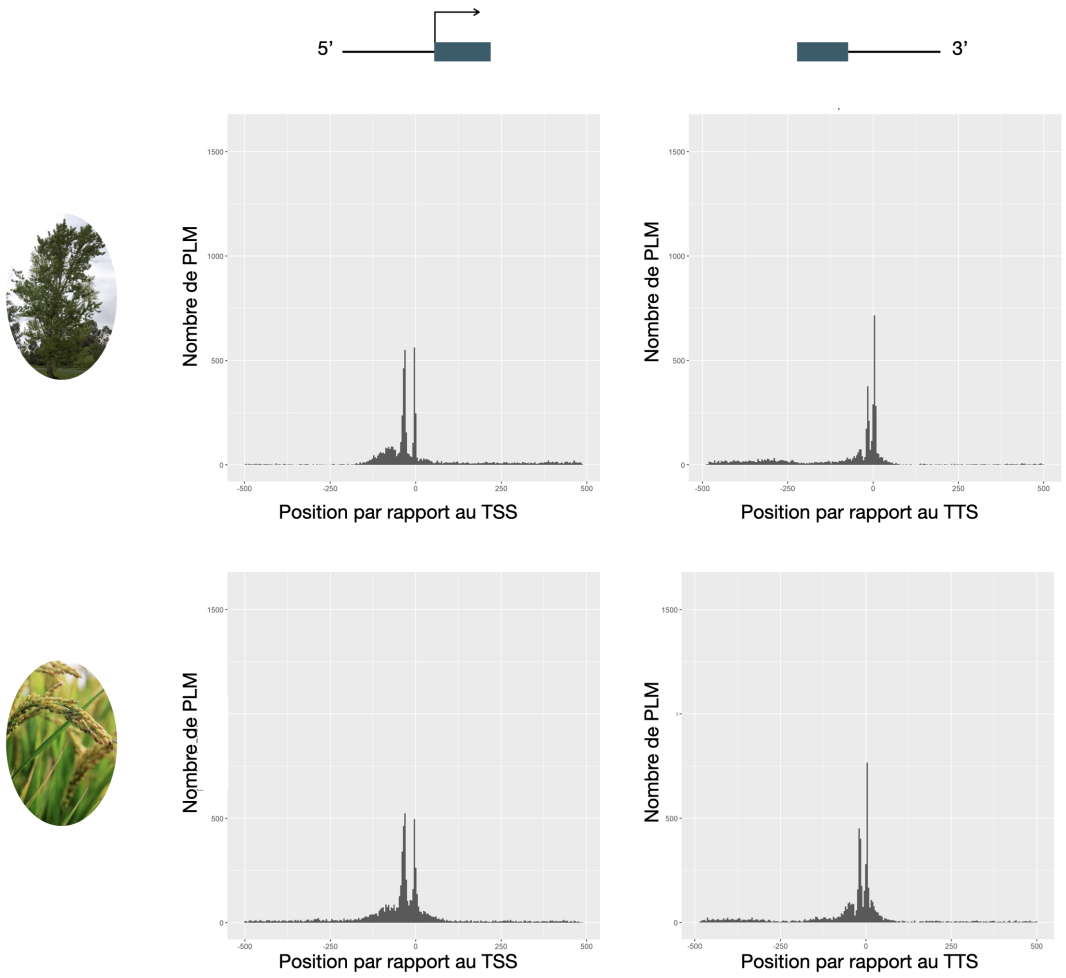
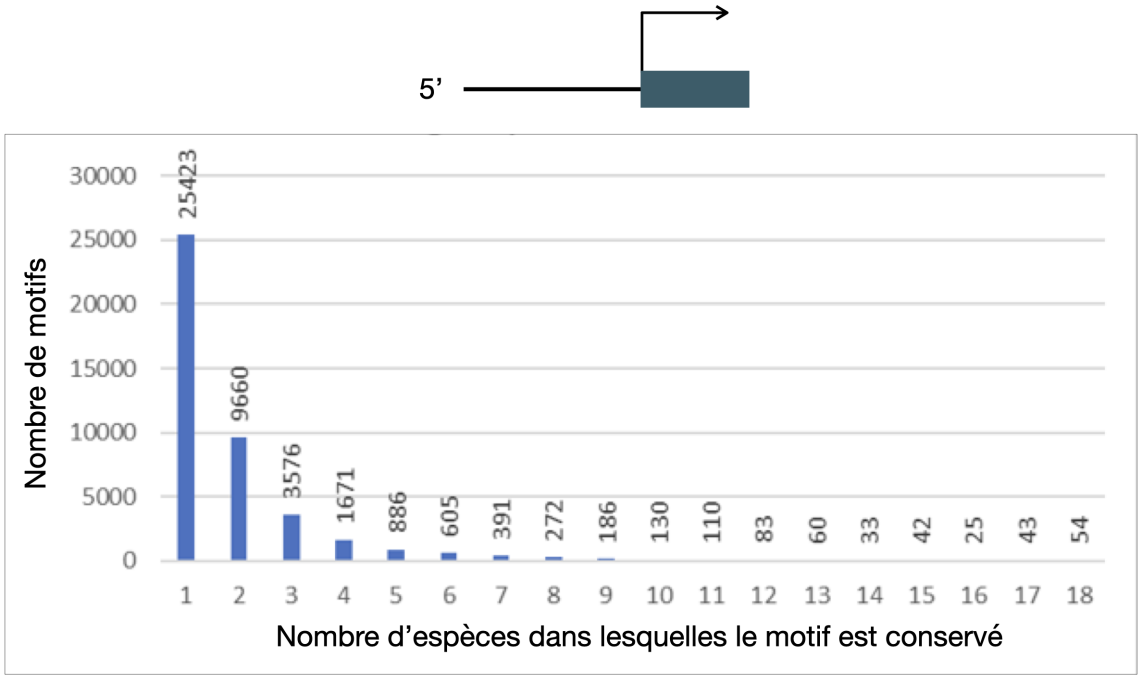


Figure 3.2 – Distribution des PLM selon leur position préférentielle chez *Populus trichocarpa* et *Oryza sativa*. *Les distributions des PLM selon leur position préférentielle sont présentées dans les régions 5' et 3'-proximales pour chacune des espèces.*

Dans un deuxième temps, une comparaison générale des PLM détectés a permis de mettre en avant que les séquences *cis*-régulatrices conservées au sein de l'ensemble des espèces étaient (1) les boîtes TATA dans la région 5'-proximale et (2) les sites TCAT liés à la polyadénylation dans la région 3'-proximale. Ce résultat appuie l'importance de ces signaux dans la structure génomique et dans la régulation de l'expression. Par ailleurs, les comparaisons réalisées ont montré que 59% des 5'-PLM et 65% des 3'-PLM sont spécifiques d'une espèce (Figure 3.3). Ces premiers résultats suggèrent qu'une partie, au moins, de ces PLM spécifiques constituerait des signaux moléculaires explicatifs des variations phénotypiques de chaque espèce.



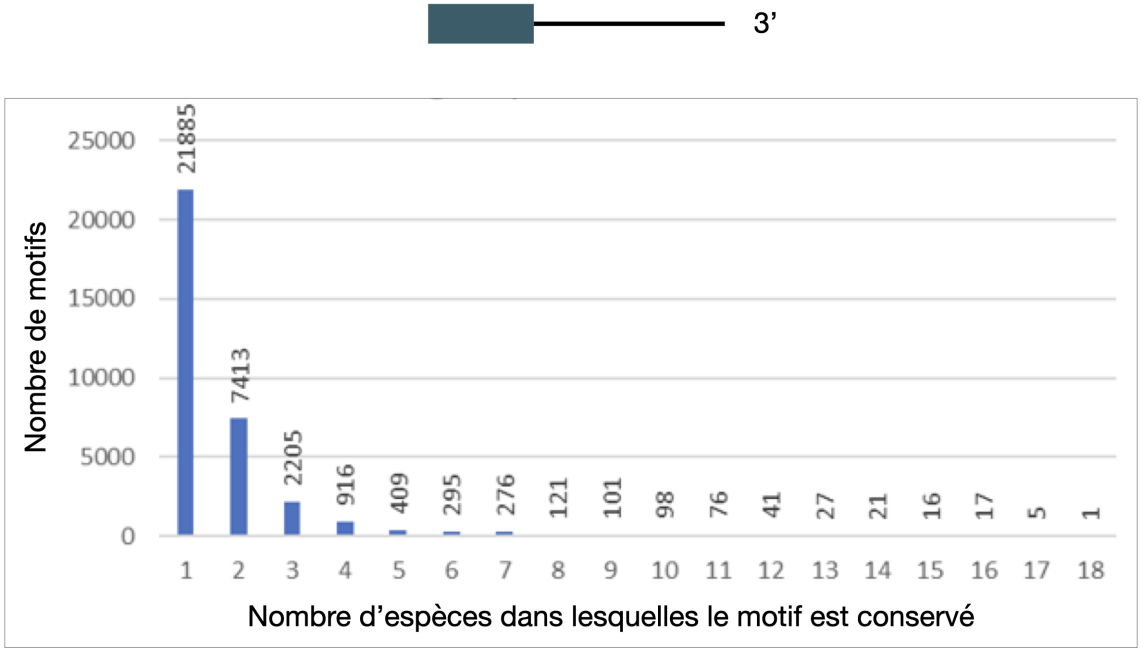


Figure 3.3 – Conservation des motifs PLM au sein des 18 espèces.

Dans un troisième et dernier temps, la stratégie a consisté à diviser en sous-régions les régions proximales des gènes sur la base des connaissances *a priori* que nous avons de celles-ci. Ainsi, la région 5'-proximale a été divisée en trois régions : (1) R1 [-500 ;-50] correspond au promoteur proximal, (2) R2 [-50 ;0] correspond au promoteur central et (3) R3 [0 ;+500] correspond à la sous-région génique. Pour son stage, Camille Lemercier s'est focalisée sur la région R2. Ces analyses ont permis d'identifier les variants de la boîte TATA et des motifs Inr (Figure Introduction 1.5) les plus conservés. Ces analyses ont aussi permis de révéler le motif AAATACC qui s'avère conservé chez 16 des 18 espèces étudiées et n'est jusqu'à présent associé à aucun TFBS connu. Sa position et sa forte conservation suggèrent que ce serait un nouveau motif du promoteur central, impliqué dans la fixation de la machinerie transcriptionnelle. Les travaux réalisés par Camille Lemercier doivent être étendus aux autres sous-régions de la région 5'-proximale et à l'ensemble de la région 3'-proximale. Une caractérisation des mécanismes régulateurs sous-jacents la présence des PLM (TFBS et séquences présentant des homologies de séquences avec les microARN) doit être également réalisée. À l'instar des pangénomes (Morneau, 2021), ce travail permettra d'élaborer une première version de "panrégulome proximal" avec l'identification du "régulome proximal coeur" conservé chez toutes les espèces et du "régulome proximal accessoire" avec la présence de PLM spécifiques des 20 espèces étudiées.

4 - Implication des PLM dans la réponse globale aux stress chez *A. thaliana*

4.1 . Contexte et objectifs

Dans un contexte de dérèglement climatique, les contraintes environnementales sont accentuées et étudier les réponses mises en place par les plantes est primordial. Pour cela, une étude menée par Marie-Laure Martin et Etienne Delannoy (IPS2) a cherché à identifier une réponse transcriptionnelle globale aux stress chez *A. thaliana* à partir d'une large étude de co-expression menée sur 18 catégories de stress (Zaag *et al.*, 2015). L'intégration de ces données de co-expression a permis d'identifier un réseau constituant la réponse transcriptomique commune aux stress. Ce réseau se compose de 3 407 gènes regroupés dans 47 sous-réseaux, eux-mêmes regroupés en 10 modules désignés M1 à M10 ci-après (Figure 4.1). Les sous-réseaux et les modules ont pu être annotés et sont associés à quatre grandes fonctions biologiques : (1) réponse aux stress, (2) fonction chloroplastique (photosynthèse, stress plastidial), (3) fonction mitochondriale (respiration cellulaire, traduction mitochondriale) et (4) fonction ribosomique (traduction, biogénèse des ribosomes). On peut aussi noter la présence de sous-réseaux enrichis en termes liés à l'organisation de la paroi cellulaire.

Dans ce chapitre, je présenterai le travail que j'ai entrepris pour exploiter cette ressource et ainsi étudier la réponse des plantes aux stress et le rôle des PLM dans cette réponse. Le travail que j'ai mené a consisté à :

- détecter les PLM *de novo* enrichis pour chaque sous-réseau afin d'identifier des séquences *cis*-régulatrices putatives, impliquées dans la co-régulation des gènes d'un même sous-réseau. Je leur ai ensuite assigné des mécanismes régulateurs (tPLM, miPLM ou uPLM) comme décrit précédemment dans le Chapitre 2 (Rozière *et al.*, 2022b).
- utiliser les tPLM identifiés pour inférer des TF impliqués dans la co-régulation de chacun des sous-réseaux et ainsi mettre en évidence des modules "TF-sous-réseau".
- développer une méthodologie pour identifier les uPLM susceptibles d'être des séquences *cis*-régulatrices (Rozière *et al.*, 2022b) et déterminer les expériences humides à mettre en oeuvre pour les valider.

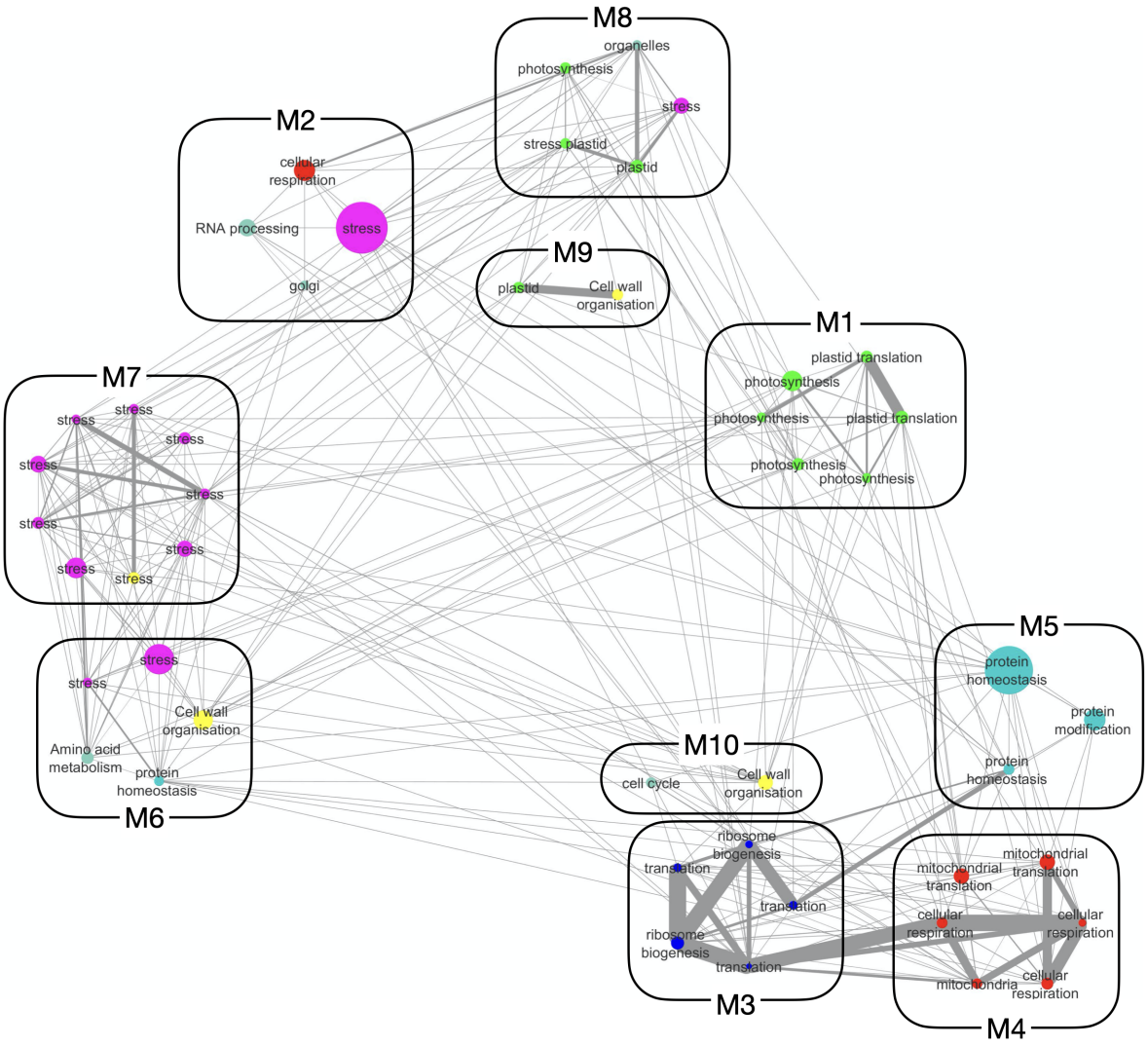


Figure 4.1 – Représentation du réseau de co-expression de réponse globale aux stress chez *A. thaliana* (adaptée de Delannoy *et al.*, en préparation).

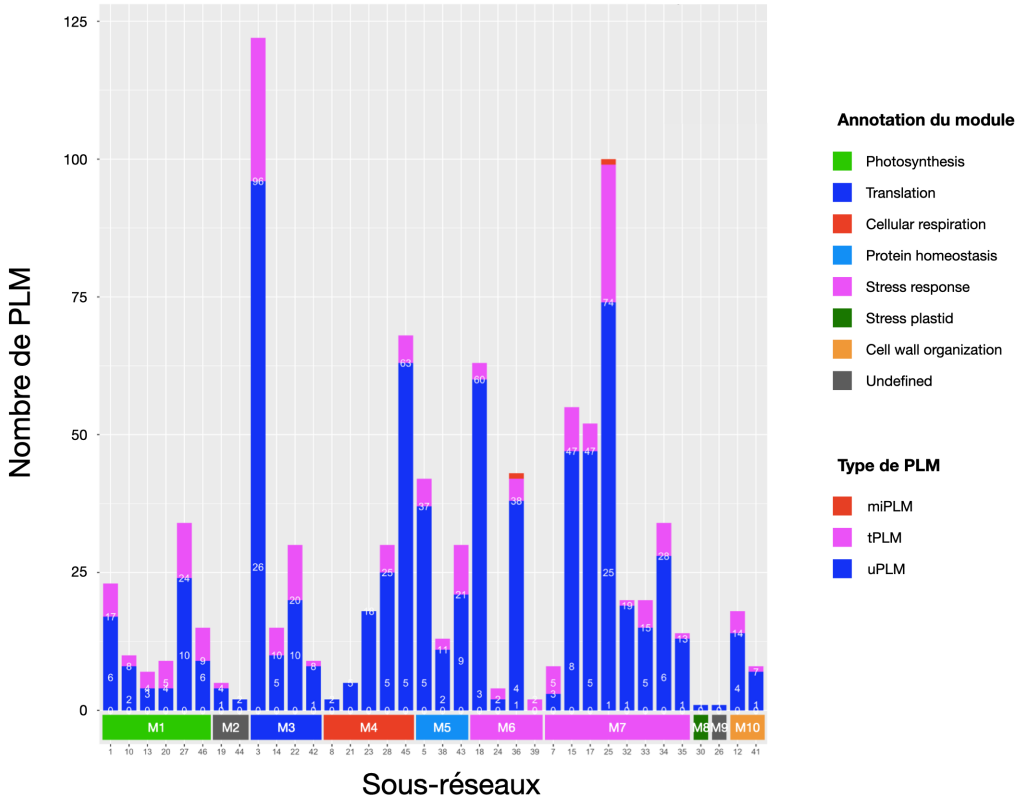
4.2 . Détection des PLM enrichis dans les sous-réseaux de co-expression

Dans l'objectif d'identifier des séquences *cis*-régulatrices proximales de la co-régulation de chaque sous-réseau de gènes, j'ai cherché à identifier des PLM *de novo* enrichis pour chacun d'entre eux. Dans un premier temps, j'ai réalisé une détection de PLM similaire à celle réalisée sur le génome d'*A. thaliana* (Chapitre 2), en exploitant l'ensemble des 4 à 8-mers, puis en filtrant les PLM identifiés sur la

base de leur score (score ≥ 2). Afin d'identifier les PLM enrichis dans chaque sous-réseau, j'ai comparé l'occurrence relative de chaque PLM dans son sous-réseau par rapport à son occurrence relative dans l'ensemble du réseau en réalisant un test hypergéométrique. Après un contrôle du nombre de faux positifs avec un ajustement de Benjamini-Hochberg des probabilités critiques (FDR $\leq$ 0,05), j'ai identifié (1) 932 5'-PLM enrichis et répartis sur 36 sous-réseaux et (2) 112 3'-PLM enrichis et répartis sur 20 sous-réseaux (Figure 4.2). Les PLM ont ensuite été assignés à des TFBS ou des séquences présentant des homologies avec des microARN comme décrit précédemment dans le Chapitre 2 (Rozière *et al.*, 2022b). Ces comparaisons ont mis en avant 2 (0,2%) miPLM, 173 (18,6%) tPLM et 757 (81,2%) uPLM dans les régions 5'-proximales (Figure 4.2). Pour la région 3'-proximale, je n'ai pu identifier aucun miPLM. En revanche, 19 (17%) tPLM et 93 (83%) uPLM ont été détectés (Figure 4.2).

Il est intéressant d'observer que la répartition du nombre de PLM n'est pas homogène entre les sous-réseaux et que les 5'-PLM sont plus abondants que les 3'-PLM. Cependant, les sous-réseaux 11, 16, 31 et 40 ne présentent que des 3'-PLM enrichis, sachant que les sous-réseaux 11 et 16 n'ont pas d'annotations fonctionnelles associées, alors que les sous-réseaux 31 et 40 sont liés aux fonctions chloroplastiques et annotées "stress plastidial" (Figure 4.2).

a.



b.

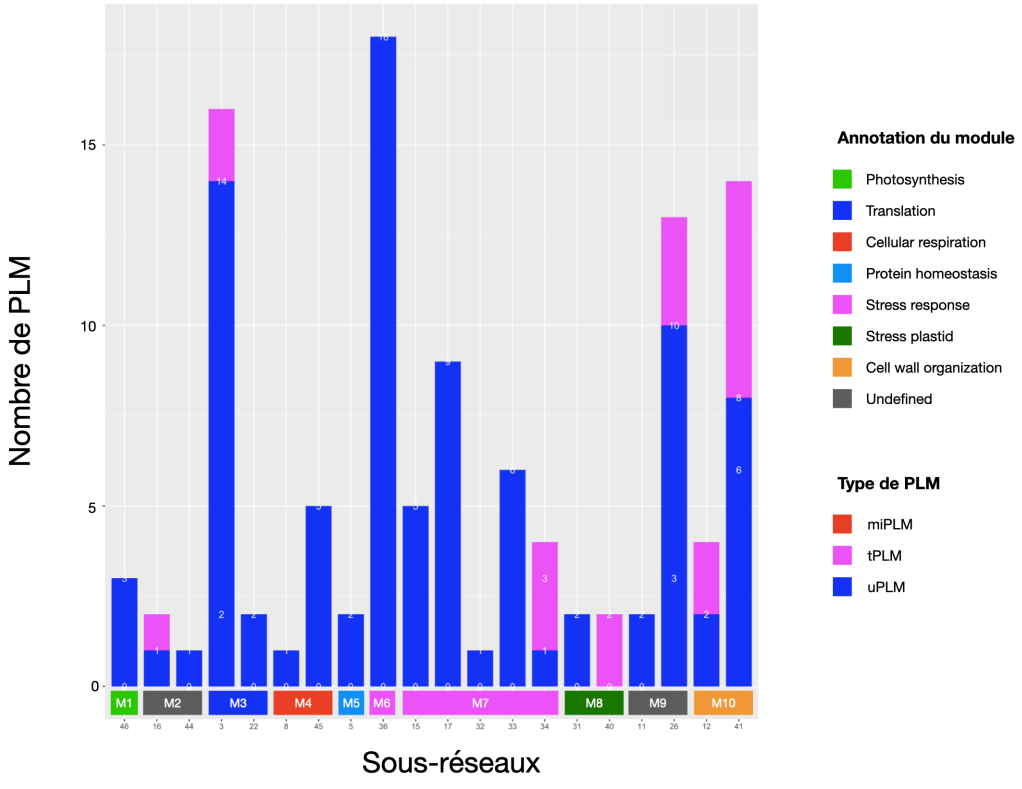


Figure 4.2 – Distribution des PLM enrichis parmi les 47 sous-réseaux du réseau de réponse au stress. *(a) 5'-PLM enrichis. (b) 3'-PLM enrichis. Les rectangles M1 à M10 indiquent les modules dans lesquels sont regroupés les sous-réseaux et les couleurs correspondent aux fonctions biologiques associées.*

4.3 . Inférence des TF impliqués dans la co-régulation des sous-réseaux

Je me suis ensuite servi de la présence des tPLM enrichis pour identifier des TF impliqués dans la co-régulation des gènes d'un même sous-réseau. En effet, la présence des tPLM, similaires en séquence à un/des TFBS caractérisés, permet d'inférer une interaction avec les TF capables de reconnaître les TFBS. J'ai ainsi pu identifier 257 TF appartenant à 24 familles pour la région 5'-proximale et 48 TF appartenant à 8 familles pour la région 3'-proximale. Les modules de régulation "TF-sous-réseaux" sont représentés à la Figure 4.3.

Dans la région 5'-proximale, les familles de TF les plus abondantes sont les WRKY, AP2/ERF, bHLH, bZIP et MYB-related. Il est intéressant de noter que toutes ces familles sont connues pour intervenir dans de nombreux processus biologiques incluant les réponses aux stress biotiques et abiotiques (Hong, 2016). Les analyses descriptives des modules "TF-sous-réseau" (Figure 4.3) révèlent que les TF de la famille :

- des WRKY sont prédits pour réguler trois sous-réseaux annotés stress, un sous-réseau lié à l'homéostasie des protéines et un sous-réseau associé à l'organisation des parois cellulaires.
- des AP2/ERF peuvent interagir fortement avec des sous-réseaux liés à la respiration cellulaire et à l'homéostasie des protéines. Ils présentent aussi quelques interactions avec un sous-réseau lié à la traduction et un autre associé à l'organisation de la paroi cellulaire.
- des bHLH forment des modules avec des sous-réseaux annotés stress et un sous-réseau lié à la traduction.
- des bZIP sont également liés à des sous-réseaux stress et associés à la respiration cellulaire.
- des MYB-related semblent avoir une position plus centrale dans la régulation du réseau se manifestant par des interactions avec 18 sous-réseaux annotés stress, photosynthèse, homéostasie des protéines et organisation de la paroi cellulaire.

Enfin, il est intéressant de noter que les TF de la famille des MYB ont également un rôle central dans la régulation du réseau et sont inférés pour être liés à 20 sous-réseaux.

Dans la région 3'-proximale, les modules "TF-sous-réseau" identifiés sont bien moins nombreux que ceux observés pour la région 5'-proximale. Les cinq familles de TF identifiées avec le plus de représentants sont les MYB-related, les Homeo domain factors, les C2H2 zing finger fators, les HD-ZIP factors et les MYB. À l'instar de la région 5'-proximale, les TF de la famille des MYB sont centraux dans la régulation des sous-réseaux et interagissent avec 6 sous-réseaux (Figure 4.3b).

De manière intéressante, trois sous-réseaux présentent uniquement des tPLM enrichis dans la région 3'-proximale. Il s'agit des sous-réseaux 16, 26 et 40. Les sous-réseaux 16 et 26 ne présentent pas d'enrichissements fonctionnels. Le sous-

réseau 16 est lié à des TF de la famille des MYB et MYB-related. Le sous-réseau 26 est lié à des TF des familles MYB, NAC/NAM et G2-like. Le sous-réseau 40 est enrichi en termes liés au stress plastidial et est lié à des TF MYB.

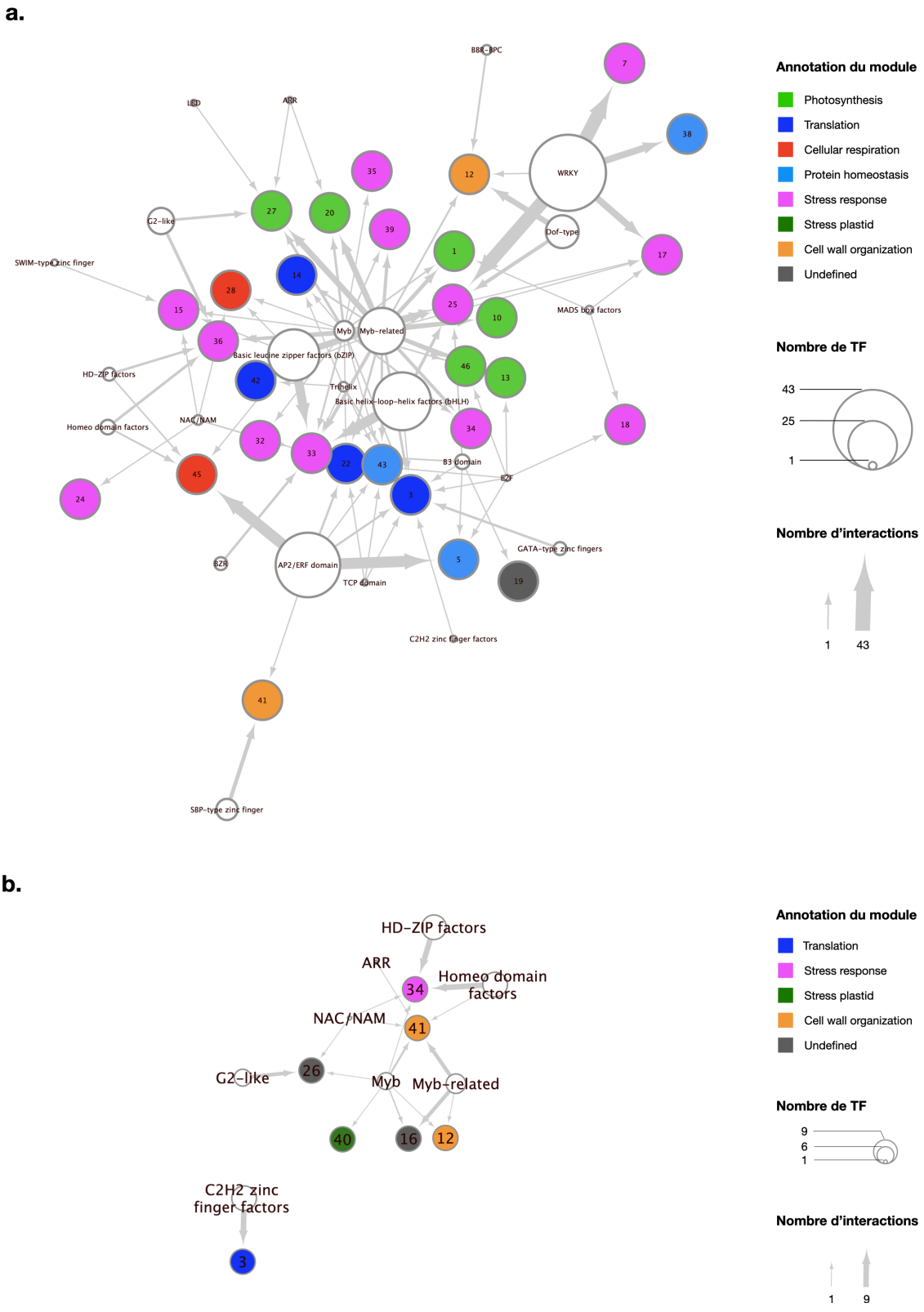


Figure 4.3 – Représentation des modules "TF-sous-réseaux" dans chaque région proximale étudiée. (a) Région 5'-proximale. (b) Région 3'-proximale. *Chaque sous-réseau est coloré selon les enrichissements fonctionnels du module dans lequel il se trouve (voir Figure 4.2). Les familles de TF sont indiquées en blanc et la taille du cercle est relative au nombre de TF inférés. La largeur des flèches est relative au nombre d'interactions prédites entre la famille de TF et le sous-réseau.*

4.4 . Identification de séquences *cis*-régulatrices putatives dans la région 5'-proximale et développement d'une méthodologie pour les valider

A côté des tPLM, j'ai pu identifier un grand nombre de uPLM enrichis susceptibles d'être des éléments *cis*-régulateurs impliqués dans la co-régulation transcriptionnelle des sous-réseaux. Afin de tester cette hypothèse et procéder à des validations humides des uPLM candidats, j'ai développé une méthodologie capable :

- d'identifier les uPLM les plus susceptibles d'avoir un impact transcriptionnel sur les gènes des sous-réseaux.
- de caractériser les conditions expérimentales idéales à reproduire pour entreprendre des validations humides. Cela est nécessaire car les séquences *cis*-régulatrices, tout comme les gènes qu'elles régulent, ne sont pas toujours actives et dépendent de plusieurs facteurs, comme le stade de développement, le tissu ou le stress appliqué à la plante (Schmitz *et al.*, 2021). Ainsi, il est nécessaire de déterminer des conditions expérimentales pour lesquelles la présence du uPLM est supposée être explicative d'une modulation de l'expression.
- de prendre en considération la présence de tPLM pouvant fonctionner en combinaison avec le uPLM testé car il est connu que les séquences *cis*-régulatrices fonctionnent très largement en module (Schmitz *et al.*, 2021).
- d'identifier les gènes sur lesquels entreprendre les différentes constructions moléculaires pour mesurer l'impact de mutations des uPLM testés et des tPLM associés.

4.4.1 . Identification de 5'-uPLM candidats

Afin d'identifier des séquences *cis*-régulatrices non caractérisées jusqu'à présent, j'ai établi les liens d'inclusion pour l'ensemble des motifs PLM identifiés dans la région 5'-proximale, générant ainsi des arbres d'inclusion de motifs. Cela m'a permis d'identifier neuf 5'-uPLM ne faisant pas partie d'une famille présentant des TFBS (Figure 4.4). Faute de temps, je n'ai pas pu réaliser ce travail pour la région 3'-proximale.

Pour s'assurer de leur potentielle implication dans la régulation de l'expression des gènes, une sous-sélection a été effectuée parmi les 9 5'-uPLM candidats. Pour faire cette sélection, trois facteurs ont été pris en compte simultanément :

- le score du 5'-uPLM qui devait être supérieur à 3.
- une homologie forte du 5'-uPLM avec un TFBS d'espèces non végétales. Pour cela, les séquences des 5'-uPLM ont été comparées à celles des TFBS répertoriés dans JASPAR 2020 (non plante) en suivant le même protocole de comparaison que celui suivi pour identifier les tPLM (section 4.2). Si la similarité de séquence était significative (qvalue $\leq$ 0,05), un tblastn était ensuite réalisé entre le TF associé et les gènes d'*A. thaliana* pour identifier de potentiels candidats fixant le 5'-uPLM *in planta*. J'ai enfin fait le choix

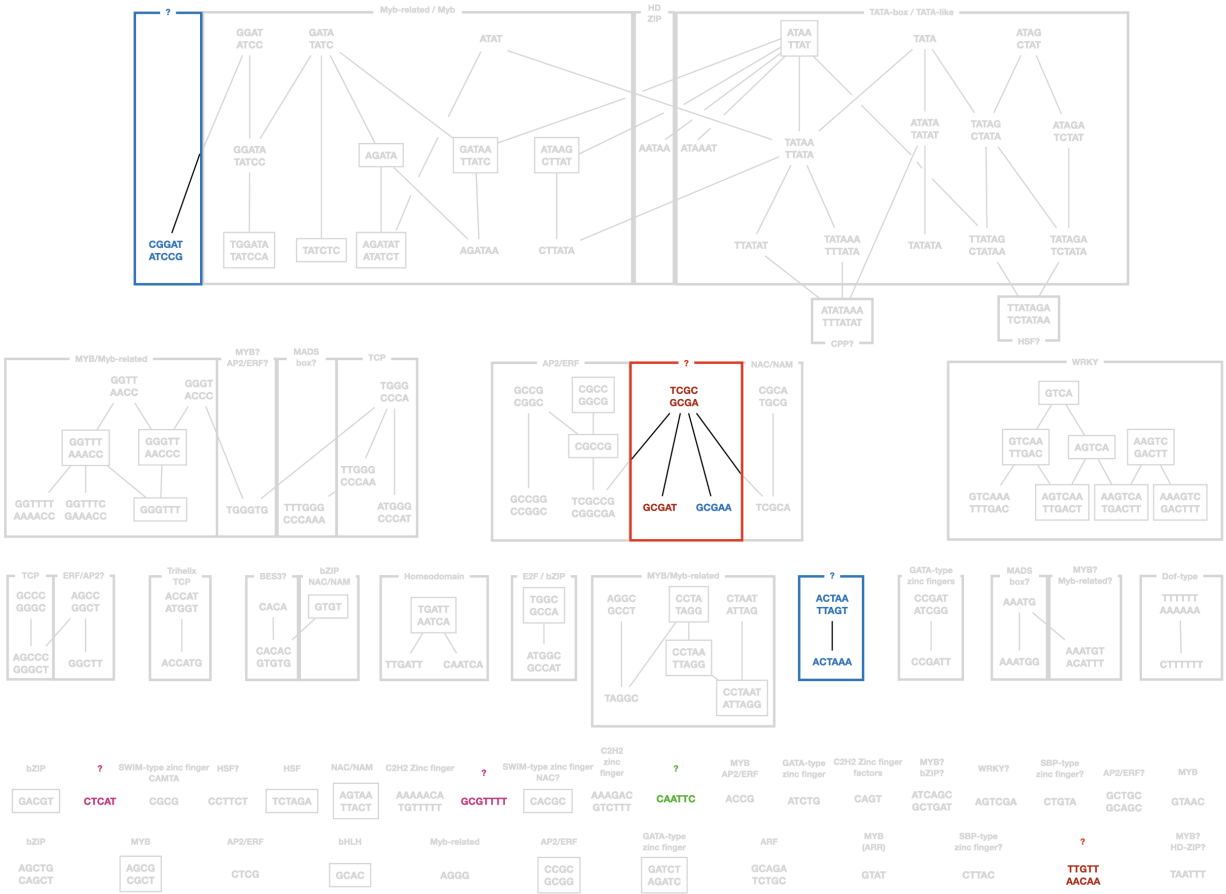


Figure 4.4 – Représentation de l'arbre d'inclusion des PLM identifiés dans le réseau de réponse globale aux stress. *Les cadres regroupent des membres d'une même famille de motif avec des TF associés identiques. Les motifs grisés sont ceux non retenus comme candidats sur la base de l'arbre. Les couleurs des motifs sont relatives aux annotations fonctionnelles des modules dans lesquels les sous-réseaux sont regroupés. Bleu : Ribosome; Rouge : Mitochondrie; Vert : Chloroplaste; Rose : Stress.*

de regrouper les 5'-uPLM d'une même famille de motif pour ne garder que le représentant avec le score le plus élevé.

— une différence significative du niveau d'expression entre les gènes possédant le 5'-uPLM et ceux ne le possédant pas pour au moins une expérience répertoriée dans la base de données CATdb (Gagnot *et al.*, 2008).

Cette base de données répertorie des données transcriptomiques obtenues à l'aide de puces CATMA pour 21 736 gènes d'*A. thaliana* étudiés dans 2 216 expériences. Pour chaque expérience, un test de Student a été réalisé pour comparer la moyenne des niveaux d'expression des gènes possédant le 5'-uPLM et ceux ne le possédant pas. J'ai retenu les expériences pour lesquelles les niveaux d'expression entre le groupe de gènes ayant le 5'-uPLM et celui ne l'ayant pas, étaient significativement différents après ajustement des p-valeurs par la procédure de Benjamini-Hochberg pour contrôler le taux de faux positifs à 5%. Les 5'-uPLM candidats étaient ceux pour lesquels au moins une expérience dans laquelle les niveaux d'expression des gènes avec le 5'-uPLM étaient significativement différents de ceux des gènes sans le 5'-uPLM (Table 4.1). En conciliant tous ces critères, les 5'-uPLM GCGAA et CAATTC ont été considéré candidats (Table 4.1). Le 5'-uPLM GCGAA est retrouvé dans les sous-réseaux 3 et 45 annotés respectivement "Traduction" et "Respiration cellulaire". Le 5'-uPLM CAATTC est retrouvé dans le sous-réseau 27 annoté "Photosynthèse".

Table 4.1 – Caractéristiques des 9 uPLM candidats.

5'-uPLM	Score	TF	Espèce	Gène *A. thaliana* (tblastn)	Nom	Evalue (tblastn)	Nb exp signif
GCGAA	8,08	SWI4	*S. cerevisiae*	AT2G14255	-	0,004	297
ACTAAA	2,37	-	-	-	-	-	17
CGGAT	2,54	SPDEF	*H. sapiens*	-	-	-	37
CTCAT	2,29	POU6F1	*H. sapiens*	AT4G00730	ANL2	0,004	48
		JUNB	*H. sapiens*	AT5G11260	HY5	4	
GCGTTTT	4,21	-	-	-	-	-	-
CAATTC	3,27	ZNF140	*H.sapiens*	AT5G04240	ELF6	7,00E-10	8
				AT4G06634	YY1	2,00E-08	
GCGA	3,96	SWI4	*S. cerevisiae*	AT2G14255	-	0,004	595
		ZBED1	*H. sapiens*	-	-	-	
		ZBTB33	*H. sapiens*	AT4G06634	YY1	0,084	
GCGAT	4,49	ZBED1	*H. sapiens*	-	-	-	27
		TOD6	*S. cerevisiae*	AT5G40430	MYB22	0,003	
		DOT6	*S. cerevisiae*	AT2G13960	-	0,011	
TTGTT	2,85	FOXN3	*H.sapiens*	-	-	-	194
		SOX14	*H. sapiens*	AT1G20693	HMGB2	6,00E-04	
		SOX5	*M. musculus*	AT4G23800	3XHMG-BOX2	4,00E-04	

4.4.2 . Développement d'une méthodologie pour valider les 5'-uPLM candidats sélectionnés

Pour valider l'implication des trois 5'-uPLM candidats dans la régulation transcriptionnelle des sous-réseaux dans lesquels ils ont été détectés, j'ai entrepris de réaliser une dissection fonctionnelle des promoteurs des gènes possédant les 5'-uPLM

d'intérêt. Pour cela, plusieurs étapes de caractérisation *in silico* furent nécessaires avant d'entreprendre les expérimentations humides *in planta*.

Dans un premier temps, il a été nécessaire de déterminer les conditions expérimentales, parmi celles présentent dans CATdb, pour lesquelles la présence du 5'-uPLM testé était explicative de la réponse transcriptionnelle des gènes. Sachant que les séquences *cis*-régulatrices fonctionnent en modules (Schmitz *et al.*, 2021), il a été aussi nécessaire de déterminer si le 5'-uPLM était suceptible de fonctionner en collaboration avec un tPLM dans les conditions expérimentales mises en avant. Pour cela, l'ensemble des couples (5'-uPLM, tPLM) a été testé pour chaque condition expérimentale. Une décomposition linéaire des valeurs des logratios a ainsi été réalisée en considérant deux variables explicatives : (1) la présence du 5'-uPLM dans le gène et (2) la présence du tPLM dans le gène.

$$y_i = \beta_0 x_{i1} + \beta_1 x_{i2} + \varepsilon_i$$

où y_i la valeur de logratio du gène i, x_{i1} la présence (1) ou l'absence (0) du uPLM dans le gène i et x_{i2} la présence (1) ou l'absence (0) du tPLM dans le gène i, enfin $i = 1, ..., n$ et n désigne le nombre de gènes dont l'expression est mesurée.

Ainsi, les expériences pour lesquelles les coefficients β_0 et β_1 sont significativement non nuls ont été retenues (p-valeur ajustée avec la procédure de bonferroni pour contrôler le FWER à 5%) car elles suggèrent que le couple (5'-uPLM,tPLM) testé y est explicatif de la réponse transcriptionnelle des gènes.

Dans l'objectif d'identifier une condition expérimentale idéale pour les études de validation humide, il a été nécessaire de filtrer manuellement les expériences identifiées selon plusieurs critères :

- Déterminer s'il était possible de remettre en place facilement le protocole correspondant aux expériences identifiées.
- S'assurer que les gènes ayant le 5'-uPLM soient fortement sur-exprimés et ce, de manière homogène dans la condition de l'expérience, afin d'être en mesure de détecter une différence entre des séquences avec le 5'-uPLM natif et celles ayant le 5'-uPLM muté.
- S'assurer de la reproductibilité de l'expérience afin que les corrélations entre les réplicats biologiques soient fortes.
- Privilégier les expériences ayant permis de construire le réseau de réponse globale aux stress, ce qui n'est pas le cas de l'ensemble des expériences répertoriées dans CATdb.

La combinaison de ces différents critères a permis d'identifier 9 expériences (Table 4.2) et de sélectionner celle conduite par l'équipe de Mathilde Fagard (IJPB) (Moreau *et al.*, 2012) impliquant le uPLM CAATTC du sous-réseau 27. Cette expérience correspond à l'infection (6h après l'inoculation) des feuilles d'*A. thaliana* par *Erwinia amylovora*, une bactérie de la famille des *Enterobacteriaceae* responsable de la maladie du feu bactérien chez les *Rosaceae*. Dans cette condition, les 23 gènes du sous-réseau 27 avec le uPLM CAATTC ont leur expression induite 6

heures après inoculation de la bactérie.

Table 4.2 – Expériences considérées pour la validation des 5'-uPLM candidats. *Les couleurs des noms d'expériences indiquent le niveau de corrélations entre les réplicats biologiques (rouge $\leq$ 0,4; 0,4 $<$ orange $\leq$ 0,6; vert $>$ 0,6; gris : pas de réplicats). Les couleurs des motifs sont relatives aux annotations fonctionnelles des modules dans lesquels les sous-réseaux sont regroupés. Bleu : Traduction; Rouge : Respiration cellulaire; Vert : Photosynthèse. La colonne "Expérience réponse aux stress" indique si l'expérience identifiée a été considéré dans la construction du réseau de réponse globale aux stress. La colonne "Nb tPLM dans le sous-réseau" indique le nombre de tPLM détectés enrichis dans le sous-réseau correspondant. La colonne "Nb de tPLM avec associations significatives pour l'expérience" indique le nombre de tPLM pour lesquels le couple (5'-uPLM,tPLM) s'est révélé significatif dans la première étape de sélection pour l'expérience correspondante.*

uPLM	Sous-réseau	Nb gènes du sous-réseau avec uPLM	Nom expérience	Expérience réponse aux stress	Nb tPLMs dans le sous-réseau	Nb de tPLM avec associations significatives pour l'expérience
GCGAA	3	27/62	sbe1-sbe2 mutant sucrose/mannitol Erwinia	Non Oui Oui	26	5 3 0
GCGAA	45	22/80	Vip1-HA mutant myb77 mutant mpk6 flg22 1h	Oui Non Non	5	2 0 4
CAATTC	**27**	23/111	Erwinia **Erwinia** Erwinia	Oui **Oui** Oui	**10**	3 **2** 4

Une fois la condition expérimentale déterminée, j'ai sélectionné les gènes sur lesquels initier les validations humides afin d'avoir un nombre réaliste de constructions moléculaires à faire, de plantes d'*A. thaliana* à transformer et de tests d'infection à conduire avec *E. amylovora*. Pour cela, j'ai cherché à identifier les gènes les plus fortement exprimés (i) dans la condition contrôle (absence de la bactérie), (ii) dans la condition stressée et (iii) répondant fortement à l'infection (i.e. logratios forts). Ces critères ont permis d'identifier trois gènes. Deux d'entre eux (AT2G30390 et AT1G17220) ont finalement été sélectionnés afin d'avoir un nombre plus limité de tests fonctionnelles à conduire. Pour chacun des gènes, cinq constructions (C1 à C5) ont été réalisées par Christine Horlow (IJPB) (Figure 4.5) :

— C1 : séquence promotrice pleine longueur. Elle sert de contrôle positif et doit induire l'expression de la séquence codante du gène de la β-glucuronidase (GUS) placée en aval et dont le substrat clivé produit un précipité bleu et insoluble qui confère une coloration bleue.

— C2 : séquence courte proche de la fenêtre fonctionnelle du 5'-uPLM mais néanmoins assez grande pour contenir le(s) tPLM susceptibles de fonctionner avec le uPLM sur la base des tests statistiques avec le couple (uPLM, tPLM). Elle doit également induire l'expression du gène rapporteur et constitue un autre contrôle positif.

— C3 : séquence courte similaire à C2 avec le uPLM muté.
— C4 : séquence courte similaire à C2 avec le(s) tPLM muté(s).
— C5 : séquence courte similaire à C2 avec le uPLM et le(s) tPLM mutés.

Des plantes d'*A. thaliana* ont ensuite été transformées mi-octobre 2022 avec chacune des constructions afin d'obtenir des grains T1. Les plantes T1 transformées seront inoculées fin décembre avec la bactérie pendant 6 heures par l'équipe de Mathilde Fagard avec laquelle une collaboration a été initiée. Pour chacune des constructions, l'activité du gène rapporteur GUS sera ensuite mesurée par (1) colorimétrie, (2) dosage enzymatique et (3) détermination de l'abondance relative du transcrit *GUS* par RT-PCR quantitative. Si le 5'-uPLM marche en combinaison avec le tPLM, sa mutation peut n'avoir aucun effet. Les résultats obtenus avec les plantes transformées avec la construction C5 et la comparaison avec les plantes transformées avec les constructions C3 et C4 permettront de déterminer si le 5'-uPLM a un effet cis-régulateur et si cet effet requiert ou pas la présence du(des) tPLM.

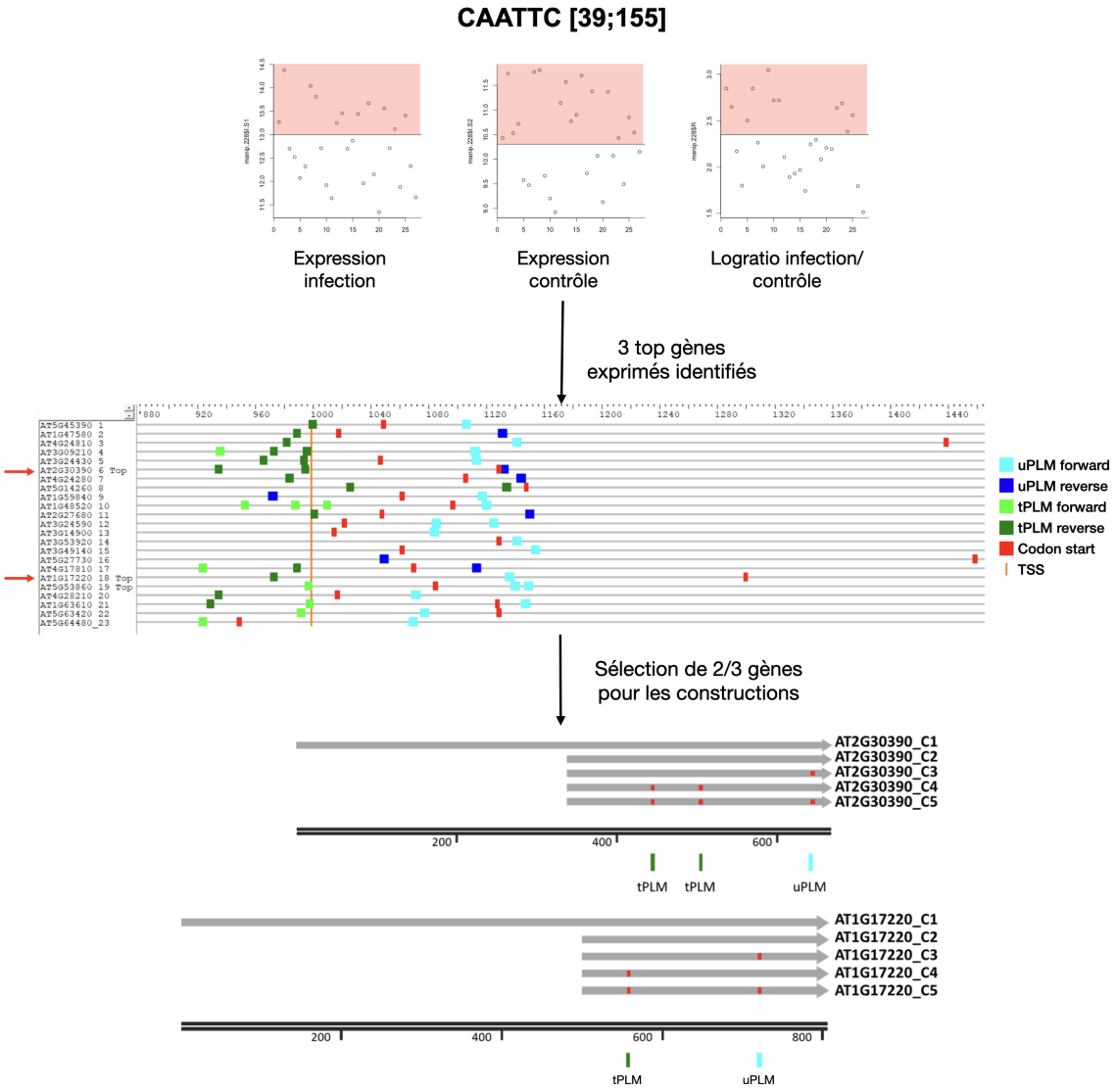


Figure 4.5 – Sélection et construction des séquences pour valider le 5'-uPLM CAATTC. *Les séquences promotrices initiales sélectionnées sont basées sur les niveaux d'expression des gènes en condition contrôle (Expression contrôle), en réponse à l'infection par E. amylovora (Expression infection) et à la différence d'expression entre les deux conditions (Logratio infection/contôle). Les gènes identifiés sont appelés "top gènes". Deux gènes parmi les trois top gènes ont ensuite été sélectionnés sur la base des répartitions des 5'-uPLM et tPLM le long des séquences promotrices. Les constructions finales réalisées se basent sur les promoteurs des gènes AT2G30390 et AT1G17220 et consistent en une séquence pleine longueur (C1), une séquence courte proche de la fenêtre fonctionnelle du 5'-uPLM (C2), la séquence courte avec le 5'-uPLM muté (C3), la séquence courte avec le(s) tPLM muté(s) (C4) et la séquence courte avec le 5'-uPLM et le(s) tPLM mutés.*

4.5 . Conclusion et perspectives

Dans l'objectif de mieux caractériser l'implication des séquences *cis*-régulatrices proximales dans la réponse aux stress, j'ai réalisé une détection *de novo* de PLM sur un réseau de gènes représentant la réponse transcriptionnelle d'*A. thaliana* face aux stress. Cette détection m'a permis d'inférer des modules de régulation "TF-sous-réseau" impliqués dans cette réponse, ainsi que d'identifier de nouvelles séquences *cis*-régulatrices putatives pour lesquelles des validations expérimentales sont en cours.

L'identification de tPLM a permis d'inférer l'implication de plusieurs centaines de TF dans la régulation du réseau. Les familles majoritaires sont connues pour intervenir dans les réponses aux stress biotiques et abiotiques (Hong, 2016), ce qui appuie les prédictions initiales conduites par Marie-Laure Martin et Etienne Delannoy. Cependant, ce travail doit être poursuivi pour appuyer les interactions TF-tPLM. À court terme, il s'agira de comparer les sites tPLM identifiés avec les données expérimentales de ChIP-seq et DAP-seq répertoriées dans la base de données ReMap (Hammal *et al.*, 2022). La ressource génomique que constitue l'ensemble des TF prédits pourra ensuite être exploitée afin d'identifier de nouveaux acteurs impliqués dans plusieurs processus biologiques. Il sera ainsi intéressant d'étudier plus spécifiquement l'une des voies présente dans le réseau de réponse globale aux stress. Le choix pourrait se porter sur la voie associée à l'organisation de la paroi cellulaire identifiée dans les deux régions proximales.

L'identification de uPLM a aussi permis de développer une méthodologie pour identifier des séquences *cis*-régulatrices putatives dans la région 5'-proximale et développer une approche pour les valider en partant de prédictions réalisées sur des données intégrées. Ce travail devra être étendu aux 3'-uPLM afin de mettre en évidence des marqueurs de co-régulation originaux au niveau de cette région.

5 - Discussion

5.1 . Propriétés générales des régions proximales de 20 plantes à fleurs

5.1.1 . Une organisation des PLM globalement conservée dans les deux régions proximales

Les analyses comparatives réalisées avec les 20 plantes à fleurs sélectionnées (chapitres 2 et 3) ont montré que certaines espèces ont beaucoup moins de PLM que d'autres. De manière intéressante, cette différence ne s'explique pas par le nombre de séquences étudiées étant donné que ce dernier n'est pas corrélé au nombre de PLM identifiés (Table 3.1, annexe B). Cependant, quelque soit l'espèce considérée, le nombre de PLM identifiés dans la région 5'proximale est du même ordre de grandeur que celui observé dans la région 3'-proximale (Table 3.1). Par ailleurs, l'analyse comparative des PLM d'*A. thaliana* et de *Z. mays* a montré que les 3'-PLM sont plus conservés que les 5'-PLM. L'ensemble de ces résultats suggère donc que la région 3'-proximale a un rôle structural important chez les plantes à fleurs.

Les analyses comparatives ont aussi montré que la distribution des PLM est relativement similaire pour 78% des espèces étudiées. Dans la région 5'-proximale, des pics ont ainsi été globalement observés au niveau des boîtes TATA (-30 bases) et du TSS, alors que dans la région 3'proximale, les pics ont été observés au niveau des signaux FUE/NUE (-10 à -30 bases) et du TTS. Cependant, cette distribution n'est pas partagée par l'ensemble des espèces, comme l'illustre la distribution des 5'-PLM chez *C. maxima* (Figure 5.1). Ces différences peuvent être d'origine biologique ou technique :

- D'un point de vue technique, il est possible que la qualité d'annotation des TSS/TTS des génomes concernés soit moins bonne que celle des autres espèces. Étant donné que les sur-représentations locales déterminées par la méthode PLMdetect repose sur la position des TSS/TTS, une mauvaise annotation de ces derniers impacterait l'identification des PLM. Pour évaluer rapidement la qualité d'annotation des TSS/TTS, il serait envisageable de caractériser la qualité (score élevé et fenêtre fonctionnelle petite ; Figure introduction 1.6) des PLM associés aux boîtes TATA et aux sites de polyadénylation. Si la position de ces derniers varie fortement, cela suggère que les bornes des gènes ne sont pas correctement annotées.
- D'un point de vue biologique, il est possible que les espèces qui diffèrent soient plus proches phylogénétiquement les unes des autres et que les différences observées soient conservées entres elles (Jegga and Aronow, 2013; Burgess and Freeling, 2014; Van de Velde *et al.*, 2016). Pour appuyer cette hypothèse, il sera nécessaire de répertorier l'ensemble des espèces présen-

tant ces différences et d'identifier les familles auxquelles elles appartiennent. Il est aussi possible que les différences observées soient dues à la composition nucléotidique des génomes de ces espèces qui se distinguerait de celle des autres espèces. Néanmoins, cette hypothèse semble peu probable dans la mesure où une partie des Poacées, telles que *B. distachyon* ou *O. Sativa*, a un génome globalement riche en GC mais présente des distributions proches de celles observées pour les Brassicacées, telles que *A. thaliana* et *A. lyrata* qui ont des génomes riches en AT (Singh *et al.*, 2016). Enfin, il est envisageable que cette modification de l'organisation du contenu en PLM soit causée par des contenus en éléments transposables (TE) différents au sein des régions proximales. En effet, les espèces de plantes à fleurs présentent une grande variabilité de leur contenu en TE et les portions du génome occupées par ces derniers peuvent varier de 20% à plus de 80% pour certaines espèces (Oliver *et al.*, 2013). Une partie de ces TE est également décrite pour s'insérer dans les régions proximales des gènes et influer sur le contenu en séquences *cis*-régulatrices (Quesneville, 2020). Ainsi, pour évaluer cette possibilité, il sera nécessaire de caractériser les contenus en TE des régions proximales des espèces présentant des distributions de PLM différentes des autres. Cependant, cette hypothèse semble peu probable dans la mesure où des distributions similaires de PLM ont été obtenues pour *A. thaliana* et *Z. mays* (Rozière *et al.*, 2022b), deux espèces ayant des contenus en TE fortement contrastés (Oliver *et al.*, 2013).

5.1.2 . Relation entre les PLM et les phénotypes

L'étude comparative des PLM chez les 20 espèces de plantes à fleurs a révélé que plus de la moitié d'entres eux sont spécifiques d'une seule espèce (Figure 3.3). Ce résultat surprenant ouvre des pistes de réflexion quant à l'implication de ces PLM dans la variabilité phénotypique des espèces étudiées. À court terme, il serait intéressant de poursuivre la caractérisation de ces PLM (1) en réalisant des analyses d'enrichissements fonctionnels afin de déterminer les processus biologiques associés et (2) en déterminant les entités moléculaires impliquées dans la régulation de ces motifs (TFBS, petits ARN) afin d'identifier des acteurs *trans* putatifs impliqués dans ces processus. À moyen terme, il serait intéressant d'exploiter les PLM comme des marqueurs moléculaires associés à la variabilité de caractères phénotypiques. Des discussions ont été initiées, au cours de ma thèse, afin de réfléchir à l'élaboration de méthodes d'associations génomiques (GWAS) (Alseekh *et al.*, 2021) prenant en considération des PLM à la place de "Single Nucleotide Polymorphism" (SNP). Dans la mesure où le PLM peut rendre compte d'un environnement ou d'un caractère particulier, sa prise en compte pourraient améliorer les prédictions, comme cela a déjà été montré pour la structure des réseaux de régulation (Dong *et al.*, 2012; Wade *et al.*, 2022).

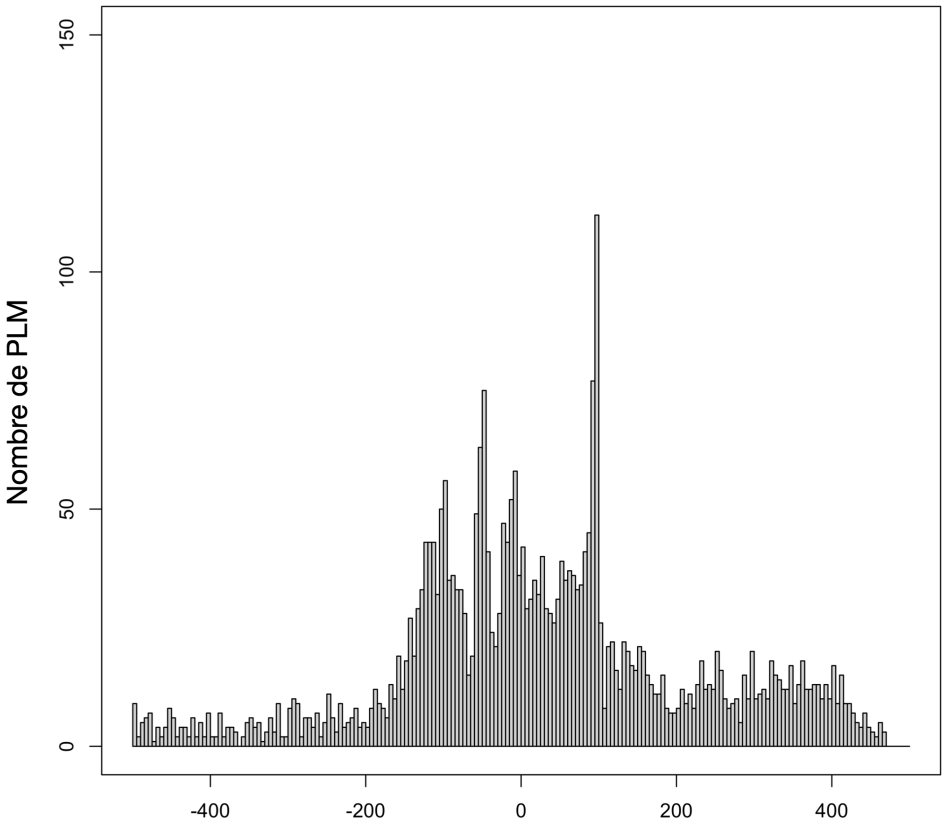


Figure 5.1 – Distribution des 5'-PLM par rapport au TSS chez *C. maxima*.

5.1.3 . Appliquer PLMdetect à des groupes d'orthologues pour poursuivre la caractérisation des régions proximales

Pour étudier la conservation des PLM chez les plantes à fleurs, je me suis placé à l'échelle du génome de chacune des espèces sélectionnées. Cette détection s'est révélée remarquable pour la quantité d'informations apportée, ainsi que pour sa qualité, en permettant l'identification de signaux spécifiques à de petits ensembles de gènes. Ainsi, dans la région 5'-proximale d'*A. thaliana*, le nombre de gènes médian associé à un PLM est de seulement 95 alors que 19 736 gènes ont été fournis en entrée. Toutefois, il est certain que certains signaux ne sont pas identifiés par cette détection "à l'échelle du génome" : ils sont en effet noyés dans le "bruit génomique". Ainsi, seuls 35% (229/659) des 5'-PLM et 45% (50/110) des 3'-PLM identifiés dans les sous-réseaux du réseau de réponse globale aux stress d'*A. thaliana* sont communs aux 5'-PLM et 3'-PLM identifiés à l'échelle du génome. Il est exclu que tous les PLM identifiés à l'échelle du génome et non retrouvés à l'échelle du réseau soient des faux-positifs. En effet, la comparaison des sites tPLM avec les données expérimentales de ChIP-seq et DAP-seq (Rozière *et al.*, 2022b; Hammal *et al.*, 2022) a pu valider 579 5'-tPLM et 374 3'-tPLM d'*A. thaliana* et parmi eux 98% (930/953) ne sont pas des tPLM au sein du réseau.

Dans ce contexte, il pourrait être pertinent de reprendre les analyses conduites à l'échelle du génome en se plaçant au sein de sous-ensemble de gènes orthologues afin d'identifier des signaux "indétectables" à l'échelle génomique. Ce type de détection permettrait aussi de révéler des séquences conservées dans les régions proximales des gènes orthologues et par conséquent indétectables dans des analyses espèce par espèce. À court/moyen terme, il est donc envisagé d'associer la base de données Plant-PLMview à des base de données répertoriant des groupes d'orthologues de plantes telles que Phytozome(Goodstein *et al.*, 2012), OrthoDB (Zdobnov *et al.*, 2021), PLAZA (Van Bel *et al.*, 2022) ou encore SyntenyViewer (non publiée) qui répertorie des orthogroupes comprenant les gènes des 20 espèces traitées au cours de cette thèse.

5.2 . Caractéristiques des PLM des régions proximales des plantes

Cette thèse a permis d'apporter des éclaircissements sur la diversité des acteurs de la régulation en *cis* dans les régions proximale des gènes des plantes.

5.2.1 . L'identification de tPLM ouvrent de nouvelles perspectives mécanistiques pour les TF

Comme attendu, les analyses conduites au cours de cette thèse ont permis d'identifier de nombreux TFBS dans les régions proximales. Bien que déjà décrite dans la littérature, la présence de nombreux TFBS dans la région 3'-proximale soulève la question du rôle mécanistique des TF dans la régulation de l'expression des gènes en 3'. Plusieurs hypothèses peuvent être émises :

— La première est liée aux rôles des TF dans le choix du site de polyadényla-

tion alternatif (APA) du transcrit. En se fixant à proximité du TTS, les TF peuvent en effet stopper l'élongation et ainsi conduire le complexe CPMC à choisir un APA. À l'inverse, les TF qui se fixent plus en aval du gène, peuvent maintenir l'accessibilité de l'ADN et permettre ainsi le maintien de l'élongation et le choix d'un APA plus en aval. Cette hypothèse mécanistique est étayée par des travaux récents démontrant le rôle d'un "enhancer" qui interagit avec la partie terminatrice du gène cible et induit la production d'un isoforme 3'UTR (Kwon *et al.*, 2022). Ainsi, le grand nombre de tPLM identifiés dans la région 3'-proximale (Rozière *et al.*, 2022b) suggère que ce mécanisme de régulation pourrait impacter l'expression d'une large partie des gènes de plantes.

— Il est aussi possible que le rôle des TF en 3' soit lié au phénomène de "gene-looping" qui est possible grâce à l'interaction promoteur-terminateur (Al-Husini *et al.*, 2020) et permet une transcription rapide du gène cible. Ce phénomène est aujourd'hui envisageable dans la mesure où le TF général TFIIB est connu pour interagir avec une partie du complexe CPMC (Wang *et al.*, 2010). De part leur rôle dans la conformation de la chromatine et des repliements 3D permettant les interactions "enhancer/silencer-promoteur" (Schmitz *et al.*, 2021), il est possible que les TF présents dans les promoteurs et les terminateurs soient aussi impliqués dans le phénomène de "gene-looping".

— Enfin, il est probable que les TF identifiés dans la partie terminatrice d'un gène donné puissent jouer un rôle dans le promoteur d'un gène situé en aval. En effet, bien que ce phénomène n'ait pas encore été décrit chez les plantes, plusieurs arguments plaident en sa faveur, comme la proximité de nombreux gènes (Berardini *et al.*, 2015), l'existence de promoteurs partagés (Yang *et al.*, 2008) et la complexité des mécanismes *cis*-régulateurs (Schmitz *et al.*, 2021).

L'ensemble de ces hypothèses ne sont pas antinomiques. En effet, vue la complexité d'action des séquences *cis*-régulatrices et l'existence d'éléments multifonctionnels (Schmitz *et al.*, 2021) pouvant être des activateurs ou des répresseurs, une même région 3'-proximale peut remplir l'ensemble des rôles selon les acteurs en *trans* qui s'y fixent.

5.2.2 . L'identification de miPLM suggère l'intervention de nombreux microARN au niveau transcriptionnel

Les analyses conduites suggèrent également que de nombreux microARN interviennent au niveau transcriptionnel. Aujourd'hui, les interactions ADN-petits ARN impliquent principalement les petits ARN interférents (siARN) qui sont capables d'initier la voie "RNA-directed DNA methylation" (RdDM) (Erdmann and Picard, 2020). Cette voie spécifique des plantes induit la méthylation de l'ADN. Les microARN sont quant à eux surtout connus pour leur rôle post-transcriptionnel et leur implication dans la dégradation des transcrits ciblés. Néanmoins, il existe des

exemples de microARN capables d'interagir avec les promoteurs et de moduler (positivement ou négativement) l'expression de gènes (Kim *et al.*, 2008; Place *et al.*, 2008; Wu *et al.*, 2010; Axtell, 2013). Les résultats obtenus au cours de cette thèse laissent donc penser que les interactions ADN-microARN sont plus fréquentes que ne le suggère le nombre d'exemples que nous avons à ce jour. Il est notable que des développements expérimentaux basés sur une technique de capture d'affinité de microARN biotinylé à l'aide de perles de streptavidine, ont récemment été mis en place pour identifier des cibles génomiques de microARN chez l'homme (Xun *et al.*, 2019). L'adaptation et le déploiement de cette approche à des organismes végétaux permettraient de conforter ou non les miPLM.

5.2.3 . Élucider les mécanismes sous-jacents à la présence des uPLM

J'ai pu mettre en avant que près de 80% des PLM identifiés sont encore à ce jour non assignés (uPLM). Une partie de ces uPLM a pu être appuyée avec des données expérimentales de MOA-seq (Rozière *et al.*, 2022b). Par ailleurs, les caractérisations *in silico* réalisées à partir des sous-réseaux du réseau de réponse globale aux stress chez *A. thaliana*, ont permis de mettre en avant que certains uPLM sont explicatifs d'une différence d'expression observées dans de nombreuses conditions expérimentales. Cependant, les mécanismes sous-jacents à la présence de ces uPLM restent encore inconnues. Plusieurs hypothèses peuvent être émises :

— Il est vraisemblable qu'une partie des uPLM identifiés soient des TFBS encore non caractérisés. En effet, la majorité des TF n'ont pas encore de données expérimentales caractérisant leurs sites de fixation dans les bases de données. Ainsi, si on prend l'exemple d'*A. thaliana*, seuls 372 TF sur les 2000 que contient cette espèce, ont jusqu'ici été caractérisés par des données de ChIP-seq ou DAP-seq (Hammal *et al.*, 2022). Dès lors, l'ajout de données expérimentales supplémentaires (Tu *et al.*, 2020) à la base de donnée JASPAR (Fornes *et al.*, 2020) a permis d'augmenter de 16% le nombre de tPLM identifiés chez *A. thaliana* (Rozière *et al.*, 2022b). Par ailleurs, la correspondance entre des motifs MOA et des uPLM constitue un argument supplémentaire pour soutenir cette hypothèse.

— Les PLM détectés dans les régions transcrites pourraient aussi être reconnus par des protéines RBP ("RNA Binding Protein") et donc impliqués dans la régulation post-transcriptionnelle. Les protéines RBP interviennent dans la régulation de nombreux processus biologiques, telles que le développement et la croissance des plantes ou la réponse aux stress abiotiques (Lee and Kang, 2016). Aujourd'hui, des techniques NGS, telles que le CLIP-seq (Van Nostrand *et al.*, 2016), permettent d'identifier les interactions protéines-ARN et les sites de liaison impliqués. À l'instar de la base de données JASPAR (Castro-Mondragon *et al.*, 2022) pour les TFBS, le développement de bases de données intégrant les sites de fixation de protéines RBP sur l'ARN messager serait d'un grand intérêt pour mieux caractériser les PLM impliqués dans la régulation post-transcriptionnelle.

— Il est également possible que les uPLM soient des séquences indispensables à la bonne fixation d'un TF sur un TFBS proche, c'est à dire des séquences "flanking anchor" (Stringham *et al.*, 2013). Pour appuyer cette hypothèse, l'étude de la co-occurence de PLM se révèle d'un grand intérêt, en permettant d'identifier des modules de PLM spécifiques, ainsi que des associations uPLM/tPLM fonctionnelles. Ainsi, les résultats obtenus dans le Chapitre 4 montrent que des associations uPLM-tPLM peuvent être explicatives de la réponse transcriptionnelle des gènes qui les possèdent. À court terme, il serait intéressant de poursuivre ces analyses de co-occurences de PLM pour participer à l'élucidation des mécanismes liés aux uPLM.
— Enfin, une partie des uPLM pourrait correspondre à des oligonucléotides structurant la conformation 3D des régions proximales des gènes. Des approches d'apprentissage profond, telles que les "convolutional neural network" (CNN), ont été entraînées afin de prédire la structure 3D de la chromatine sur la base de séquences d'ADN (Piecyk *et al.*, 2022). Les données d'entraînement se composent dans la grande majorité des cas de Hi-C (Belton *et al.*, 2012) et peuvent être complétées par des données annexes comme du ChIP-seq (Barski *et al.*, 2007), de l'ATAC-seq (Buenrostro *et al.*, 2015), du DNase-seq (Boyle *et al.*, 2008), des données de méthylomes (Lay *et al.*, 2018) ou bien simplement l'annotation structurale du génome considéré. Ces modèles servent à déterminer les régions de la chromatine en contact les unes des autres, et notamment les interactions "enhancer-promoteurs" (Schmitz *et al.*, 2021). Ces approches proposent ainsi des solutions pour évaluer l'impact de SNP dans la conformation chromatinienne et les interactions qui en découlent. Pour cela, les modèles sont entraînés pour prédire la structure chromatinienne en réalisant de la mutagénèse *in silico*, c'est à dire en considérant chaque modalité du polymorphisme. Si la prédiction d'interactions diffèrent selon la modalité, alors cela suggère que le variant est en partie responsable de la structure 3D de la chromatine. Il serait donc très intéressant d'utiliser ce procédé pour caractériser les uPLM et leur rôle avec les éléments régulateurs distaux, en appliquant la démarche aux uPLM à la place des SNP.

5.3 . Rôle des PLM dans la réponse aux stress chez *A. thaliana*

5.3.1 . Les miPLM et les 3'-PLM sont peu impliqués dans la co-régulation des sous-réseaux de gènes

Au cours de l'étude de l'implication des séquences *cis*-régulatrices proximales dans la réponse commune aux stress (chapitre 4), il a été possible d'identifier des centaines de PLM enrichis pour des sous-réseaux de gènes co-exprimés. De manière intéressante, cette étude de cas a révélé des différences importantes avec les recherches de PLM à l'échelle des génomes (Rozière *et al.*, 2022b).

Tout d'abord, contrairement à l'étude réalisée au niveau génomique, aucun miPLM (ou presque) n'a été impliqués dans la co-expression des sous-réseaux. Deux hypothèses non antinomiques peuvent être avancées pour expliquer cela. Il est possible que la régulation *via* les microARN soit mis en place "gène par gène", c'est-à-dire n'implique qu'un seul gène et pas un ensemble de gènes en simultané. Une alternative est que si ce type de régulation est impliquée dans la co-expression de gènes, cela ne soit pas en réponse aux stress.

Par ailleurs, le nombre de 5'-PLM enrichis détectés au sein des sous-réseaux est bien plus élevé que le nombre de 3'-PLM enrichis. De manière corrélée, le nombre de TF inférés pour être des régulateurs en *trans* dans la région 5'-proximale est bien supérieur à celui obtenu dans la région 3'-proximale. Ainsi, la région 5'-proximale est plus fortement impliquée dans la co-expression des sous-réseaux en réponse aux stress que la région 3'-proximale. Ces résultats laissent penser que la région 3'-proximale est le lieu de nombreux événements de régulation en *cis* qui n'interviennent que peu dans la co-régulation d'ensemble de gènes. Néanmoins, il est intéressant de noter que certains sous-réseaux ne contiennent que des 3'-PLM enrichis. Dés lors, la région 3'-proximale semble majoritairement impliquée dans des régulations intervenant "gène à gène" et, plus rarement, dans la co-régulation d'ensemble de gènes.

5.3.2 . Le pipeline de validation des uPLM candidats : des perspectives d'application au-delà du réseau de réponse aux stress

La validation expérimentale de données prédites est toujours une tâche difficile. Ceci est d'autant plus vrai lorsque les prédictions sont réalisées à partir de données intégrant de nombreuses expériences. Dans le cas de la recherche de séquences *cis*-régulatrices, la validation est encore plus complexe car elle met en jeu des caractéristiques tissulaires, développementales ou liées au stress subit (Hernandez-Garcia and Finer, 2014; Schmitz *et al.*, 2021). En outre, les séquences *cis*-régulatrices sont largement connues pour fonctionner en combinaison (Schmitz *et al.*, 2021). La validation humide des uPLM candidats (Chapitre 4) a donc demandé de sélectionner des conditions expérimentales idéales en prenant en considération l'ensemble des caractéristiques pré-citées. Les résultats obtenus ont permis de mettre en évidence des couples (uPLM,tPLM) dont la présence est explicative de différences d'expression dans certaines conditions expérimentales, suggérant ainsi une asso-

ciation fonctionnelle entre la paire de PLM identifiée. Cette étude de cas souligne l'importance d'étudier les PLM en modules et incite à poursuivre ces études de co-occurences fonctionnelles à l'échelle des génomes.

Le pipeline développé pour valider les uPLM candidats ouvre aussi des perspectives pour valider les prédictions génomiques (Rozière *et al.*, 2022b). La validation humide de ces prédictions est une tâche complexe dans la mesure où il n'y a pas de connaissance *a priori* sur les conditions expérimentales dans lesquelles la séquence *cis*-régulatrice putative est active. Ainsi, l'application de ce pipeline pour l'ensemble des PLM identifiés peut permettre de déterminer des conditions expérimentales dans lesquels la présence du PLM est explicative d'une différence d'expression et ainsi, par extension, de sélectionner des conditions idéales à reproduire pour valider la prédiction.

Au-delà de l'étude des PLM, la méthodologie développée présente des perspectives d'applications plus larges :

- Cette démarche peut s'appliquer à tout élément *cis*-régulateur proximal identifié via d'autres méthodes prédictives et dont l'impact transcriptionnel n'est pas connu.
- Cette approche peut aussi être pertinente dans le cas de l'étude d'un TF spécifique dont on souhaite identifier l'ensemble des interactions génomiques proximales directes. Cela permet d'inférer les interactions fonctionnelles et de déterminer les conditions expérimentales dans lesquelles l'interaction semble avoir un impact transcriptionnel.
- Enfin, il serait envisageable d'appliquer cette méthodologie aux séquences *cis*-régulatrices distales et ainsi déterminer *in silico* dans quelles conditions les interactions "enhancer/promoteur" sont fonctionnelles. Pour cela, il faudrait réaliser des ajustements de la méthode en remplaçant les variables explicatives uPLM et tPLM dans le modèle section 4.4.2 par les "enhancers" eux-mêmes ou les TFBS qu'ils contiennent. Puis, il serait nécessaire d'inférer les cibles de l'enhancer testé ce qui pourrait être fait (1) à partir de données expérimentales, telles que du Hi-C (Belton *et al.*, 2012), (2) sur la base d'une limite de distance entre l'enhancer et les gènes cibles potentiels ou (3) sur la base des probabilités d'interactions obtenues à partir des méthodes d'apprentissage profond évoquées dans la section 5.2.3. Cette perspective de recherche permettrait d'accélérer grandement la caractérisation d'éléments *cis*-régulateurs distaux et d'apporter de nouvelles pistes de compréhension de la régulation de l'expression génique.

6 - Conclusion générale

Durant ma thèse, j'ai cherché à mieux comprendre la structure et la fonction des séquences *cis*-régulatrices présentes dans les régions proximales en 5' et 3' des gènes de 20 espèces de plantes à fleurs.

Dans un premier temps, j'ai étendu la méthode PLMdetect pour identifier *de novo* l'ensemble des PLM présents dans les régions proximales des 20 espèces sélectionnées. J'ai montré que les 3'-PLM étaient aussi abondants que les 5'-PLM chez toutes les espèces étudiées (Chapitres 2 et 3). L'analyse comparative des régions proximales d'*A. thaliana* et de *Z. mays* a aussi révélé que les 3'-PLM étaient plus conservés que les 5'-PLM chez ces deux espèces (Chapitre 2). L'ensemble de ces résultats a donc permis de conforter l'importance de la région 3'-proximale dans la structure des génomes des plantes à fleurs. Par ailleurs, les analyses préliminaires conduites avec les 20 espèces ont montré que plus de la moitié des PLM identifiés sont spécifiques d'une espèce donnée, suggérant qu'une partie au moins de ces PLM correspondrait à des signaux moléculaires explicatifs des variations phénotypiques propre à chaque espèce. Enfin, la quantité de données générées a conduit au développement de la base de données Plant-PLMview qui met la méthode PLMdetect à la disposition de l'ensemble de la communauté scientifique (Chapitre 3). Cette base de données permet à l'utilisateur de rechercher des séquences *cis*-régulatrices préférentiellement localisées dans un groupe de gènes d'intérêt et d'obtenir une carte pour visualiser les modules de PLM potentiellement impliqués dans la co-régulation du groupe.

Dans un deuxième temps, j'ai caractérisé les PLM identifiés chez *A. thaliana* et *Z. mays* (Chapitres 2 et 4). J'ai ainsi mis en évidence trois types de PLM dans chaque région proximale étudiée : (1) des TFBS, (2) des PLM présentant des homologies de séquences avec des microARN et (3) des séquences *cis*-régulatrices putatives qui constituent 79% des PLM identifiés et dont une partie est supportée par des données expérimentales de type MOA-seq obtenues chez *Z. mays* (Savadel *et al.*, 2021). Par ailleurs, j'ai développé une méthodologie pour identifier *in silico* des couples (uPLM, tPLM) et déterminer les conditions expérimentales dans lesquelles ils sont supposés être fonctionnels (Chapitre 4).

Dans un troisième temps, je me suis intéressé au rôle des éléments *cis*-régulateurs proximaux dans la réponse commune aux stress chez *A. thaliana* (Chapitre 4). Pour cela, je me suis basé sur une ressource génomique consistant en un réseau de gènes composé de sous-réseaux de co-expression. J'ai identifié l'ensemble des PLM enrichis et les ai caractérisés en suivant le protocole mis en place pour l'étude génomique (Chapitre 2). J'ai montré qu'une grande partie des PLM identifiés sont des tPLM qui ont permis d'inférer plus de 250 TF impliqués dans la régulation du réseau de réponse commune aux stress. J'ai aussi mis en évidence que la majorité des PLM sont des uPLM susceptibles d'être des séquences *cis*-régulatrices puta-

tives. Ce travail a également permis de poser les bases d'expériences *in planta* pour valider un uPLM candidat dans la réponse d'*A. thaliana* à *E. amylovora*.

Enfin, cette thèse ouvre de nouvelles perspectives pour comprendre la régulation de l'expression des gènes au niveau des régions distales. En effet, aujourd'hui, les analyses des séquences *cis*-régulatrices distales se focalisent principalement sur les marques épigénétiques et les TF que l'on peut y retrouver (Oka *et al.*, 2017). Il est vraisemblable que les acteurs que j'ai identifiés puissent être impliqués dans la modulation de l'activité des éléments distaux. À titre d'exemple, l'identification de séquences présentant des homologies avec de petits ARN dans ces séquences *cis*-régulatrices distales pourraient expliquer le phénomène de "silencing" de ces dernières *via* leur méthylation (Erdmann and Picard, 2020). Il en est de même pour les uPLM, dont les mécanismes sous-jacents leur présence, ne sont pas encore caractérisés. Une partie de ces uPLM pourraient correspondre à des acteurs dans les séquences *cis*-régulatrices distales ou encore être impliqués dans l'interaction "régions proximales-éléments distaux" (Schmitz *et al.*, 2021).

Bibliographie

Al-Husini, N., Medler, S., and Ansari, A. (2020). Crosstalk of promoter and terminator during RNA polymerase II transcription cycle. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1863**(12), 194657.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., Aganezov, S., Ranallo-Benavidez, T. R., Lemmon, Z. H., Kim, J., Robitaille, G., Kramer, M., Goodwin, S., McCombie, W. R., Hutton, S., Van Eck, J., Gillis, J., Eshed, Y., Sedlazeck, F. J., van der Knaap, E., Schatz, M. C., and Lippman, Z. B. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**(1), 145–161.e23.

Alseekh, S., Kostova, D., Bulut, M., and Fernie, A. R. (2021). Genome-wide association studies : assessing trait characteristics in model and crop plants. *Cellular and molecular life sciences : CMLS*, **78**(15), 5743–5754.

Axtell, M. J. (2013). Classification and Comparison of Small RNAs from Plants. *Annual Review of Plant Biology*, **64**(1), 137–159. _eprint : https ://doi.org/10.1146/annurev-arplant-050312-120043.

Azodi, C. B., Lloyd, J. P., and Shiu, S.-H. (2020). The cis-regulatory codes of response to combined heat and drought stress in Arabidopsis thaliana. *NAR Genomics and Bioinformatics*, **2**(3).

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, **2**, 28–36.

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 Research*, **37**(Web Server issue), W202–208.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, **129**(4), 823–837.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi–C : A comprehensive technique to capture the conformation of genomes. *Methods*, **58**(3), 268–276.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The arabidopsis information resource : Making and mining the "gold standard" annotated reference plant genome : Tair : Making and Mining the "Gold Standard" Plant Genome. *genesis*, **53**(8), 474–485.

Berger, M. F. and Bulyk, M. L. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods in Molecular Biology (Clifton, N.J.)*, **338**, 245–260.

Bernard, V., Lecharny, A., and Brunaud, V. (2010a). Improved detection of motifs with preferential location in promoters. *Genome*, **53**(9), 739–752.

Bernard, V., Brunaud, V., and Lecharny, A. (2010b). TC-motifs at the TATA-box expected position in plant genes : a novel class of motifs involved in the transcription regulation. *BMC Genomics*, **11**(1), 166.

Bernardes, W. S. and Menossi, M. (2020). Plant 3' Regulatory Regions From mRNA-Encoding Genes and Their Uses to Modulate Expression. *Frontiers in Plant Science*, **11**, 1252.

Bi, W., Wu, L., Coustry, F., Crombrugghe, B. d., and Maity, S. N. (1997). DNA Binding Specificity of the CCAAT-binding Factor CBF/NF-Y *. *Journal of Biological Chemistry*, **272**(42), 26562–26572. Publisher : Elsevier.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**(2), 311–322.

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. *Nature Methods*, **10**(12), 1213–1218. Number : 12 Publisher : Nature Publishing Group.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq : A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, **109**, 21.29.1–21.29.9.

Bueso, E., Muñoz-Bertomeu, J., Campos, F., Brunaud, V., Martínez, L., Sayas, E., Ballester, P., Yenush, L., and Serrano, R. (2014). ARABIDOPSIS THALIANA HOMEOBOX25 Uncovers a Role for Gibberellins in Seed Longevity. *Plant Physiology*, **164**(2), 999–1010.

Bueso, E., Muñoz-Bertomeu, J., Campos, F., Martínez, C., Tello, C., Martínez-Almonacid, I., Ballester, P., Simón-Moya, M., Brunaud, V., Yenush, L., Ferrándiz, C., and Serrano, R. (2016). Arabidopsis COGWHEEL1 links light perception

and gibberellins with seed tolerance to deterioration. *The Plant Journal*, **87**(6), 583–596. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13220.

Burgess, D. and Freeling, M. (2014). The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates. *The Plant Cell*, **26**(3), 946–961.

Burgess, S., Reyna-Llorens, I., Stevenson, S., Singh, P., Jaeger, K., and Hibberd, J. (2019). Genome-Wide Transcription Factor Binding in Leaves from C 3 and C 4 Grasses. *The Plant Cell*, **31**, 2297–2314.

Burke, T. W. and Kadonaga, J. T. (1996). Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes & Development*, **10**(6), 711–724. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Calatayud, P.-A., Garrec, J.-P., and Nicole, M. (2013). Chapitre 14. Adaptation des plantes aux stress environnementaux. In P.-A. Calatayud, F. Marion-Poll, N. Sauvion, and D. Thiéry, editors, *Interactions insectes-plantes*, pages 229–245. IRD Éditions.

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 Research*, **45**(13), e119.

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2022). JASPAR 2022 : the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **50**(D1), D165–D173.

Charoensawan, V., Wilson, D., and Teichmann, S. A. (2010). Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Research*, **38**(21), 7364–7377.

Chen, Z.-Y., Guo, X.-J., Chen, Z.-X., Chen, W.-Y., and Wang, J.-R. (2017). Identification and positional distribution analysis of transcription factor binding sites for genes from the wheat fl-cDNA sequences. *Bioscience, Biotechnology, and Biochemistry*, **81**(6), 1125–1135. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/09168451.2017.1295803.

Choudhary, P. and Muthamilarasan, M. (2022). Modulating physiological and transcriptional regulatory mechanisms for enhanced climate resilience in cereal crops. *Journal of Plant Physiology*, **278**, 153815.

Chow, C.-N., Lee, T.-Y., Hung, Y.-C., Li, G.-Z., Tseng, K.-C., Liu, Y.-H., Kuo, P.-L., Zheng, H.-Q., and Chang, W.-C. (2019). PlantPAN3.0 : a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Research*, **47**(D1), D1155–D1163.

Couvillion, M., Harlen, K. M., Lachance, K. C., Trotta, K. L., Smith, E., Brion, C., Smalec, B. M., and Churchman, L. S. (2022). Transcription elongation is finely tuned by dozens of regulatory factors. *eLife*, **11**, e78944. Publisher : eLife Sciences Publications, Ltd.

Cuello, C., Baldy, A., Brunaud, V., Joets, J., Delannoy, E., Jacquemot, M.-P., Botran, L., Griveau, Y., Guichard, C., Soubigou-Taconnat, L., Martin-Magniette, M.-L., Leroy, P., Méchin, V., Reymond, M., and Coursol, S. (2019). A systems biology approach uncovers a gene co-expression network associated with cell wall degradability in maize. *PloS One*, **14**(12), e0227011.

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget : a plant small RNA target analysis server (2017 release). *Nucleic Acids Research*, **46**(W1), W49–W54.

Defrance, M. and van Helden, J. (2009). info-gibbs : a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics (Oxford, England)*, **25**(20), 2715–2722.

Del Prete, S., Molitor, A., Charif, D., Bessoltane, N., Soubigou-Taconnat, L., Guichard, C., Brunaud, V., Granier, F., Fransz, P., and Gaudin, V. (2019). Extensive nuclear reprogramming and endoreduplication in mature leaf during floral induction. *BMC Plant Biology*, **19**(1), 135.

Deng, W. and Roberts, S. G. (2005). A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes & Development*, **19**(20), 2418–2423.

DeRidder, B. P., Shybut, M. E., Dyle, M. C., Kremling, K. A. G., and Shapiro, M. B. (2012). Changes at the 3'-untranslated region stabilize Rubisco activase transcript levels during heat stress in Arabidopsis. *Planta*, **236**(2), 463–476.

Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., and Cooper, M. (2012). A Gene Regulatory Network Model for Floral Transition of the Shoot Apex in Maize and Its Dynamic Modeling. *PLOS ONE*, **7**(8), e43450. Publisher : Public Library of Science.

Erdmann, R. M. and Picard, C. L. (2020). RNA-directed DNA Methylation. *PLOS Genetics*, **16**(10), e1009034. Publisher : Public Library of Science.

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., and Mathelier, A. (2020). JASPAR 2020 : update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **48**(D1), D87–D92.

Frei dit Frey, N., Garcia, A. V., Bigeard, J., Zaag, R., Bueso, E., Garmier, M., Pateyron, S., de Tauzia-Moreau, M.-L., Brunaud, V., Balzergue, S., Colcombet, J., Aubourg, S., Martin-Magniette, M.-L., and Hirt, H. (2014). Functional analysis of Arabidopsis immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences. *Genome Biology*, **15**(6), R87.

Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., and Brunaud, V. (2008). CATdb : a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, **36**(Database issue), D986–990.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, **17**(6), 877–885. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Godoy Herz, M. A. and Kornblihtt, A. R. (2019). Alternative Splicing and Transcription Elongation in Plants. *Frontiers in Plant Science*, **10**.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome : a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**(Database issue), D1178–D1186.

Grace, M. L., Chandrasekharan, M. B., Hall, T. C., and Crowe, A. J. (2004). Sequence and spacing of TATA box elements are critical for accurate initiation from the beta-phaseolin promoter. *The Journal of Biological Chemistry*, **279**(9), 8102–8110.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO : scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, **27**(7), 1017–1018.

Haag, J. R. and Pikaard, C. S. (2011). Multisubunit RNA polymerases IV and V : purveyors of non-coding RNA for plant gene silencing. *Nature Reviews. Molecular Cell Biology*, **12**(8), 483–492.

Hammal, F., de Langen, P., Bergon, A., Lopez, F., and Ballester, B. (2022). ReMap 2022 : a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Research*, **50**(D1), D316–D325.

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., Toomajian, C., Roux, F., and Bergelson, J. (2011). Adaptation to Climate Across the *Arabidopsis thaliana* Genome. *Science*, **334**(6052), 83–86.

Helden, J. v., Olmo, M. d., and Pérez-Ortín, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Research*, **28**(4), 1000–1010.

Hernandez-Garcia, C. M. and Finer, J. J. (2014). Identification and validation of promoters and cis-acting regulatory elements. *Plant Science*, **217-218**, 109–119.

Herr, A. J., Jensen, M. B., Dalmay, T., and Baulcombe, D. C. (2005). RNA polymerase IV directs silencing of endogenous DNA. *Science (New York, N.Y.)*, **308**(5718), 118–120.

Hill, C. B. and Li, C. (2022). Genetic Improvement of Heat Stress Tolerance in Cereal Crops. *Agronomy*, **12**(5), 1205. Number : 5 Publisher : Multidisciplinary Digital Publishing Institute.

Hong, J. C. (2016). General Aspects of Plant Transcription Factor Families. pages 35–56.

Hunt, A. G., Xing, D., and Li, Q. Q. (2012). Plant polyadenylation factors : conservation and variety in the polyadenylation complex in plants. *BMC Genomics*, **13**(1), 641.

Jegga, A. G. and Aronow, B. J. (2013). Evolutionarily Conserved Noncoding DNA. In John Wiley & Sons, Ltd, editor, *eLS*. Wiley, 1 edition.

Ji, G., Chen, M., Ye, W., Zhu, S., Ye, C., Su, Y., Peng, H., and Wu, X. (2018). TSAPA : identification of tissue-specific alternative polyadenylation sites in plants. *Bioinformatics (Oxford, England)*, **34**(12), 2123–2125.

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 Research*, **45**(D1), D1040–D1045.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, **20**(6), 861–873.

Jores, T., Tonnies, J., Wrightsman, T., Buckler, E. S., Cuperus, J. T., Fields, S., and Queitsch, C. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nature Plants*, **7**(6), 842–855.

Kim, D. H., Sætrom, P., Snøve, O., and Rossi, J. J. (2008). MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proceedings of the National Academy of Sciences*, **105**(42), 16230–16235. Company : National Academy of Sciences Distributor : National Academy of Sciences Institution : National Academy of Sciences Label : National Academy of Sciences Publisher : Proceedings of the National Academy of Sciences.

Kiran, K., Ansari, S. A., Srivastava, R., Lodhi, N., Chaturvedi, C. P., Sawant, S. V., and Tuli, R. (2006). The TATA-Box Sequence in the Basal Promoter Contributes to Determining Light-Dependent Gene Expression in Plants. *Plant Physiology*, **142**(1), 364–376.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, **20**(4), 207–220. Number : 4 Publisher : Nature Publishing Group.

Ksouri, N., Castro-Mondragón, J. A., Montardit-Tarda, F., van Helden, J., Contreras-Moreira, B., and Gogorcena, Y. (2021). Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants : the peach example. *Plant Physiology*, **185**(3), 1242–1258.

Kwon, B., Fansler, M. M., Patel, N. D., Lee, J., Ma, W., and Mayr, C. (2022). Enhancers regulate 3' end processing activity to control expression of alternative 3'UTR isoforms. *Nature Communications*, **13**(1), 2709. Number : 1 Publisher : Nature Publishing Group.

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., and Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription : sequence-specific DNA binding by transcription factor IIB. *Genes & Development*, **12**(1), 34–44.

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. *Molecular Plant*, **12**(6), 743–763.

Landick, R. (2009). Functional divergence in the growing family of RNA polymerases. *Structure (London, England : 1993)*, **17**(3), 323–325.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals : a Gibbs sampling strategy for multiple alignment. *Science (New York, N.Y.)*, **262**(5131), 208–214.

Lay, F. D., Kelly, T. K., and Jones, P. A. (2018). Nucleosome Occupancy and Methylome Sequencing (NOMe-seq). *Methods in Molecular Biology (Clifton, N.J.)*, **1708**, 267–284.

Lee, K. and Kang, H. (2016). Emerging Roles of RNA-Binding Proteins in Plant Growth, Development, and Stress Responses. *Molecules and Cells*, **39**(3), 179–185.

Lifton, R. P., Goldberg, M. L., Karp, R. W., and Hogness, D. S. (1978). The organization of the histone genes in Drosophila melanogaster : functional and evolutionary implications. *Cold Spring Harbor Symposia on Quantitative Biology*, **42 Pt 2**, 1047–1051.

Lin, Z., Wu, W.-S., Liang, H., Woo, Y., and Li, W.-H. (2010). The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. *BMC Genomics*, **11**, 581.

Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS computational biology*, **9**(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 systems*, **3**(3), 278–286.e4.

Mayr, C. (2019). What Are 3' UTRs Doing ? *Cold Spring Harbor Perspectives in Biology*, **11**(10), a034728.

Minnoye, L., Marinov, G. K., Krausgruber, T., Pan, L., Marand, A. P., Secchia, S., Greenleaf, W. J., Furlong, E. E. M., Zhao, K., Schmitz, R. J., Bock, C., and Aerts, S. (2021). Chromatin accessibility profiling methods. *Nature Reviews Methods Primers*, **1**(1), 1–24. Number : 1 Publisher : Nature Publishing Group.

Mitra, A., Kataki, S., Singh, A. N., Gaur, A., Razafindrabe, B. H. N., Kumar, P., Chatterjee, S., and Gupta, D. K. (2021). Plant Stress, Acclimation, and Adaptation : A Review. In D. K. Gupta and J. M. Palma, editors, *Plant Growth and Stress Physiology*, Plant in Challenging Environments, pages 1–22. Springer International Publishing, Cham.

Mogno, I., Vallania, F., Mitra, R. D., and Cohen, B. A. (2010). TATA is a modular component of synthetic promoters. *Genome Research*, **20**(10), 1391–1397. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Moreau, M., Degrave, A., Vedel, R., Bitton, F., Patrit, O., Renou, J.-P., Barny, M.-A., and Fagard, M. (2012). EDS1 contributes to nonhost resistance of Arabidopsis thaliana against Erwinia amylovora. *Molecular plant-microbe interactions : MPMI*, **25**(3), 421–430.

Morneau, D. (2021). Pan-genomes : moving beyond the reference. *Nature Research*. Bandiera_abtest : a Cg_type : Milestones Publisher : Nature Publishing Group.

Oka, R., Zicola, J., Weber, B., Anderson, S. N., Hodgman, C., Gent, J. I., Wesselink, J.-J., Springer, N. M., Hoefsloot, H. C. J., Turck, F., and Stam, M. (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biology*, **18**(1), 137.

Oliver, K. R., McComb, J. A., and Greene, W. K. (2013). Transposable Elements : Powerful Contributors to Angiosperm Evolution and Diversity. *Genome Biology and Evolution*, **5**(10), 1886–1901.

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**(5), 1280–1292.

Piecyk, R. S., Schlegel, L., and Johannes, F. (2022). Predicting 3D chromatin interactions from DNA sequence using Deep Learning. *Computational and Structural Biotechnology Journal*, **20**, 3439–3448.

Place, R. F., Li, L.-C., Pookot, D., Noonan, E. J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences*, **105**(5), 1608–1613. Publisher : Proceedings of the National Academy of Sciences.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. (2006). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biology*, **7**(8), R78.

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., Braun, B. L., Dietze, M., Dunfee, B., Elsik, C. G., Manchanda, N., Coe, E., Sachs, M., Stinard, P., Tolbert, J., Zimmerman, S., and Andorf, C. M. (2019). MaizeGDB

2018 : the maize multi-genome genetics and genomics database. *Nucleic Acids Research*, **47**(D1), D1146–D1154.

Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A., and Mathelier, A. (2021). UniBind : maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics*, **22**(1), 482.

Quesneville, H. (2020). Twenty years of transposable element analysis in the Arabidopsis thaliana genome. *Mobile DNA*, **11**(1), 28.

Rensink, W. A., Lee, Y., Liu, J., Iobst, S., Ouyang, S., and Buell, C. R. (2005). Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genomics*, **6**(1), 124.

Rozière, J., Guichard, C., Brunaud, V., Martin, M.-L., and Coursol, S. (2022a). A comprehensive map of preferentially located motifs reveals distinct proximal cis-regulatory elements in plants. Pages : 2022.01.17.476590 Section : New Results.

Rozière, J., Guichard, C., Brunaud, V., Martin, M.-L., and Coursol, S. (2022b). A comprehensive map of preferentially located motifs reveals distinct proximal cis-regulatory sequences in plants. *Frontiers in Plant Science*, **13**.

Ruan, S., Swamidass, S. J., and Stormo, G. D. (2017). BEESEM : estimation of binding energy models using HT-SELEX data. *Bioinformatics*, **33**(15), 2288–2295.

Saad, C., Noé, L., Richard, H., Leclerc, J., Buisine, M.-P., Touzet, H., and Figeac, M. (2018). DiNAMO : highly sensitive DNA motif discovery in high-throughput sequencing data. *BMC Bioinformatics*, **19**, 223.

Savadel, S. D., Hartwig, T., Turpin, Z. M., Vera, D. L., Lung, P.-Y., Sui, X., Blank, M., Frommer, W. B., Dennis, J. H., Zhang, J., and Bass, H. W. (2021). The native cistrome and sequence motif families of the maize ear. *PLoS Genetics*, **17**(8), e1009689.

Schbath, S. and Hoebeke, M. (2011). R'MES : A Tool to Find Motifs with a Significantly Unexpected Frequency in Biological Sequences. In *Advances in Genomic Sequence Analysis and Pattern Discovery*, volume Volume 7 of *Science, Engineering, and Biology Informatics*, pages 25–64. WORLD SCIENTIFIC.

Schmitz, R. J., Grotewold, E., and Stam, M. (2021). Cis-regulatory sequences in plants : Their importance, discovery, and future challenges. *The Plant Cell*, **34**(2), 718–741.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, **132**(5), 887–898. Publisher : Elsevier.

Shelake, R. M., Kadam, U. S., Kumar, R., Pramanik, D., Singh, A. K., and Kim, J.-Y. (2022). Engineering drought and salinity tolerance traits in crops through CRISPR-mediated genome editing : Targets, tools, challenges, and perspectives. *Plant Communications*, page 100417.

Shen, Y., Liu, Y., Liu, L., Liang, C., and Li, Q. Q. (2008). Unique Features of Nuclear mRNA Poly(A) Signals and Alternative Polyadenylation in Chlamydomonas reinhardtii. *Genetics*, **179**(1), 167–176.

Singh, K. (2002). Transcription factors in plant defense and stress responses. *Current Opinion in Plant Biology*, **5**(5), 430–436.

Singh, R., Ming, R., and Yu, Q. (2016). Comparative Analysis of GC Content Variations in Plant Genomes. *Tropical Plant Biology*, **9**(3), 136–149.

Smale, S. T. and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, **72**, 449–479.

Stormo, G. D. (2000). DNA binding sites : representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Stringham, J. L., Brown, A. S., Drewell, R. A., and Dresch, J. M. (2013). Flanking sequence context-dependent transcription factor binding in early Drosophila development. *BMC Bioinformatics*, **14**, 298.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., and Moreau, Y. (2002). A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, **9**(2), 447–464.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012). RSAT peak-motifs : motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, **40**(4), e31.

Tian, B. and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nature Reviews. Molecular Cell Biology*, **18**(1), 18–30.

Tu, X., Mejía-Guerra, M. K., Valdes Franco, J. A., Tzeng, D., Chu, P.-Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E. S., and Zhong, S. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications*, **11**(1), 5089. Number : 1 Publisher : Nature Publishing Group.

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, **3**(10), 1578–1588.

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., and Vandepoele, K. (2022). PLAZA 5.0 : extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research*, **50**(D1), D1468–D1474.

Van de Velde, J., Van Bel, M., Vaneechoutte, D., and Vandepoele, K. (2016). A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. *Plant Physiology*, **171**(4), 2586–2598.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, **281**(5), 827–842.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., and Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, **13**(6), 508–514.

Wade, A. R., Duruflé, H., Sanchez, L., and Segura, V. (2022). eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction. *BMC Genomics*, **23**(1), 476.

Wang, Y., Fairley, J. A., and Roberts, S. G. E. (2010). Phosphorylation of TFIIB Links Transcription Initiation and Termination. *Current Biology*, **20**(6), 548–553.

Waters, A. J., Makarevitch, I., Noshay, J., Burghardt, L. T., Hirsch, C. N., Hirsch, C. D., and Springer, N. M. (2017). Natural variation for gene expression responses to abiotic stress in maize. *The Plant Journal*, **89**(4), 706–717. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13414.

Weirauch, M. T., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, **158**(6), 1431–1443.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA Methylation Mediated by a MicroRNA Pathway. *Molecular Cell*, **38**(3), 465–475.

Xing, D. and Li, Q. Q. (2011). Alternative polyadenylation and gene expression regulation in plants. *Wiley interdisciplinary reviews. RNA*, **2**(3), 445–458.

Xun, Y., Tang, Y., Hu, L., Xiao, H., Long, S., Gong, M., Wei, C., Wei, K., and Xiang, S. (2019). Purification and Identification of miRNA Target Sites in Genome Using DNA Affinity Precipitation. *Frontiers in Genetics*, **10**.

Yamamoto, Y. Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., and Abe, T. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC genomics*, **8**, 67.

Yang, W., Ng, P., Zhao, M., Wong, T. K., Yiu, S.-M., and Lau, Y. (2008). Promoter-sharing by different genes in human genome – CPNE1 and RBM12 gene pair as an example. *BMC Genomics*, **9**(1), 456.

Yu, C.-P., Lin, J.-J., and Li, W.-H. (2016). Positional distribution of transcription factor binding sites in Arabidopsis thaliana. *Scientific Reports*, **6**(1), 25164. Number : 1 Publisher : Nature Publishing Group.

Zaag, R., Tamby, J. P., Guichard, C., Tariq, Z., Rigaill, G., Delannoy, E., Renou, J.-P., Balzergue, S., Mary-Huard, T., Aubourg, S., Martin-Magniette, M.-L., and Brunaud, V. (2015). GEM2Net : from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response. *Nucleic Acids Research*, **43**(Database issue), D1010–D1017.

Zdobnov, E. M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., and Kriventseva, E. V. (2021). OrthoDB in 2020 : evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, **49**(D1), D389–D393.

Zhou, P., Enders, T. A., Myers, Z. A., Magnusson, E., Crisp, P. A., Noshay, J. M., Gomez-Cano, F., Liang, Z., Grotewold, E., Greenham, K., and Springer, N. M. (2022). Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *The Plant Cell*, **34**(1), 514–534.

Annexes

A . Rapport de stage de L2 d'Anne Duveau - Extraction de séquences proximales aux gènes pour différents génomes de plantes : développement d'un pipeline pour la création des fichiers d'entrées de PLMdetect

RAPPORT DE STAGE

du 1er au 30 Juin 2021

EXTRACTION DE SÉQUENCES PROXIMALES AUX GÈNES POUR DIFFÉRENTS GÉNOMES DE PLANTES DÉVELOPPEMENT D'UN PIPELINE POUR LA CRÉATION DES FICHIERS D'ENTRÉE DE PLMDETECT

Anne DUVEAU
L2 - Double Licence Sciences de la Vie, Informatique

Enseignante référente : Hélène DEBAT
Encadrants de stage : Julien ROZIERE et Marie-Laure MARTIN-MAGNIETTE

Table des matières

Remerciements

Je voudrais remercier tout d'abord Marie-Laure Martin-Magniette et Julien Rozière d'avoir accepté de me prendre comme stagiaire et de m'avoir donné l'opportunité d'en connaître plus sur le métier de bioinformaticien, ainsi que sur la vie au sein d'un laboratoire. J'ai pu aussi réaliser le stage en présentiel malgré les conditions sanitaires.

Je remercie Julien Rozière, qui a été mon maître de stage, qui m'a encadrée et guidée durant ce stage malgré son emploi du temps chargé, et qui s'est montré disponible à chaque fois que j'avais besoin d'aide pour la réalisation du projet de stage ainsi que pour la rédaction du rapport.

Je remercie aussi Marie-Laure Martin-Magniette pour sa disponibilité, ses conseils et son accompagnement durant ce stage.

Je remercie également toute l'équipe Gnet : ceux que j'ai pu rencontrer durant ce mois de stage pour leur accueil et leur sympathie.

1 Introduction

1.1 Contexte biologique

Phénomène de la transcription : expression des gènes

La transcription est la première étape pour l'expression génique. Elle consiste en la synthèse d'une molécule d'ARN, aussi appelé transcrit, à partir du brin d'ADN matrice grâce à l'ARN-polymérase. La transcription s'effectue du site d'initiation de la transcription, aussi appelé TSS (Transcription Start Site) jusqu'au site de terminaison de la transcription, le TTS (Transcription Termination Site). Pour la grande majorité des gènes, leur expression n'est pas constante, elle varie en fonction de l'environnement de la plante et du stade de développement de la cellule. Ainsi, chez les plantes, des gènes peuvent être sur-exprimés ou réprimés en réponse aux différents stress liés à leur environnement.

La régulation de l'expression des gènes et facteurs de transcription

La transcription est régulée majoritairement par l'intervention de protéines appelées facteurs de transcription (FT) (Figure 1). Un FT est une protéine régulatrice produite par des gènes régulateurs et peut induire ou non la transcription du gène. Si le FT est un activateur, il va agir positivement et permettre l'initiation de la transcription en se fixant sur une séquence amplificatrice appelée "enhancer". Dans le cas d'un répresseur, celui-ci va au contraire avoir une fonction négative et va empêcher la transcription et donc inhiber l'expression du gène. Les FT interviennent en se liant à l'ADN sur des motifs spécifiques avec lesquels ils ont une forte affinité. On les appelle des motifs *cis*-régulateurs ou plus simplement motifs de régulation.

Sur la figure 1, on peut observer des TFBS (Transcriptor Factor Binding Site) ou CRM (Cis-Regulatory Module) proches ou distants correspondant à des motifs de régulation.

Définition d'un motif de régulation

Un motif de régulation est une séquence particulière d'ADN variant de 4 à 15 bases. Elle est reconnue de manière spécifique par certains FT. La spécificité de la reconnaissance des FT ne veut pour autant pas dire que les motifs de régulation sont stricts. Un FT peut reconnaître un motif avec une variabilité sur certaines bases. Les bases qui ne permettent pas de variabilité et qui sont nécessaires pour la bonne fixation du FT constitue ce que l'on appelle le *core motif*.

Ces motifs se situent majoritairement en amont des gènes, en aval des gènes, les introns et dans les parties transcrites non traduites (UTR). Ils peuvent également se situer à de longues distances génomiques par rapport aux gènes régulés, ils interviennent dans la régulation des gènes cibles grâce à des repliements chromatiniens.

L'identification des motifs de régulation

Plusieurs méthodes sont utilisées pour identifier ces motifs régulateurs. Il existe des méthodes expérimentales (*in vivo*, *in vitro*) et bio-informatique (*in silico*) qui s'avèrent être complémentaires. Les deux méthodes expérimentales les plus utilisées sont les suivantes.

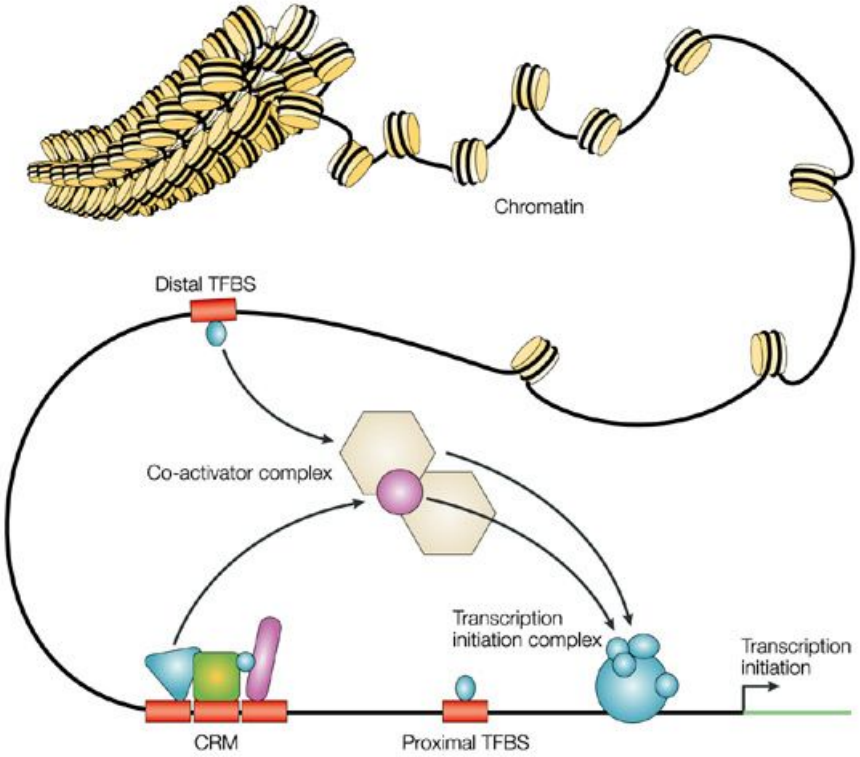


FIGURE 1 – Régulation de la transcription des gènes par les FT [13]

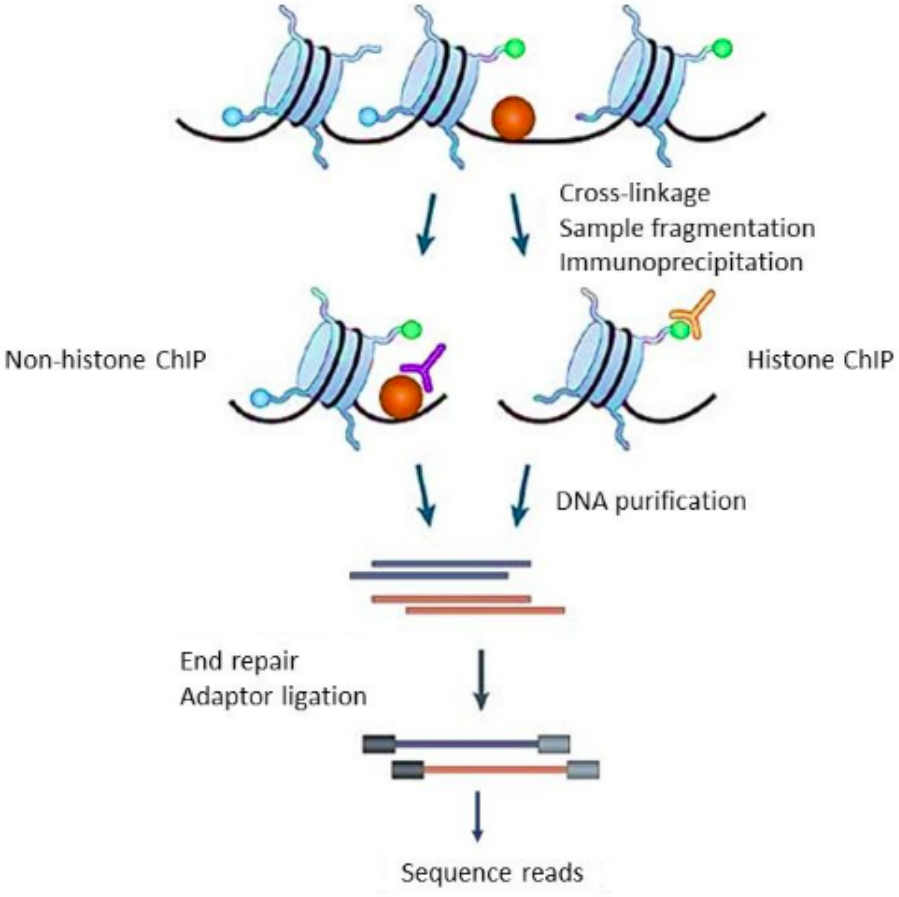


FIGURE 2 – Étapes de la technique de ChIP-seq [10]

ChIP-seq : *in vivo*

Le ChIP seq [10] (Chromatin ImmunoPrecipitation Sequencing) permet d'analyser les interactions entre les protéines et l'ADN *in vivo* : ici la protéine en question serait un facteur de transcription. Cette méthode (Figure 2) se base sur une immunoprécipitation de la chromatine (ChIP), en utilisant des anticorps spécifiques de la protéine d'intérêt et le séquençage haut débit. Elle permet de cartographier tous les sites de liaison sur l'ADN de la protéine d'intérêt, c'est-à-dire ici, tous les motifs de l'ADN reconnus par le facteur de transcription.

Les étapes de la technique sont les suivantes : il y a, dans un premier temps, une fixation des protéines sur l'ADN par un traitement chimique puis une fragmentation de la chromatine. Une immunoprécipitation sera ensuite effectuée pour isoler les complexes ADN-protéines en présence de l'anticorps d'intérêt. Enfin, après l'élimination des protéines et la purification des fragments d'ADN, les fragments d'ADN obtenus sont séquencés et alignés sur le génome de référence.

DAP-seq : *in vitro*

Cette autre méthode (Figure 3) appelée DAP-Seq [9] (DNA Affinity Purification Sequencing), contrairement au ChIP-Seq, ne nécessite pas d'anticorps spécifiques pour chaque facteur de transcription étudié. Le principe se fonde sur l'expression d'une protéine de fusion, ici le facteur de transcription avec un tag ou un GST (Glutathion -S Transférase) grâce à un ADN recombinant. Le GST sera lié aux billes de glutathion. La protéine de fusion va ensuite se lier aux motifs d'ADN qu'elle reconnaît. Après élution et lavage, on pourra ainsi isoler les séquences contenant les motifs d'intérêt spécifiques au facteur de transcription. Ces séquences seront ensuite séquencées et alignées sur le génome de référence de la même manière que le ChIP-seq.

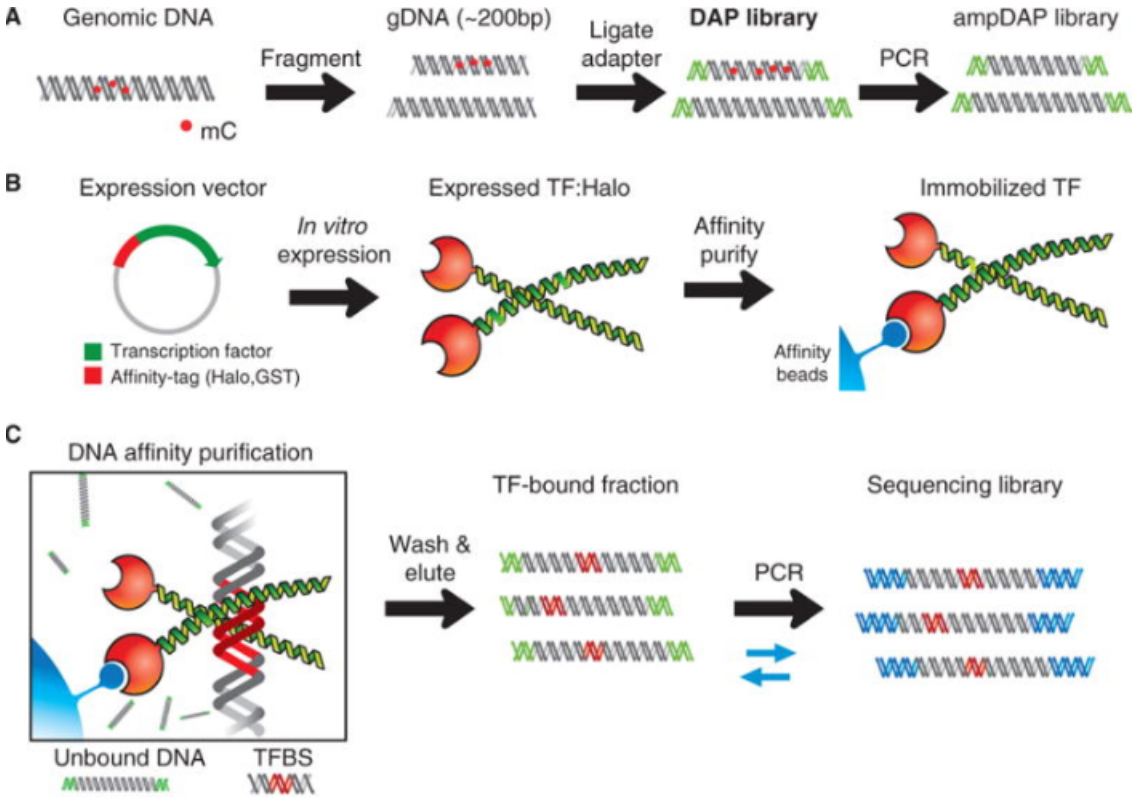


FIGURE 3 – Étapes de la technique du DAP-seq [9]

Les méthodes bio-informatiques : *in silico*

Enfin, il existe de nombreuses méthodes bio-informatiques permettant de détecter ces motifs. Elles peuvent être rangées en deux grandes catégories (Annexe I).

PWM et séquences consensus

Les méthodes peuvent être distinguées de cette manière : celles utilisant la représentation des motifs sous forme de séquences consensus (Figure 4) et celles représentant les motifs sous forme de matrices, majoritairement des *Position Weight Matrix* (PWM) (Figure 5).

Les séquences consensus sont une représentation des motifs de régulation considérant les nucléotides les plus abondants à chaque position du motif (Figure 4). Elles sont utilisées depuis de nombreuses années mais ne permettent que peu de variabilité à chaque position. C'est pour cela que le code IUPAC (Annexe II) a été créé afin de considérer des polymorphismes au sein des motifs.

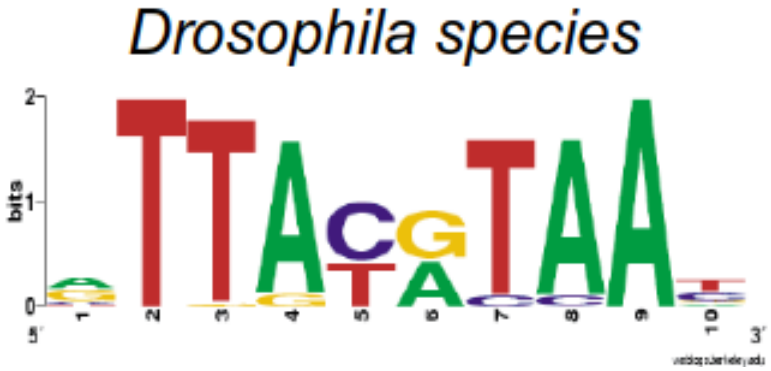


FIGURE 4 – Séquences consensus [11]

Les PWM, quant à elles, sont une représentation plus fine des motifs de régulation. Elle permet de présenter la proportion de chaque base nucléotidique à chaque position du motif. Cette représentation indique donc plus finement les polymorphismes au sein d'un motif que les séquences consensus.

Position weight matrix

A	421	0	0	873	0	466	0	885	954	91
C	106	0	0	0	547	0	76	66	0	311
G	357	0	32	81	0	477	0	1	0	105
T	70	954	922	0	407	11	878	2	0	447

FIGURE 5 – Exemple de PWM [11]

Parmi l'ensemble des méthodes proposées, RSAT (Regulatory Sequence Analysis Tools) et MEME (Multiple EM for Motif Elicitation) sont des références dans la recherche de motifs régulateurs.

RSAT (Regulatory Sequence Analysis Tools)

RSAT [8] était initialement un algorithme pour rechercher des motifs surreprésentés dans les séquences d'ADN, à l'origine sur le génome de S. cerevisiae (van Helden et al., 1998 [5]). Il permettait de mettre en évidence des motifs sur-représentés en comparant leur comptage dans des régions promotrices à celui dans le génome de l'organisme étudié. Il faisait appel à la représentation sous forme de séquences consensus des motifs.

Depuis RSAT a beaucoup évolué et n'est plus seulement un algorithme. C'est aujourd'hui une suite d'outils permettant l'extraction, la manipulation de séquences tout en proposant de la recherche de motifs *de novo* ou connus, sous forme de séquences consensus ou PWM, pour près de 10 000 organismes.

MEME (Multiple EM for Motif Elicitation)

Cette méthode [1] représente les motifs sous forme de PWM et permet d'identifier des motifs récurrents *de novo* dans un groupe de séquences. MEME utilise des techniques de modélisation statistique afin de déterminer les nucléotides appartenant aux motifs de régulation. Il peut également se servir d'un groupe de séquences contrôles afin de déterminer des motifs enrichis dans le groupe de séquences étudié.

1.2 Présentation de la méthode PLMdetect

D'autres méthodes se servent d'*a priori* biologiques afin de limiter le nombre de faux positifs identifiés lors de la détection. C'est le cas de la méthode sur laquelle j'ai travaillé durant ce stage : PLMdetect.

PLMdetect [2] (Preferentially Located Motif detection), est une méthode développée par mon équipe d'accueil Genomic Networks depuis 2010. Cette méthode détecte des motifs consensus préférentiellement localisés (PLM) dans des séquences proximales de gènes. Pour cela, elle déterminera si un motif est préférentiellement localisé par rapport à des références qui sont soient le TSS ou le TTS. Les motifs peuvent être identifiés *de novo* ou sur la base de motifs connus.

A l'origine cette méthode a été développée pour la recherche de motifs régulateurs dans les promoteurs d'*Arabidopsis thaliana*. Elle est depuis récemment étendue à d'autres espèces d'intérêt agronomiques, comme *Zea mays*, et à une autre région génomique à savoir les parties terminatrices des gènes.

Détail du fonctionnement de PLMdetect

La méthode prend en entrée un jeu de séquences d'intérêt et un ensemble de motifs à rechercher. Pour le jeu de séquences, cela consiste généralement en un jeu de séquences proximales de gènes différentiellement exprimés dans une même expérience. Cela permet d'identifier de potentiels éléments régulateurs communs expliquant la différence d'expression. Il peut également s'agir de l'ensemble des régions proximales du génome d'une espèce afin de faire une détection globale (aussi qualifiée de *genome-wide*). Pour les motifs, il peut s'agir de motifs de régulation déjà connus ou bien l'ensemble des motifs d'une taille donnée pour réaliser une recherche *de novo*. Dans son fonctionnement, la méthode récupère les séquences alignées sur une référence (TSS ou TTS) et réalise un comptage du nombre d'occurrences du motif considéré à chaque position de notre jeu de séquences. Elle obtient alors une distribution du motif. La méthode se sert d'une région d'apprentissage pour réaliser une régression linéaire qui est ensuite étendue à la région d'étude. Cela permet de détecter si un pic est observable dans la distribution du motif dans la région d'étude. Si c'est le cas, le motif est alors détecté comme un PLM.

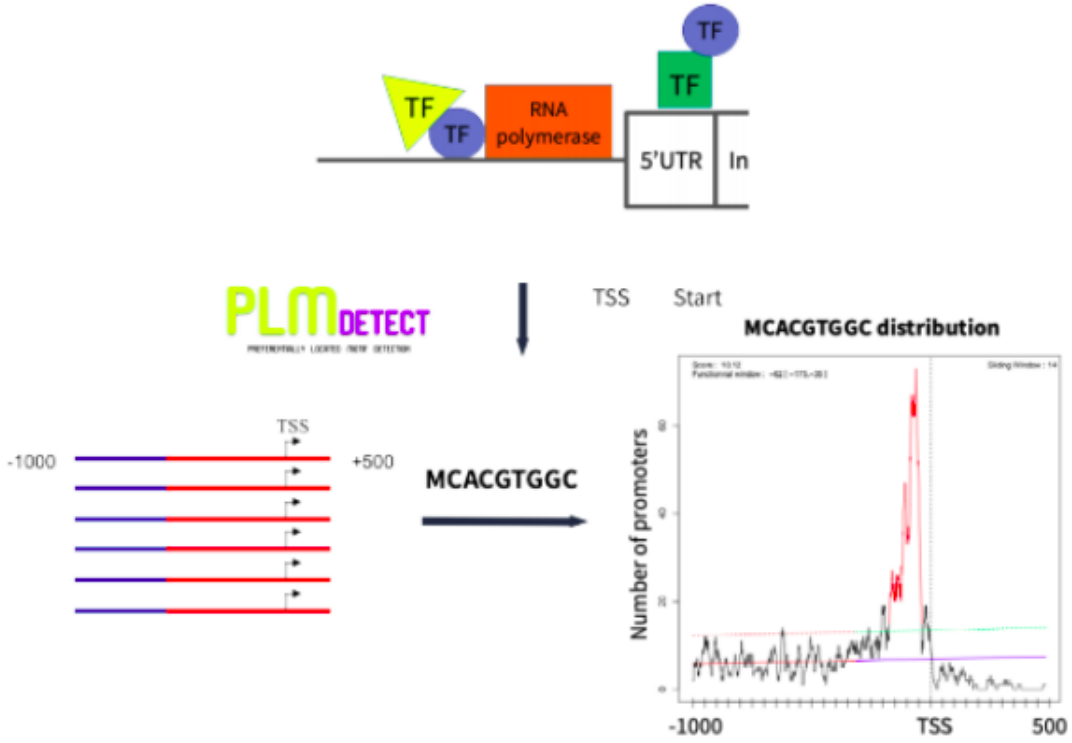


FIGURE 6 – Schéma explicatif du fonctionnement de PLMdetect

1.3 Présentation de l'Institut et de l'équipe

Le laboratoire dans lequel j'ai effectué mon stage est le laboratoire de l'Institut des Sciences des Plantes de Paris-Saclay (IPS2). L'institut regroupe 3 départements : le Département Génomique et Génétique du développement (DGG), le Département Physiologie et signalisation (DIPHY) et enfin le département Interactions Plantes Micro-organismes et réseaux (PMIN) dans lequel j'ai intégré l'équipe Genomic Networks.

L'équipe Genomic Network (GNet) est une équipe composée de bio-informaticiens et de biostatisticiens. L'équipe travaille sur le développement de modèles statistiques et d'approches bioinformatiques pour améliorer l'annotation fonctionnelle et relationnelle des gènes des plantes. Les thèmes principaux abordés au sein de l'équipe sont : les méthodes statistiques pour l'analyse de données omiques, l'intégration de données pour l'annotation fonctionnelle et l'étude des gènes impliqués dans la réponse aux stress.

Julien Rozière et Marie-Laure Martin-Magniette sont les personnes qui m'ont encadrés durant mon stage. Marie-Laure Martin-Magniette est directrice de recherche de l'INRAE et responsable de l'équipe GNet. Julien Rozière est doctorant en bio-informatique au sein de l'équipe GNet et également de Qualibiosec de l'Institut Jean-Pierre Bourgin (IJPB).

J'ai effectué mon stage une semaine en distanciel puis en présentiel au laboratoire. J'avais à disposition un ordinateur sur lequel je pouvais accéder à un des serveurs de développement de l'IPS2 et sur lequel je codais et réalisais les différentes extractions.

1.4 Objectifs du Stage & Contributions

Ce projet de stage s'inscrit dans le projet PLMviewer qui vise à étendre PLMdetect à de nombreuses espèces de plantes et à développer une interface web associée à la méthode nommée PLMview. Jusqu'à ce jour, 4 espèces ont pu être ajoutées sur l'interface web : *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Sorghum bicolor* et *Zea mays*.

Sujet du stage

Dans ce contexte, l'objectif de mon stage est de développer un pipeline informatique capable d'extraire les séquences proximales aux gènes pour un panel de 20 espèces sélectionnées pour être implémentées dans PLMview. Les séquences extraites constituent les fichiers d'entrée pour l'outil PLMdetect.

Ma contribution au projet : les tâches effectuées

Pendant ce mois de stage, j'ai réalisé différentes tâches afin de parvenir aux objectifs cités précédemment :

- Tout d'abord, il m'a fallu comprendre et étudier la structure des fichiers de séquences et d'annotation des différentes espèces végétales.
- Pour pouvoir faire tout cela, il a fallu d'abord récupérer les fichiers de séquences et d'annotation sur les bases de données dont les principales sont Phytozome 13 et Ensembl Plants 51.
- Ensuite, à partir d'un script d'extraction existant écrit en Perl et non général, c'est à dire non applicable à l'ensemble du panel, j'ai retranscrit le script en Python en le généralisant afin de traiter de manière automatique l'ensemble des espèces végétales étudiées.
- Après cela, j'ai pu commencer l'extraction des séquences promotrices avec mon code et ce d'abord sur les 4 espèces déjà extraites avant mon arrivée.
- Ensuite j'ai comparé les fichiers de sortie avec les fichiers de séquences existants afin de confirmer que mon code fonctionnait correctement. Pour les espèces dont les séquences ont été extraites pour la première fois (16 espèces), j'ai réalisé plusieurs contrôles afin de vérifier que les extractions s'étaient bien déroulées, dont la vérification des bornes d'extraction.
- J'ai aussi vérifié que les en-têtes pour les fichiers fasta en sortie étaient bien notées pour chaque espèce.
- Mon script étant conçu pour l'extraction de séquences de fichiers provenant de Phytozome, j'ai écrit un nouveau script permettant d'uniformiser les fichiers GFF qui n'ont pas exactement la même nomenclature.
- Après avoir uniformisé tous les fichiers d'annotation, l'extraction des séquences a pu être faite sur toutes les espèces du panel de plantes.

2 Matériels et méthodes

2.1 Génomes et annotations

Fichier de séquences du génome (fasta)

Pour chacune des 20 espèces du panel, j'ai téléchargé leur génome au format fasta. Le format fasta est caractérisé par un intitulé contenant le nom de la séquence suivi de cette même séquence nucléotidique.

Exemple :
"> nom du gène ou de l'espèce d'où provient la séquence et commentaires
AGGTCATACATAACTATTA......"

Fichier d'annotation du génome (gff3)

J'ai également téléchargé l'annotation structurale de chaque génome. Les fichiers d'annotations peuvent être sous plusieurs formats, les plus connus étant bed, gtf et gff3. Je dispose de fichiers gff3, GFF pour General Feature Format. Ce sont des fichiers de type tabulaire avec 9 champs dans chaque ligne qui permettent d'obtenir les coordonnées de l'ensemble des gènes (Annexe III). Ils me permettent d'obtenir les coordonnées de l'ensemble des gènes. Les fichiers de séquences et d'annotation proviennent des bases de données Phytozome 13 [4] et Ensembl [6].

2.2 Langage de programmation

Le script d'extraction de séquences proximales que j'ai réalisé a été codé en Python3. J'ai pu utiliser deux modules de Python :

- Le module argparse : pour la lecture d'options en ligne de commandes. C'est un parseur d'arguments, d'options et de sous-commandes. Il permet de prendre en compte les options données par l'utilisateur pour l'exécution du programme.

- J'ai aussi importé le module os.path afin de pouvoir utiliser la fonction path.exist(). Cette fonction permet de vérifier l'existence d'un fichier fourni en argument

2.3 Outil bio-informatique

Bedtools

Bedtools [12] regroupe un ensemble d'outils bio-informatiques permettant de réaliser de nombreuses analyses génomiques. Cela permet entre autres de faire l'extraction de séquences souhaitées d'un génome. Plusieurs options existent dans l'utilisation de Bedtools, dont "getfasta".

La fonction "getfasta" permet d'extraire des séquences d'un fichier fasta pour les intervalles donnés dans un fichier bed/gff/vcf (formats de fichiers d'annotation). Son utilisation se fait de la manière suivante :

```
$ bedtools getfasta [OPTIONS] -fi <input FASTA> -bed <BED/GFF/VCF>
```

Exemple de commande :

```
$ bedtools getfasta -fi fic_input.fa -bed tempo.gff3 -fo fic_output.fa -s -name
```

Les options que j'ai utilisées pour la fonction getfasta sont les suivantes (Tableau 1) :

Tableau 1 – Options de getfasta utilisées

-bed	Fichier d'annotation structurale contenant les bornes d'extraction
-name	Option pour mettre le nom du gène dans l'intitulé du fasta
-fi (file input)	Fichier fasta de référence (d'où les séquences seront extraites)
-fo (file output)	Nom du fichier de sortie contenant les séquences proximales des gènes extraites
-s	Oriente l'ensemble des séquences de 5' vers 3'

3 Résultats et discussions

Résultats

Dans le cadre de mon projet de stage, mon travail porte sur une partie du pipeline informatique qui permet l'extraction de séquences proximales puis l'analyse PLMdetect pour un panel de 20 espèces (Annexe IV). Les espèces notées "*" sont les espèces dont j'ai extrait les séquences. Ces espèces ont été choisies dans le but d'avoir un large échantillon avec des espèces de groupes phylogénétiques différents.

Uniformisation des fichiers d'annotation gff3 en fonction des bases de données

Contrairement au script d'extraction Perl mis en place avant mon arrivée, le script que j'ai développé en Python doit être général, c'est-à-dire capable de prendre en compte l'ensemble des annotations. Cependant, les fichiers gff3 ont de très nombreuses variations au sein de la neuvième colonne. Cette colonne m'est indispensable pour réaliser mon extraction correctement. J'ai ainsi fait le choix de ne travailler que sur une nomenclature particulière, celle des fichiers provenant de Phytozome 13. Il a donc été nécessaire d'uniformiser les fichiers d'annotation et de modifier les fichiers pour les espèces suivantes : *Cucumis melo*, *Zea mays* et *Arabidopsis thaliana*. Cela a été possible grâce à un script que j'ai nommé uniform_gff.py. C'est avec l'ensemble des fichiers uniformisés que j'ai pu procéder à l'extraction des séquences proximales. J'ai pour cela développé le script script_extraction.py (Annexe V).

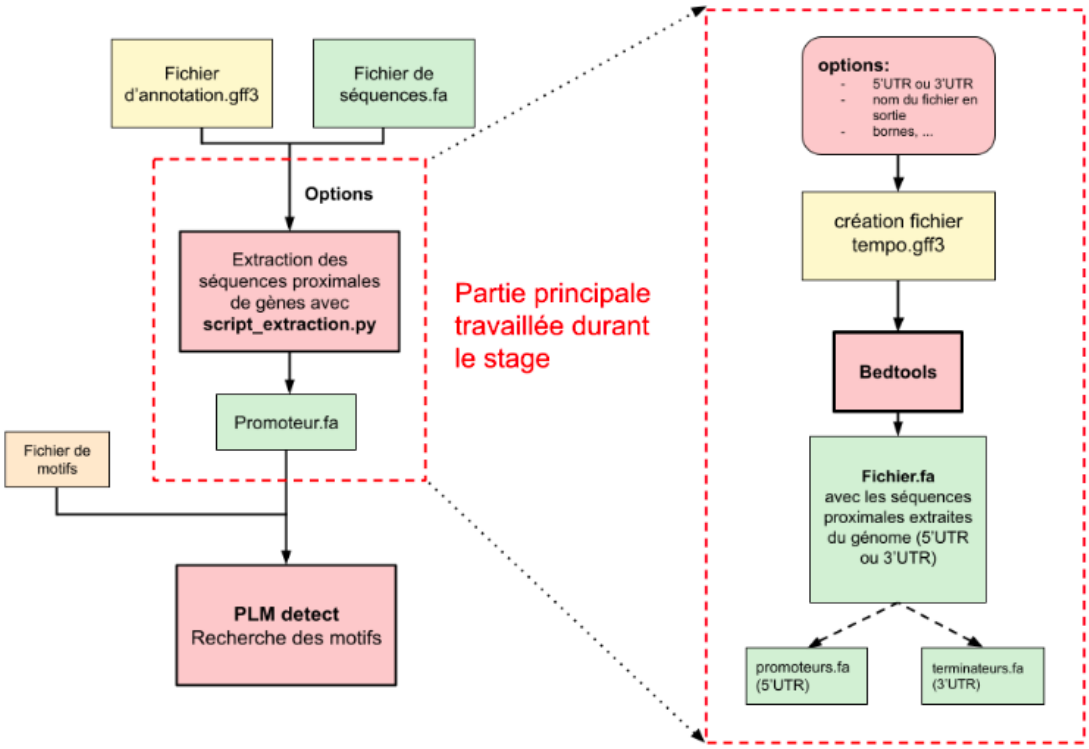


FIGURE 7 – Schéma fonctionnel du pipeline et du programme script_extraction.py

Étapes principales du programme d'extraction

Le programme d'extraction (Figure 7) se divise en 4 parties différentes

Etape 1 : cela concerne la prise en compte des fichiers d'entrée et des options d'extraction saisies dans la ligne de commande par l'utilisateur. Les différentes options sont les suivantes :

- **"–fichier_fasta" et "–gffFile" :** permettent de fournir les fichiers de séquences et d'annotation de l'espèce étudiée.
- **"–UTR" :** permet à l'utilisateur de définir la région proximale à extraire, 5' ou 3' (avec "5UTR" ou "3UTR"). Par défaut, on a $UTR = 5UTR$.
- **"–nom" :** permet de donner le nom du fichier des séquences extraites (par défaut, on a "promoteur").
- **"–borneinf" et "–bornesup" :** définissent les bornes d'extraction par rapport au TSS ou TTS (par défaut $borneinf$ = -1000 et $bornesup = 500$).
- **"–PLM" :** donne à l'utilisateur le choix de poursuivre l'analyse de ses séquences après l'extraction par la méthode PLMdetect (avec "yes" ou "no"). Par défaut, l'analyse ne se fait pas.
- **"–PLMrepository" :** permet de saisir l'adresse menant au répertoire où se trouve PLMdetect, nécessaire pour le lancement de l'option précédente.

usage : script_extraction.py [-h] [–*nom* NOM] [–*UTR* UTR] [–*fichier_fasta* FICHIER_FASTA] [–*gffFile* GFFFILE] [–*borneinf* BORNEINF] [–*bornesup* BORNESUP] [–*PLM* PLM] [–*PLMrepository* REPOSITORY]

Etape 2 : ici le script se focalise sur la gestion d'erreurs afin de détecter d'éventuels problèmes lors d'ouverture de fichiers et de vérifier si les paramètres saisis par l'utilisateur sont corrects.

Etape 3 : La méthode PLMdetect considérant comme *a priori* biologique la position du TSS ou TTS, il est primordial qu'ils soient correctement annotés. Cette étape, permet la lecture du fichier d'annotation et la création d'un fichier temporaire nommé tempo.gff3 dans lequel les bornes d'extraction sont modifiées afin de ne considérer que les régions à extraire pour chaque gène. Ceci dépend de la région proximale indiquée par l'utilisateur. Cependant, tous les gènes ne sont pas considérés dans les fichiers de sortie. En effet, le programme vérifie que les TSS ou TTS des gènes sont correctement annotés avant d'extraire les séquences. Un TSS ou un TTS bien annoté est issu d'un gène comportant ce qu'on appelle une rubrique (en anglais "feature") "five_prime_UTR" et ou un "three_prime_UTR" associée. Dans le cas de gènes codant pour des protéines, si l'UTR n'est pas annoté alors cela signifie que les positions des références que sont le TSS ou le TTS ne seront pas correctes. Le codon start et stop sont alors choisis comme extrémités des gènes.

Etape 4 : le programme utilise la fonction getfasta de Bedtools afin de générer le fichier de sortie au format fasta, contenant les séquences proximales des gènes extraites. En appliquant ce programme d'extraction à 20 génomes de plantes, on s'aperçoit que l'on garde au minimum environ 40% des gènes pour l'espèce de *Triticum aestivum* et *Medicago truncatula* et au maximum 90% des gènes environ chez par exemple *Arabidopsis lyrata*. Cela provient de l'application du filtre sur l'annotation des UTR qui amène PLMdetect à ne considérer qu'un sous-échantillon des gènes de chaque génome. Le tableau 2 donne la proportion de gènes considérés en fonction de l'espèce.

Tableau 2 – Résultats des extractions des séquences proximales pour chaque espèces du panel

Nom de l'espèce	Taille du génome	Nombre de Loci	5'UTR nbre gènes considérés	3'UTR nbre gènes considérés
Arabidopsis thaliana	135 Mb	27 416	19 736 (~70%)	19 573 (~70%)
Arabidopsis lyrata	207 Mb	31 073	29 792 (~95%)	29 726 (~95%)
Oryza sativa	372 Mb	42 189	22 122 (~50%)	23 182 (~55%)
Sorghum bicolor	732.2 Mb	34 129	26 545 (~80%)	28 073 (~80%)
Triticum aestivum	17 Gb	99 386	37 640 (~40%)	48 429 (~50%)
Zea mays	2.3 Gb	40 557	25 879 (~60%)	25 239 (~60%)
Helianthus annuus	3.6 Gb	52 243	32 581 (~60%)	35 930 (~70%)
Brachypodium distachyon	272 Mb	32 439	26 787 (~80%)	27 315 (~80%)
Hordeum vulgare	4,6 Gb	39 734	31 486 (~80%)	33 407 (~80%)
Malus domestica	688 Mb	45 116	28 834 (~60%)	30 583 (~70%)
Medicago truncatula	~360 Mb	50 894	21 022 (~40%)	21 849 (~40%)
Fragaria vesca	220 Mb	34 006	19 710 (~60%)	20 117 (~60%)
Phaseolus vulgaris	~600 Mb	27 433	22 367 (~80%)	22 281 (~81%)
Populus trichocarpa	392.2 Mb	34 699	30 661 (~90%)	30 846 (~90%)
Prunus persica	225.7 Mb	26 873	21 129 (~80%)	21 817 (~80%)
Solanum lycopersicum	~900 Mb	35 768	17 683 (~50%)	18 408 (~50%)
Vitis vinifera	487 Mb	31 845	14 412 (~45%)	16 155 (~50%)
Cucurbita maxima	279 Mb	32 076	14 780 (~50%)	16 054 (~50%)
Lupinus albus	451 Mb	38 258	32 122 (~85%)	31 914 (~80%)
Cucumis melo	375 Mb	27 427	18 886 (~70%)	19 731 (~70%)

Discussions

La généralisation de l'extraction des séquences à l'ensemble des espèces

L'une des difficultés majeures rencontrée lors de ce stage fût la généralisation de la méthode d'extraction. En effet, il m'a été nécessaire de comparer les fichiers d'annotation gff3 afin d'identifier quelles différences pouvaient être problématiques. La principale différence se situait au niveau de l'écriture de la dernière colonne, ce qui a pu poser problème lors de l'extraction de certains fichiers si on n'uniformisait pas les gff3 au préalable.

Exemple de différence dans la dernière colonne des gff

"gene	2	2541	.	+	.	ID=AL1G10010;Name=AL1G10010;ancestorIdentifier=918720.v1"
"gene	3631	5899	.	+	.	ID=AT1G01010;Note=protein_coding_gene;Name=AT1G01010"

On a ici, 2 lignes extraites chacune d'un fichier d'annotation d'une espèce différente (*Arabidopsis lyrata* et *Arabidopsis thaliana*) dont on a gardé à chaque fois les 7 dernières colonnes. Ce sont des lignes d'annotation d'un gène. On observe que leur dernière colonne n'a pas la même nomenclature. La partie donnant le nom du gène (en rouge) avec "Name=", par exemple, n'est pas placée dans le même ordre dans les 2 fichiers. Cela est arrivé pour l'ensemble des espèces non présentes dans Phytozome, où la nomenclature n'est pas stabilisée. Ainsi, en l'état actuel, il est nécessaire de formater les fichiers gff3 lorsqu'ils ne proviennent pas de Phytozome pour procéder à l'extraction. J'ai résolu ce problème par l'écriture du script uniform_gff.py, et j'ai donc pu procéder à l'extraction des régions proximales de toutes les espèces de manière automatique avec le script script_extraction.py. C'est un point qui montre l'importance de la nomenclature des fichiers en génomique.

L'extraction dépend de la qualité d'assemblage et d'annotation des génomes

Les résultats de l'extraction sont dépendants de la qualité d'assemblage et d'annotation des génomes. Certaines espèces comme *Zea mays* ou les deux espèces du genre Arabidopsis sont très bien annotées. Cela peut se justifier par le fait que ce sont des espèces particulièrement étudiées en biologie végétale. En effet, *Arabidopsis thaliana* est l'espèce modèle en biologie végétale et *Zea mays* l'une des espèces les plus étudiées chez les céréales. De l'autre côté, certaines espèces comme *Fragaria vesca* ont un génome avec une qualité d'assemblage moindre tout comme l'annotation des gènes. En effet, son fichier d'assemblage de génome est composé de ce que l'on appelle des contigs. Il s'agit de morceaux de séquence pouvant regrouper plusieurs gènes mais n'allant pas à l'échelle de chromosome. A cause de cela, les parties intergéniques que je souhaite extraire sont parfois manquantes. De plus, l'annotation structurale de certaines espèces n'étant pas de bonne qualité ou incomplète, les filtres appliqués sur l'annotation des 5'UTR et 3'UTR amène à ne considérer qu'un faible nombre de gènes par rapport au véritable nombre de gènes dans l'espèce. On omet également les gènes non codants pour une protéine avec ce filtre. Pour certaines espèces, plus de la moitié des gènes ne sont plus considérés à cause de ces deux facteurs (ex : *Triticum aestivum*).

4 Conclusion et perspectives

Au cours de mon stage, je suis parvenue à écrire un script d'extraction de séquences proximales en Python, l'exécuter pour extraire les séquences 5' et 3' proximales pour les 16 espèces non traitées du panel. Cela a nécessité d'uniformiser les fichiers d'annotation pour certaines d'entre elles grâce à un script que j'ai également écrit. Ce stage m'a permis de pouvoir approfondir mes connaissances dans le langage Python, et en particulier d'apprendre à coder un script capable de prendre des options en entrée via des instructions de l'utilisateur. Dans le cadre de ces développements, j'ai également appris à mettre en place de la gestion d'erreurs liées à des mauvaises entrées de la part des utilisateurs. J'ai aussi utilisé un nouvel outil bio-informatique, Bedtools et appris comment me connecter et travailler sur un serveur via des commandes ssh (cf Annexe III).

J'ai pu aussi comprendre la manière dont les génomes sont annotés par le biais des fichiers GFF3. J'ai également découvert le domaine de la recherche de motifs régulateurs et en particulier une méthode bio-informatique, PLMdetect.

Les applications à venir

Suite à mon travail, les espèces nouvellement extraites feront l'objet d'une analyse avec la détection *de novo* de PLMs pour chacune des régions (5'UTR et 3'UTR). J'ai d'ailleurs pu initier ce travail en une commande à la fin du script d'extraction permettant de lancer la méthode PLMdetect si demandé par l'utilisateur. De cette manière, j'ai pu ensuite lancer la méthode PLMdetect sur une espèce, *Medicago truncatula*, mais la durée de l'exécution de PLMdetect étant assez longue (quelques jours), je n'ai pas pu l'exécuter sur d'autres espèces, ni eu le temps d'étudier les résultats de l'analyse pour *Medicago truncatula*. Le travail d'extraction de séquences des espèces du panel que j'ai réalisé sera mis à disposition sur l'interface web PLMview utilisant la méthode PLMdetect.

Pour terminer, ce mois de stage m'a permis de découvrir la vie au sein d'un laboratoire et le travail d'un bio-informaticien. Et étant étudiante en double licence biologie-informatique, cela m'a permis de voir comment l'informatique peut être appliquée à la biologie dans le domaine végétal, particulièrement en génomique.

Liste des figures

Liste des tableaux

Bibliographie

[1] T. L. Bailey et C. Elkan. “Fitting a mixture model by expectation maximization to discover motifs in biopolymers”. eng. In : Proceedings. International Conference on Intelligent Systems for Molecular Biology 2 (1994), p. 28-36. issn : 1553-0833.

[2] Virginie Bernard, Alain Lecharny et Véronique Brunaud. “Improved detection of motifs with preferential location in promoters”. eng. In : Genome 53.9 (sept. 2010), p. 739-752. issn : 1480-3321. doi : 10.1139/g10-042.

[3] General feature format. fr. Page Version ID : 184299754. Juill. 2021. url : https://fr.wikipedia.org/w/index.php?title=General_feature_format&oldid=184299754 (visité le 21/11/2021).

[4] David M. Goodstein et al. “Phytozome : a comparative platform for green plant genomics”. eng. In : Nucleic Acids Research 40.Database issue (jan. 2012), p. D1178-1186. issn : 1362-4962. doi : 10.1093/nar/gkr944.

[5] J. van Helden, B. André et J. Collado-Vides. “Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies”. eng. In : Journal of Molecular Biology 281.5 (sept. 1998), p. 827-842. issn : 0022-2836. doi : 10.1006/jmbi.1998.1947.

[6] Kevin L. Howe et al. “Ensembl 2021”. eng. In : Nucleic Acids Research 49.D1 (jan. 2021), p. D884-D891. issn : 1362-4962. doi : 10.1093/nar/gkaa942.

[7] IUPAC Codes. url : https://www.bioinformatics.org/sms/iupac.html (visité le 21/11/2021).

[8] Alejandra Medina-Rivera et al. “RSAT 2015 : Regulatory Sequence Analysis Tools”. eng. In : Nucleic Acids Research 43.W1 (juill. 2015), W50-56. issn : 1362-4962. doi : 10.1093/nar/gkv362.

[9] Ronan C. O’Malley et al. “Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape”. eng. In : Cell 165.5 (mai 2016), p. 1280-1292. issn : 1097-4172. doi : 10.1016/j.cell.2016.04.038.

[10] Peter J. Park. “ChIP–seq : advantages and challenges of a maturing technology”. en. In : Nature Reviews Genetics 10.10 (oct. 2009), p. 669-680. issn : 1471-0056, 1471-0064. doi : 10.1038/nrg2641. url : http://www.nature.com/articles/nrg2641 (visité le 13/05/2022).

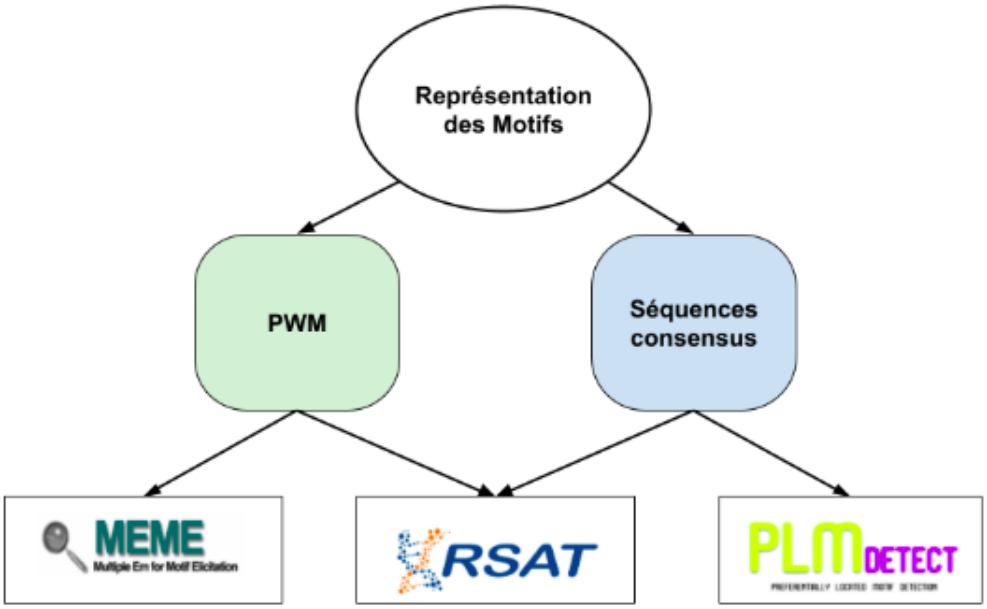
[11] Damiano Porcelli et al. “The nuclear OXPHOS genes in insecta : a common evolutionary origin, a common cis-regulatory motif, a common destiny for gene duplicates”. en. In : BMC Evolutionary Biology 7.1 (déc. 2007), p. 215. issn : 1471-2148. doi : 10.1186/1471-2148-7-215. url : https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-7-215 (visité le 13/05/2022).

[12] Aaron R. Quinlan et Ira M. Hall. “BEDTools : a flexible suite of utilities for comparing genomic features”. eng. In : Bioinformatics (Oxford, England) 26.6 (mars 2010), p. 841-842. issn : 1367-4811. doi : 10.1093/bioinformatics/btq033.

[13] Wyeth W. Wasserman et Albin Sandelin. “Applied bioinformatics for the identification of regulatory elements”. eng. In : Nature Reviews. Genetics 5.4 (avr. 2004), p. 276-287. issn : 1471-0056. doi : 10.1038/nrg1315.

ANNEXES

Annexe I – Schéma bilan des méthodes bio-informatiques

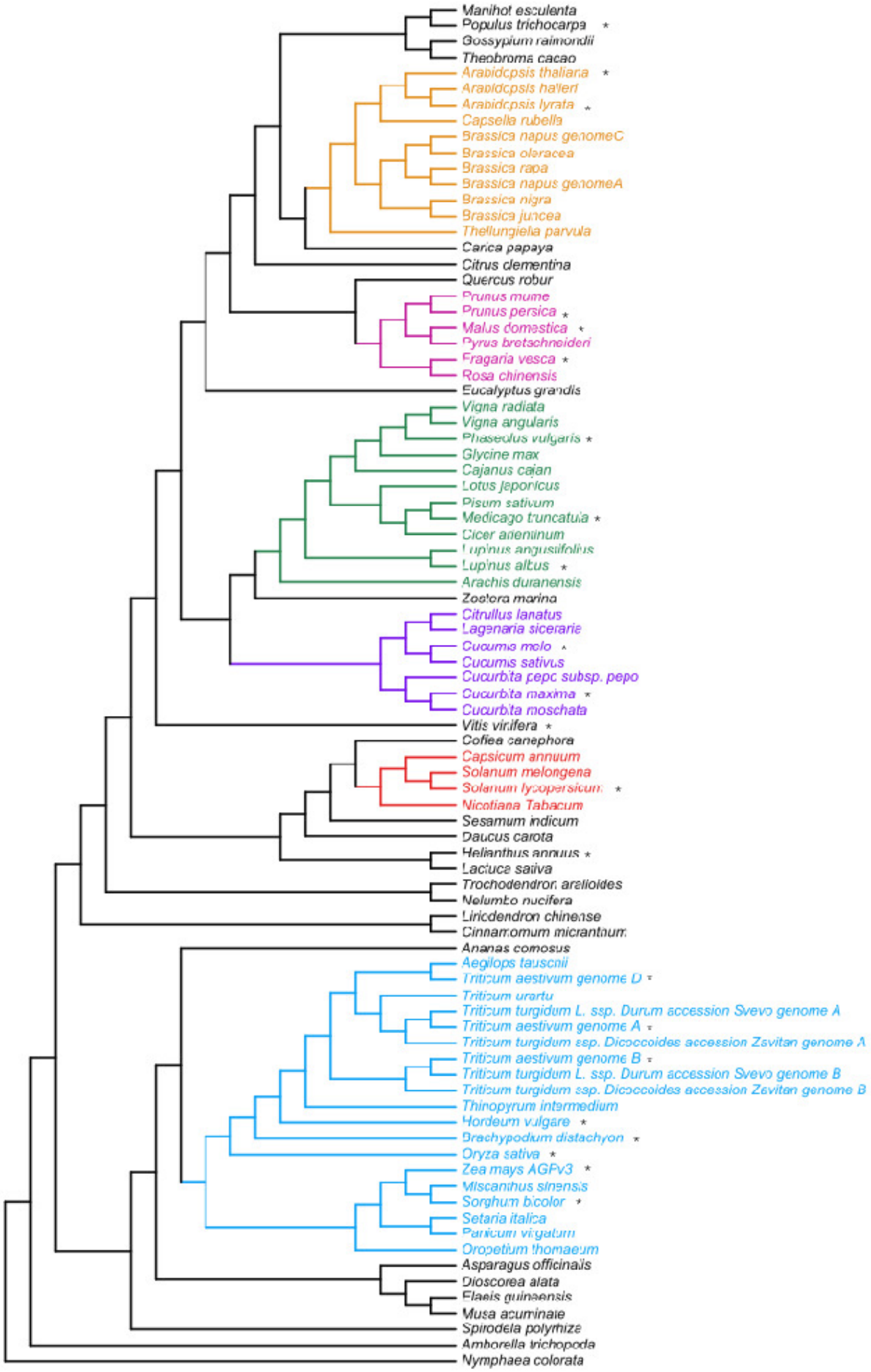


Annexe II – Code IUPAC pour les acides nucléiques [7]

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Annexe III – Tableau explicatif du format gff [3]

Indice de position	Nom de position	Description
1	séquence	Le nom de la séquence où se trouve l'élément.
2	source	Mot-clé identifiant la source de l'élément, comme un programme (par exemple Augustus ou RepeatMasker) ou une organisation (comme TAIR).
3	élément	Le nom du type d'élément, comme "gène" ou "exon". Dans un fichier GFF bien structuré, tous les éléments subordonnés suivent toujours leurs parents dans un seul bloc (ainsi, tous les exons d'un transcrit sont placés après la ligne de l'élément "transcrit" parent et avant toute autre ligne de transcrit). Dans le GFF3, tous les éléments et leurs relations doivent être compatibles avec les normes publiées par le projet Sequence Ontology [archive] .
4	début	Coordonnée génomique du début de l'élément, avec un **décalage de 1 base**. Ceci est en contraste avec d'autres formats de séquence à demi-ouverture basé sur 0, tels que les fichiers BED.
5	fin	Coordonnée génomique de Fin de l'élément, avec un **décalage de 1 base**. Il s'agit de la même coordonnée finale que dans les formats de séquence semi-ouverts à décalage 0, comme les fichiers BED. [réf. nécessaire]
6	score	Valeur numérique indiquant généralement la confiance de la source de l'élément annoté, ou son score. Une valeur de "." (un point) est utilisé pour définir une valeur nulle.
7	brin	Caractère unique qui indique le brin codant (biologie moléculaire) de l'élément; il peut prendre les valeurs de "+" (positif ou 5 '→ 3'), "-" (négatif ou 3 '→ 5'), ou "." (indéterminé).
8	phase	phase des éléments de séquence codante (CDS); il peut s'agir de 0, 1, 2 (pour les éléments CDS) ou "." (pour tout le reste). Voir la section ci-dessous pour une explication détaillée.
9	Les attributs.	Toutes les autres informations relatives à cet élément. Le format, la structure et le contenu de ce champ est celui qui varie le plus entre les trois formats de fichiers concurrents.


Manihot esculenta
Populus trichocarpa *
Gossypium raimondii
Theobroma cacao
Arabidopsis thaliana *
Arabidopsis halleri
Arabidopsis lyrata *
Capsella rubella
Brassica napus genomeC
Brassica oleracea
Brassica rapa
Brassica napus genomeA
Brassica nigra
Brassica juncea
Thellungiella parvula
Carica papaya
Citrus clementina
Quercus robur
Prunus mume
Prunus persica *
Malus domestica *
Pyrus bretschneideri
Fragaria vesca *
Rosa chinensis
Eucalyptus grandis
Vigna radiata
Vigna angularis
Phaseolus vulgaris *
Glycine max
Cajanus cajan
Lotus japonicus
Pisum sativum
Medicago truncatula *
Cicer arietinum
Lupinus angustifolius
Lupinus albus *
Arachis duranensis
Zostera marina
Citrullus lanatus
Lagenaria siceraria
Cucumis melo *
Cucumis sativus
Cucurbita pepo subsp. pepo
Cucurbita maxima *
Cucurbita moschata
Vitis vinifera *
Coffea canephora
Capsicum annuum
Solanum melongena
Solanum lycopersicum *
Nicotiana Tabacum
Sesamum indicum
Daucus carota
Helianthus annuus *
Lactuca sativa
Trochodendron aralioides
Nelumbo nucifera
Liriodendron chinense
Cinnamomum micranthum
Ananas comosus
Aegilops tauschii
Triticum aestivum genome D *
Triticum urartu
Triticum turgidum L. ssp. Durum accession Svevo genome A
Triticum aestivum genome A *
Triticum turgidum ssp. Dicoccoides accession Zavitan genome A
Triticum aestivum genome B *
Triticum turgidum L. ssp. Durum accession Svevo genome B
Triticum turgidum ssp. Dicoccoides accession Zavitan genome B
Thinopyrum intermedium
Hordeum vulgare *
Brachypodium distachyon *
Oryza sativa *
Zea mays AGPv3 *
Miscanthus sinensis
Sorghum bicolor *
Setaria italica
Panicum virgatum
Oropetium thomaeum
Asparagus officinalis
Dioscorea alata
Elaeis guineensis
Musa acuminata
Spirodela polyrhiza
Amborella trichopoda
Nymphaea colorata

```
import argparse
import os.path
from os import path


# affichage de l'usage du programme
print
↪  ("\n\n#############################################################################################################")
print ("usage : script_extraction.py [-h] [--nom NOM] [--UTR UTR] [--fichier_fasta FICHIER_FASTA] [--gffFile
↪  GFFFILE] [--borneinf BORNEINF] [--bornesup BORNESUP]")
print ("[--PLM PLM] [--PLMrepository repository]\n\n")
print ("nom           : entrez le nom de sortie du fichier de promoteurs extraits sans préciser '5UTR' ou '3UTR'
↪  dans le nom (par défaut : promoteur")
print ("UTR           : indiquez avec un 3UTR ou 5UTR la région que vous souhaitez analyser (par défaut : 5UTR)")
print ("fichier fasta : entrez le nom du fichier.fa contenant les séquences du génome de l'espèce")
print ("gffFile       : entrez le nom du fichier d'annotation.gff3")
print ("borneinf      : modifier la première coordonnée (par défaut : -1000)")
print ("bornesup      : entrez la seconde coordonnée (par défaut : 500)")
print ("PLM           : entrez si oui ou non vous souhaitez lancer PLMdetect sur vos séquences par 'yes' ou 'no'
↪  ")
print ("PLMrepository    : entrez le chemin de la méthode PLM detect ")
print
↪  ("\n#############################################################################################################\n\n")


# option en arguments à prendre en compte dans la ligne de commande
parser = argparse.ArgumentParser()

parser.add_argument("--nom", help = "entrez le nom de sortie du fichier de promoteurs extraits", default =
↪  "promoteur", dest = "nom")
parser.add_argument("--UTR", help = "indiquez avec un 3UTR ou 5UTR la région que vous souhaitez analyser",
↪  default = "5UTR", dest = "UTR")
parser.add_argument("--fichier_fasta", help = "entrez le nom du fichier.fa contenant les séquences du génome de
↪  l'espèce", required = True, dest = "fichier_fasta")
parser.add_argument("--gffFile", help = "entrez le fichier d'annotation.gff3 associé au fichier.fa", required =
↪  True, dest = "gffFile")
parser.add_argument("--borneinf", help = "modifier la borne inférieure d'extraction", default = -1000, type =
↪  int, dest = "borneinf")
parser.add_argument("--bornesup", help = "modifier la borne supérieure d'extraction", default = 500, type = int,
↪  dest = "bornesup")
parser.add_argument("--PLM", help = "entrez si oui ou non vous souhaitez lancer PLMdetect sur vos séquences par
↪  'yes' ou 'no'", default = "no", required = True, dest = "PLM")
parser.add_argument("--PLMrepository", help = "entrez le chemin de la méthode PLM detect", dest =
↪  "PLMrepository")

args = parser.parse_args()
# default => valeur donnée à la variable correspondante si l'argument est absent de la ligne de commande
# dest => la valeur retourné est mis dans args.nomdest
# required => définie si l'argument est optionnel ou non / si required = True, l'argument n'est pas facultatif, et
↪  doit être mis dans les options
# type => type de l'argument de la ligne de commande, par défaut c'est string

nom = args.nom
UTR = args.UTR
fasta = args.fichier_fasta
gffFile = args.gffFile
borneinf = args.borneinf
bornesup = args.bornesup
PLM = args.PLM
repository = args.PLMrepository

#~~~~TEST~~~~#
print("fSeq : " + nom)
print("5UTR ou 3UTR : " + UTR)
print("fichier_fasta : " + fasta)
print("gffFile : " + gffFile)
print("borneinf : " + str(borneinf))
print("bornesup : " + str(bornesup))
print("PLM : " + PLM)
print("PLMrepository : " + repository)


# vérification erreurs des paramètres

if(fasta == "" or path.exists(fasta) == False):
```

```
        print ("\nERREUR : Entrez le fichier de séquence .fasta \n\n")
        exit

if("fa" not in fasta):
        print("\nERREUR : Le type du fichier de séquence n'est pas un .fasta! \n\n")
        exit

if(gffFile == "" or path.exists(gffFile) == False):
        print("\nERREUR : Entrez le fichier d'annotations .gff3 \n\n")
        exit

if("gff" not in gffFile):
        print("\nERREUR : Le type du fichier d'annotations n'est pas un .gff! \n\n")
        exit

if(UTR != "5UTR" and UTR != "3UTR"):
        print("\nERREUR : Entrez une région d'étude valide ('3UTR' ou '5UTR') \n\n")
        exit

if(bornesup < borneinf):
        print("\nERREUR : vos bornes supérieure et inférieure entrées ne sont pas correctes \n\n")
        exit

if(PLM != "yes" and PLM != "no"):
        print("\nERREUR : l'argument PLM doit être 'yes' ou 'no' \n\n")
        exit

if(repository == "" or path.exists(repository) == False):
        print ("\nERREUR : la destination donné du Fichier PLMdetect n'est pas correcte \n\n")
        exit


f = open("tempo.gff3", "w") #ouverture en écriture d'un nouveau fichier "temporaire"

ligne = ""

g = open(gffFile, "r") #ouverture en lecture du fichier d'annotations

ligne = g.readline() #on lit une ligne


while (ligne != ""): #tant que l'on n'a pas finit de lire le fichier d'annotation

        if (ligne[0] == "#"): #si la ligne commence par un #
                f.write(ligne) #écrire la ligne dans fichier tempo.gff3
        else:
                ligne = ligne.split("\t")  #split par '\t'
                if(ligne[2] == "gene"): #si la 2e colonne contient "gene"

                        tab = {} #création d'un dictionnaire tab

                        if(UTR == "5UTR"): #si le choix de l'utilisateur est 5UTR

                                if(ligne[6] == "+"):
                                        ligne[4] = int(ligne[3]) + bornesup #redef des bornes
                                        ligne[3] = int(ligne[3]) + borneinf

                                elif(ligne[6] == "-"):
                                        ligne[3] = int(ligne[4]) - bornesup #redef des bornes
                                        ligne[4] = int(ligne[4]) - borneinf

                        if(UTR == "3UTR"): #si le choix de l'utilisateur est 3UTR
                                if(ligne[6] == "+"):
                                        ligne[3] = int(ligne[4]) - bornesup
                                        ligne[4] = int(ligne[4]) - borneinf

                                elif(ligne[6] == "-"):
                                        ligne[4] = int(ligne[3]) + bornesup
                                        ligne[3] = int(ligne[3]) + borneinf


                        name = ligne[8].rstrip() # fonction: ".rstrip() pour enlever le "\n"
                        name = name.split(";") #on split avec le ";"
                        name = name[1]
                        name = name.split("=") #split avec le "="
                        name = name[1] #on a le nom du gène!!

                        ligne[2] = name #on remplace le feature par le nom du gene
```

```
                        tab[name] = 0 #on donne la valeur de 0 à la clé du gène

                        #creation de la nouvelle ligne que le programme va devoir écrire dans tempo.gff3 contenant
                        ↪  les modifs
                        nouvelle_ligne = ligne

                        i = 0
                        l = ""
                        for column in nouvelle_ligne:
                                if(i < len(nouvelle_ligne)-1):
                                        l = l + str(column) + "\t"
                                        i += 1
                                else:
                                        l = l + str(column)
                        #on concatène les infos de la ligne dans l
                if (ligne[2] == "five_prime_UTR" and UTR == "5UTR"):
                        find = l.find(name) #on cherche name dans l
                        if(find != -1): #si il ne trouve pas name dans la ligne

                                #si le nom du gène correspondant au 5UTR existe déjà dans tab et si la valeur
                                ↪  associée est de 0 et si les coordonnées sont bien positives
                                if(name in tab and tab[name] == 0 and nouvelle_ligne[3] > 0 and nouvelle_ligne[4]
                                ↪  > 0):
                                        f.write(str(l)) #j'écris la ligne dans le fichier f (tempo.gff3)
                                        tab[name] = 1 # on change la valeur à 1

                if (ligne[2] == "three_prime_UTR" and UTR == "3UTR"):
                        find = l.find(name)
                        if(find != -1):
                                if(name in tab and tab[name] == 0 and nouvelle_ligne[3] > 0 and nouvelle_ligne[4]
                                ↪  > 0):

                                        if(nouvelle_ligne[6] == "+"):
                                                nouvelle_ligne[6] = "-"
                                        elif(nouvelle_ligne[6] == "-"):
                                                nouvelle_ligne[6] = "+"
                                        l_3UTR = ""
                                        i = 0
                                        for column in nouvelle_ligne:
                                                if(i < len(nouvelle_ligne)-1):
                                                        l_3UTR = l_3UTR + str(column) + "\t"
                                                        i += 1
                                                else:
                                                        l_3UTR = l_3UTR + str(column)
                                        f.write(str(l_3UTR))
                                        tab[name] = 1
        ligne = g.readline()

g.close()
f.close()

command = "bedtools getfasta -fi " + fasta + " -bed tempo.gff3 -fo " + nom + "_" + UTR + ".fa -s -name"
command_1 = "rm tempo.gff3" #efface le fichier temporaire

os.system(command)
os.system(command_1)


#lancer PLMdetect si demandé par l'utilisateur
if (PLM == "yes"):
        command_plm = "perl " + repository + "PLMdetect.pl --seqFile " + nom + ".fa --motifFile " + repository +
        ↪  "MotifDeNovo48.txt --UTR " + UTR + " --dir " + nom
        + "_DeNovo_" + UTR + " --thread 3"
        os.system(command_plm)
```

ANNEXE VI – **Commandes shell utilisées**

Commande	Signification
ssh -X aduveau@adresseserveur	permet la connexion à un serveur - "ssh" → Secure Shell : permet de se connecter à distance sur une machine de manière sécurisé - "-X" → interface graphique
scp (-r) Athaliana_genes.fa **aduveau@adresseserveur:/destination**	copie d'un répertoire (avec "-r", sinon copie d'un fichier) local vers le serveur
grep -c ">" Alyrata_5prime.fa	pour compter le nombre de ">" dans le fichier fasta Alyrata_5prime.fa
gunzip	pour zipper un fichier/dossier
gzip	pour dézipper un fichier/dossier

B . Rapport de stage de M1 de Camille Lemercier - Étude de la distribution de motifs *cis*-régulateurs par une analyse comparée d'une vingtaine de plantes

RAPPORT DE STAGE

Du 25 avril au 26 août
Institut des Sciences des Plantes – Paris Saclay
Bâtiment 630, rue Noetzlin, 91190 Gif-sur-Yvette

ETUDE DE LA DISTRIBUTION DE MOTIFS *CIS*-REGULATEURS PAR UNE ANALYSE COMPAREE D'UNE VINGTAINE DE PLANTES

Camille LEMERCIER

M1 GENIOMHE – Mention Bio-informatique
Université d'Evry Val d'Essonne
Année Universitaire 2021-2022

Encadrants de stage : Marie-Laure MARTIN et Julien ROZIERE
Rapporteur de stage : Cyril DALMASSO
Responsable du Master : Marie-Hélène MUCCHIELLI-GIORGI

Mots-clés :

Expression génique , Régulation transcriptionnelle, Eléments *cis*-régulateurs, Promoteur, Motifs préférentiellement localisés (PLM), Plantes

Lexique :

ADN : Acide désoxyribonucléique
ATAC-seq : Assay for Transposase-Accessibility Chromatin with highthroughput sequencing
ARN : Acide ribonucléique
ChIPseq : Chromatin ImmunoPrecipitation Sequencing
CRM : *Cis*-regulatory Module (Module *cis*-régulateur)
DAP-seq : DNA Affinity Purification Sequencing
FT : Facteur de Transcription
IUPAC : International Union of Pure and Applied Chemistry
MEME : Multiple Expectation maximizations for Motif Elicitation
pb : paires de base
PLM : Preferentially Located Motif (Motif préférentiellement localisé)
PP: Position préférentielle
PWM : Position Weight Matrix (Matrice poids-position)
TFBS : Transcription Factor Binding Site (Site de liaison de facteur de transcription)
TSS : Transcription Start Site (Site d'initiation de la transcription)
TTS : Transcription Termination Site (Site de terminaison de la transcription)

Table des matières:

1. Introduction

1.1. Contexte scientifique

1.1.1. Expression des gènes et transcription

La régulation de l'expression des gènes est indispensable pour que le métabolisme des cellules soit en adéquation avec l'environnement de la plante. La chaîne d'expression d'un gène commence par la transcription qui permet, par l'intermédiaire de l'ARN polymérase, la synthèse d'un brin d'ARN à partir de la séquence d'ADN. Elle s'effectue du site d'initiation de la transcription, appelé TSS (Transcription Start Site) au site de terminaison appelé TTS (Transcription Termination Site). L'expression des gènes peut être contrôlée à plusieurs niveaux et notamment au niveau de la transcription, c'est ce qui est appelé la régulation de la transcription [Figure 1] .

1.1.2. Régulation de la transcription

La régulation de la transcription peut s'effectuer aux étapes de démarrage, d'élongation ou de terminaison de la transcription. Elle est majoritairement régulée par l'intervention de facteurs de transcription (FT) , des protéines reconnaissant spécifiquement de courtes séquences d'ADN (entre 4 et 30 paires de bases), appelés motifs *cis*-régulateurs ou TFBS (Transcription Factor Binding Site). Quand ces fixations permettent une augmentation de la transcription, on les nomme enhancers et le FT est considéré comme un activateur. A l'inverse, quand la liaison du FT à ces séquences diminue le taux de transcription, elles sont appelées silencers et le FT est appelé répresseur [Figure 2]. On appelle module *cis*-régulateur (CRM) une région de plusieurs centaines de bases présentant une série de sites de liaison pour un ou plusieurs FT [Figure 3]. Les FT sont donc capables de réguler positivement ou négativement la transcription en se fixant sur de courtes séquences spécifiques.

Les séquences reconnues par les FT peuvent se trouver dans des régions proximales des gènes : de 0 à 50 nucléotides avant le TSS (dans le core promoteur aussi appelé promoteur central) ou à quelques centaines de bases en amont du core promoteur (promoteur proximal) ainsi que dans des régions distales en amont ou en aval, parfois à plusieurs milliers de paires de bases du promoteur proximal. Bien que les FT reconnaissent ces éléments de façon spécifique, cette reconnaissance peut se faire avec une variabilité sur certaines bases.
La spécificité des sites de liaison peut alors être représentée par la séquence consensus en utilisant le code IUPAC (International Union of Pure and Applied Chemistry), une nomenclature internationale des différentes bases nucléiques [Figure 4]. Cette séquence consensus permet ainsi d'avoir une idée des variants possibles du motif. Les motifs peuvent aussi être représentés par une matrice occurrence-position indiquant le nombre de fois où chaque nucléotide est représenté pour chaque position du TFBS. De cette matrice peut découler une matrice poids-position (Position Weight Matrix PWM) dans laquelle la fréquence de chaque nucléotide est calculée. On peut ensuite, grâce à des logiciels comme WebLogo, obtenir une représentation visuelle de ces PWM [Figure 5]. Les PWM peuvent ensuite servir à la détection de motifs *in silico*. Les bases qui n'autorisent pas de variabilité et qui sont donc nécessaires à la fixation du FT constituent le core motif [Figure 6] .
Parmi les éléments *cis*-régulateurs connus, on retrouve la boîte TATA, généralement localisée 30 bases en amont du TSS. Elle est spécifiquement reconnue par le facteur de transcription TFIID via l'une des sous unités qui le compose, la TATA-Binding Protein (TBP). Cette association forme le complexe de pré-initiation de la transcription et initie le recrutement d'autres facteurs de transcription requis pour le recrutement de l'ARN polymérase et l'initiation de la transcription.

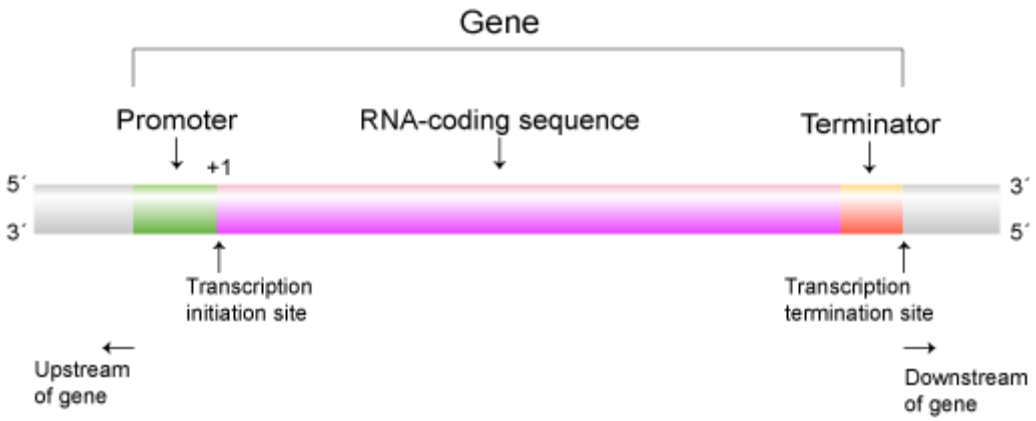


Figure 1: Représentation des sites d'initiation et de terminaison de la transcription d'un gène ainsi que de ses régions promotrices et terminatrices.

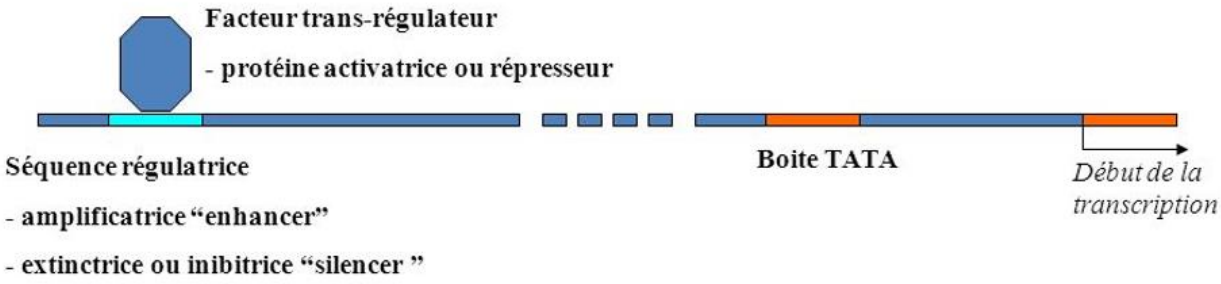


Figure 2: Fixation d'un facteur de transcription sur une séquence spécifique appelée élément cis*-régulateur localisée en amont du TSS influe, positivement ou négativement suivant la nature du FT, sur la transcription du gène sous contrôle du promoteur dans lequel l'élément* cis*-régulateur se trouve.*

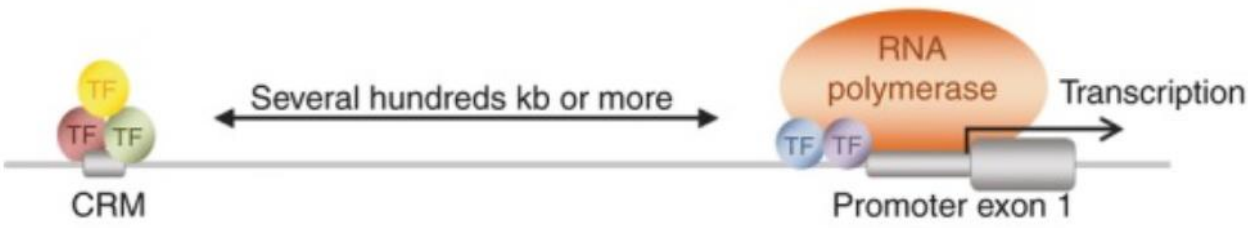


Figure 3: Fixation de plusieurs facteurs de transcription sur un module cis*-régulateur à plusieurs centaines de bases en amont du core promoteur influe sur la transcription du gène sous contrôle du dit promoteur.*

Chez les plantes, la séquence consensus des boîtes TATA serait TATAWA (où W correspond à A ou T selon le code IUPAC) et, chez l'organisme modèle *A.thaliana* l'hexamère TATAWA est conservé dans la région [-39pb ; -26pb] du TSS. [1]

Un autre motif, au sein du promoteur central et encore plus conservé que les boîtes TATA est l'élément initiateur. Il est reconnu par une protéine TBP-Associated Factor (TAF), qui s'associe à la TBP lors de l'initiation de la transcription. Cette protéine fait ainsi partie du complexe protéique du facteur d'initiation de la transcription TFIID. Il permettrait une meilleure fixation du complexe d'initiation de la transcription et donc une activité transcriptionnelle plus importante. De motif consensus YYANWYY chez les mammifères, il est localisé très proche du TSS (quelques bases en amont). Chez les plantes, c'est avant tout le dinucléotide initiateur YR ((T/C)(G/A)) qui semble conservé. Une séquence consensus plus longue du motif initiateur chez les plantes serait alors YTCAY. [2]

1.1.3. Méthodes de détection de motifs *cis*-régulateurs

Il existe trois types différents de méthodes pour identifier les motifs *cis*-régulateurs : les *in vivo*, *in vitro* et *in silico*.

Parmi les méthodes *in vivo*, on retrouve des approches expérimentales comme le ChIPseq (Chromatin ImmunoPrecipitation Sequencing), qui permet d'analyser les interactions entre les FT et l'ADN. Elle permet, par immunoprécipitation de la chromatine et séquençage à haut-débit, de cartographier tous les motifs reconnus par le FT [Figure 7]. Cependant la nécessité d'un anticorps spécifique pour chaque FT complique l'analyse de l'ensemble des FT [3].

D'autres techniques *in vitro* ne nécessitant pas d'anticorps spécifiques existent comme la méthode DAP-seq (DNA Affinity Purification Sequencing) dans laquelle une bibliothèque d'ADN est incubée avec un FT exprimé *in vitro* et marqué par une courte séquence d'acides aminés appelée étiquette d'affinité. Les complexes ADN-FT sont ensuite purifiés par séparation magnétique grâce à l'étiquette d'affinité et l'ADN lié est ensuite élué [Figure 8]. Cette méthode permet d'identifier des cibles de FT, néanmoins, elle ne permet pas de prendre en compte les cofacteurs potentiels ou le contexte génomique. D'autres stratégies reposant sur le profilage de la structure chromatinienne et permettant ainsi d'analyser l'occupation de FT existent. Parmi elles on retrouve l'ATAC-seq (Assay for Transposase-Accessibility Chromatin with highthroughput sequencing), qui permet de caractériser les régions accessibles de la chromatine grâce à la transposase [Figure 9], ou encore la recherche de sites hyper sensibles à la DNAse I [Figure 10]. Cependant ces approches dépendent des conditions expérimentales et présentent des limites de résolution.

Enfin les méthodes *in silico* peuvent se diviser en deux catégories : les méthodes énumératives ou heuristiques. Toutes deux visent à identifier de potentiels TFBS. Dans les méthodes énumératives, généralement, tous les motifs d'ADN de petite taille (<10 nucléotides) sont énumérés puis les motifs exacts surreprésentés sont recherchés. Les motifs vont ensuite être dégénérés, c'est-à-dire que position par position, chaque nucléotide du motif va être remplacé par une lettre de la nomenclature IUPAC. Si le motif est plus significatif avec la lettre dégénérée, il est conservé. Il existe toutefois d'autres méthodes énumératives, certaines utilisant des structures de données plus adaptées comme des automates ou des arbres de suffixes afin d'éviter l'énumération de tous les motifs. Néanmoins, avec ces méthodes énumératives, il n'est pas garanti de trouver un motif optimal. Ainsi, afin d'optimiser les motifs, des algorithmes probabilistes ont été développés. Les deux principaux sont l'algorithme MEME (Multiple Expectation Maximisation for Motif Elicitation) et l'algorithme de Gibbs sampling.

Code IUPAC	Bases
A	**A**denine
C	**C**ytosine
G	**G**uanine
T (ou U)	**T**hymine (ou **U**racil)
R	A ou G
Y	C ou T
S	G ou C
W	A ou T
K	G ou T
M	A ou C
B	C ou G ou T
D	A ou G ou T
H	A ou C ou T
V	A ou C ou G
N	A ou C ou G ou T

Figure 4 : Code IUPAC – Nomenclature de 15 lettres symbolisant les différentes combinaisons de bases nucléiques.

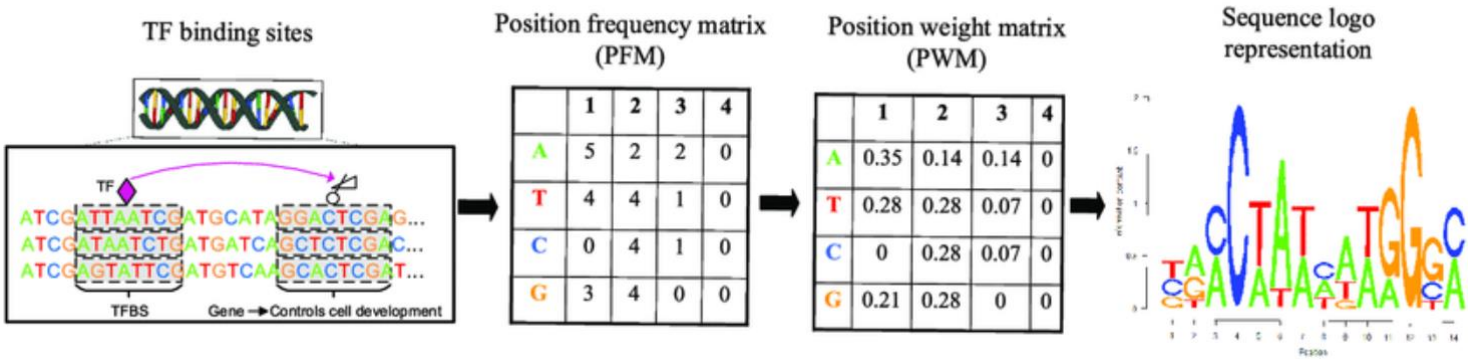


Figure 5: Calcul des différentes matrices à partir du motif lié par des FT. Matrice PFM représentant, pour chaque position du TFBS, le nombre d'occurrences d'un nucléotide. Matrice PWM représentant la fréquence des nucléotides pour chaque position et représentation graphique de la séquence du motif.

GGAC A GAT AAAC ACTT
GGAC GAC GAT AAGG ACTT
GGAC ACA GAT AAC ACTT
GGAC TCT GAT AACA TCTT
GGAC GTT GAT AAGT ACTT
GGAC GAA GAT AATC ACTT
GGAC CA GAT AAC ACTT
GGAC CGT GAT AAAC ACTT
GGAC ACC GAT AACA ACTT
GGAC GAT GAT AACA ACTT
GGAT AAA GAT AACA ACTT
GGAC GCA GAT AACA ACTT
GGAC GCA GAT ATCA ACTT

Figure 6: Ensemble de sites de fixations sur l'ADN reconnus par un même FT. Encadrée se trouve une sous séquence commune et conservée appelée core motif. Le core motif est entouré de séquences plus ou moins variables.

Le plus utilisé, l'algorithme MEME, est un algorithme EM (Expectation-Maximisation) initialisé avec une PWM aléatoire, c'est-à-dire que pour chaque position du motif, une lettre de l'alphabet est choisie aléatoirement. Chaque itération de l'algorithme comprend ensuite 2 étapes : l'étape d'expectation E qui estime la PWM à partir des nucléotides se trouvant aux positions sélectionnées à l'itération précédente en alignant les motifs et en calculant la fréquence de chaque nucléotide à chaque position. L'étape suivante est une étape de maximisation M dans laquelle chaque position de chaque motif est évaluée en calculant sa probabilité selon la PWM estimée. Les positions avec les probabilités les plus fortes sont alors retenues pour chaque séquence. Ces nouvelles positions sont alors utilisées dans l'étape E de l'itération suivante. L'algorithme continue ainsi jusqu'à converger vers un motif stable. L'algorithme Gibbs sampling est une variante de l'algorithme MEME. A chaque itération, une séquence aléatoire est choisie et l'étape d'espérance permettant d'estimer les paramètres de la PWM se fait sur toutes les séquences sauf celle choisie. A l'inverse, l'étape de maximisation se fait sur la séquence choisie uniquement et un poids est attribué à chaque position de cette séquence selon sa probabilité sous la PWM. Une étape d'échantillonnage est ensuite effectuée pour choisir une position selon les poids attribués, ainsi plus la probabilité d'une position est élevée, plus elle a de chance d'être choisie. Ce processus est ensuite répété jusqu'à convergence. [4,5,6]

1.1.4. PLMdetect et ses applications sur *Arabidopsis thaliana* et *Zea mays*

Il existe aussi des approches *ab initio* pour prédire les TFBS fondées sur le fait que ces motifs soient surreprésentés suivant leur environnement génomique. En effet ces éléments sont préférentiellement conservés durant l'évolution et retrouvés dans des promoteurs de gènes coexprimés ou impliqués dans les mêmes voies métaboliques. C'est sur ces critères que se fonde la méthode PLMdetect, permettant d'identifier les motifs surreprésentés selon leur localisation préférentielle par rapport au TSS ou au TTS des gènes. En effet, plus on se rapproche du TSS ou du TTS, et donc de la région où l'ADN est activement lié par des FT et plus le nombre d'occurrence d'un motif fixé augmente.
La méthode PLMdetect a ainsi été utilisée afin d'identifier à large échelle les motifs préférentiellement localisés (PLM) dans les régions proximales 5' et 3' de *Arabidopsis thaliana* et *Zea mays*. La distribution des PLM en fonction de leur score a alors révélé deux populations de PLM avec un score inférieur et supérieur à 2 pour chaque région proximale des 2 espèces. Plus le score d'un PLM est élevé, et plus ce PLM présente des contraintes topologiques fortes, de ce fait, seuls les PLM avec un score supérieur à 2 ont été gardés Cette étude a alors mené à l'identification de 6998 et 9768 PLM dans la région 5' de respectivement *A.thaliana* et *Z.mays* et de 7447 et 6639 PLM dans la région 3' de respectivement *A.thaliana* et *Z.mays*. Parmi ces PLM, 1063 sont retrouvés dans la région 5' proximale et 1677 dans la région 3' proximale des deux espèces , 98% de ces PLM communs étant localisés à 200pb du TSS ou du TTS. [Figure 11]
Ensuite, la distribution des positions préférentielles (PP) des PLM dans chaque région de chaque espèce a été analysée. Les quatre distributions obtenues, de profils similaires, ont révélé trois groupes pour chaque région : deux en amont du TSS (pour la région 5') ou du TTS (pour la région 3') et un groupe sur le TSS ou le TTS. Les PLM étaient toutefois plus dispersés chez le maïs *Z.mays* que chez *A.thaliana* [Figure 12]. Dans le groupe 2 de la région 5', comprenant les PLM localisés à [-50pb ; -10pb[du TSS, l'analyse en contenu nucléotidique a montré une prédominance des bases A et T, en accord avec de précédentes études indiquant la présence de la boîte TATA dans cette région. Les motifs ont ensuite été comparés aux TFBS connus et répertoriés dans la base de données JASPAR Plant 2020 ainsi qu'à 104 TFBS identifiés par ChIPseq chez le maïs.

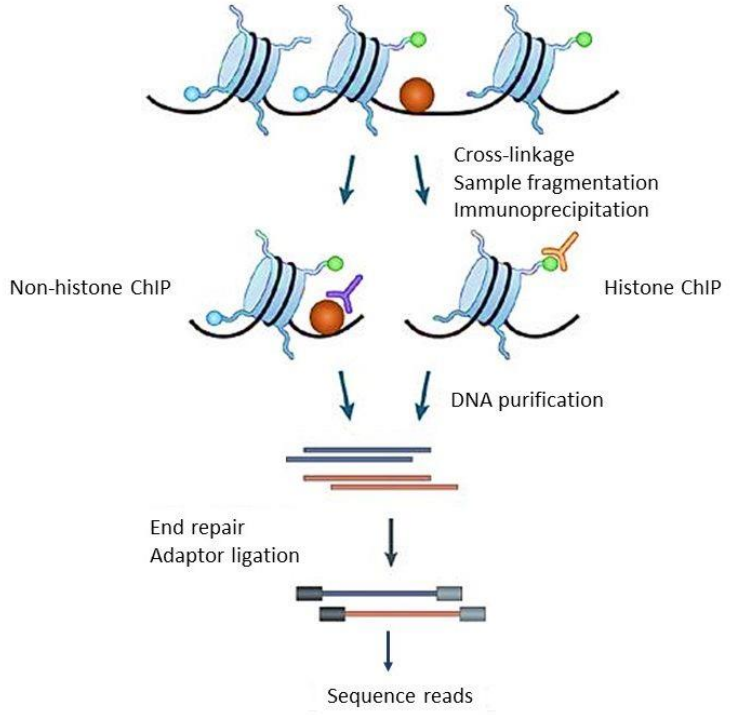


Figure 7:Méthode ChIPseq – Protéines liées à l'ADN sont fixées de façon covalente puis la chromatine est fragmentée et les complexes ADN-protéines sont isolés par immunoprécipitation en présence de l'anticorps d'intérêt. Après élimination des protéines et purification, les fragments d'ADN obtenus sont séquencés

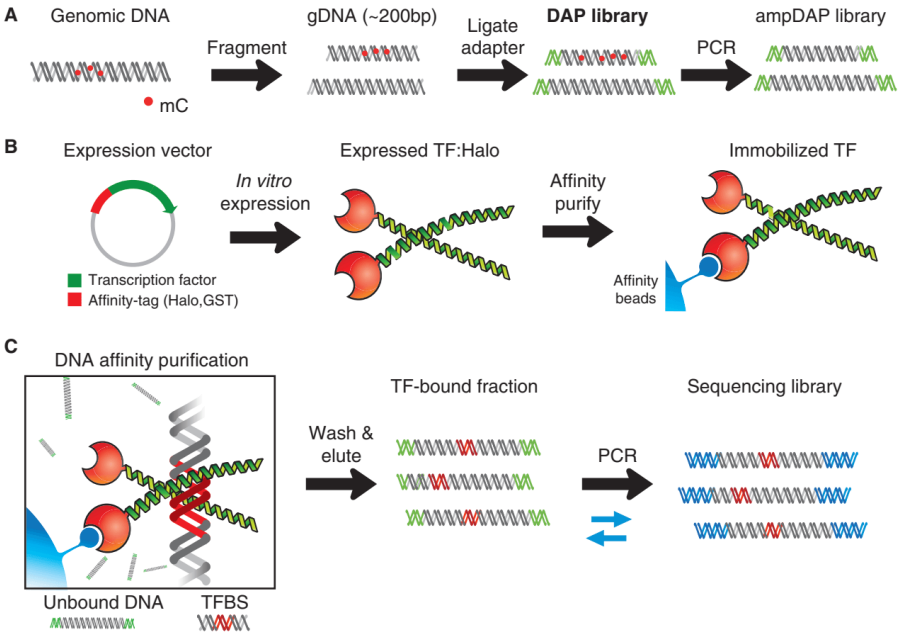


Figure 8: Méthode DAP-seq – Préparation d'une bibliothèque en fragmentant l'ADN génomique et en liant des adaptateurs aux extrémités de ces fragments. Les clones de FT, fusionnés avec un marqueur d'affinité comme HaloTag, sont exprimés in vitro et liés à des billes magnétiques. Les protéines de fusion (clones FT-HaloTag) sont ensuite incubées avec la bibliothèque précédemment préparée et les fragments d'ADN non liés par les FT sont éliminés. Les fragments d'ADN restant, donc les TFBS, vont ensuite pouvoir être séquencés.

Si les similarités étaient significatives, ces derniers étaient catégorisés tPLM. Les tPLM représentaient alors moins de 10% des PLM identifiés dans chaque région et étaient principalement localisés à [-200pb ; -50pb] en amont du TSS pour la région 5' et au niveau du TTS pour la région 3'. Les PLM identifiés ont aussi été comparés aux sites de fixation de miRNA, des petits ARN non codants qui, en se liant par complémentarité de séquence, inhibent l'expression de leurs gènes cibles. En utilisant psRNATarget, environ 8% chez l'arabette et 3% chez le maïs des PLM ont été trouvés comme associés à des sites de liaison pour miRNA.
Après ces comparaisons, 80% des PLM identifiés restaient non assignés. Afin de mieux les caractériser, une analyse d'enrichissement en utilisant les termes de Gene Ontology et les catégories fonctionnelles de MapMan a été réalisée [7].
Fondamentalement, la conservation de PLM entre ces deux espèces, de génomes lointainement apparentés, suggère que plus on se rapproche des gènes, plus le contexte génomique (dont les éléments *cis*-régulateurs) est conservé.

1.2. Objectifs du stage

Les plantes doivent s'adapter aux contraintes et aux changements de leur environnement. L'un des moteurs fondamentaux de cette adaptation est l'activation ou la répression de la transcription génique. Comme expliqué précédemment, cette régulation est contrôlée par la liaison de FT sur des motifs spécifiques localisés dans les régions proximales des gènes.
Grâce à la méthode PLMdetect, de nombreux motifs ont été identifiés comme préférentiellement localisés dans les régions proximales 5' et 3' d'*Arabidopsis thaliana* et *Zea mays* et les distributions des positions préférentielles des PLM pour chaque région proximale présentaient des similarités. En effet, chez ces deux espèces de plantes, trois groupes distincts avaient été obtenus dans chaque région.

L'objectif de mon stage est donc d'analyser la conservation des éléments *cis*-régulateurs du promoteur central au cours de l'évolution au travers des PLM identifiés chez une vingtaine d'espèces de plantes et appartenant à des familles botaniques différentes.

Dans un premier temps, j'ai étudié s'il y avait des motifs conservés chez toutes les espèces de plantes et si les trois groupes identifiés chez *Arabidopsis thaliana* et *Zea mays* étaient retrouvés chez toutes ou certaines espèces.
Je me suis ensuite intéressée à la distribution des positions préférentielles des PLM pour savoir s'il existait une homogénéité au sein d'une famille et s'il existait des motifs conservés spécifiquement par famille.
Enfin, j'ai analysé plus particulièrement les PLM localisés sur le promoteur central (correspondant à la région [-50pb ; 0pb]) pour étudier les séquences des boîtes TATA ainsi que les séquences de motifs initiateurs et identifier d'autres motifs au sein du promoteur central, n'appartenant ni au groupe des boîtes TATA ni aux motifs initiateurs.

2. Matériel et Méthodes

2.1. PLMdetect

La méthode PLMdetect permet d'identifier les motifs préférentiellement localisés par rapport au TSS ou au TTS des régions proximales. L'identification se fait en réalisant une régression linéaire sur la distribution d'apparition d'un motif dans la région neutre, déterminée comme les 500 bases précédant la région d'étude. Cette régression est ensuite étendue à la région d'étude correspondant à la région proximale du TSS ou du TTS. Si l'occurrence d'apparition du motif n'est pas incluse dans un intervalle de confiance à 99%, alors on observe la présence d'un pic et le motif est dit préférentiellement localisé [Figure 13].

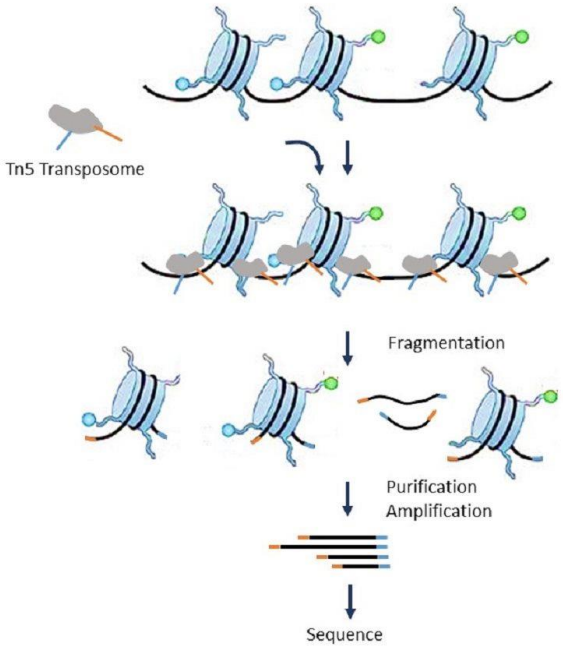


Figure 9: Méthode ATAC-seq – Incubation de la chromatine en présence de transposomes Tn5 pour fragmenter et indexer les fragments d'ADN exposés c'est-à-dire les régions avec chromatine ouverte, transcriptionnellement actives et accessibles.

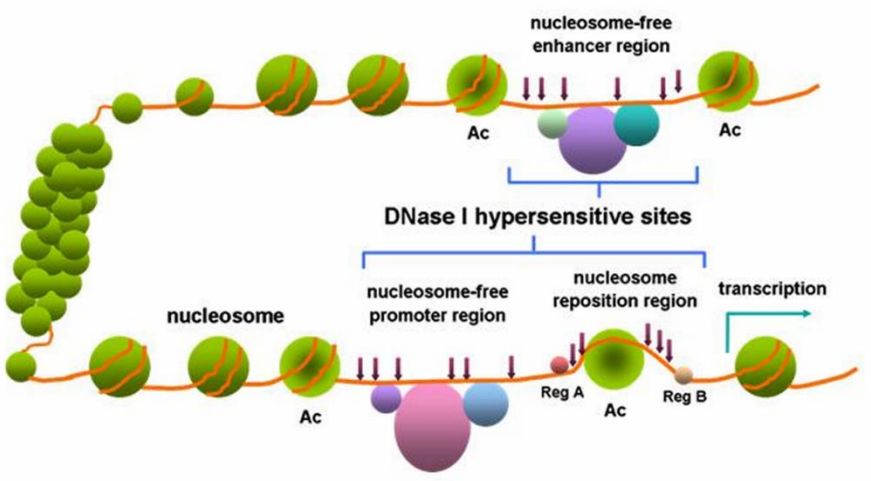


Figure 10: Méthode utilisant la DNAse I – L'enzyme va cliver tous les sites hypersensibles, représentés ici par des flèches, ces sites sont situés dans des régions de chromatine ouverte et donc accessibles pour les protéines venant se lier à l'ADN. Les sites hypersensibles à la DNAse I sont générés après liaison de facteurs de transcription, déplaçant les octamères d'histone.

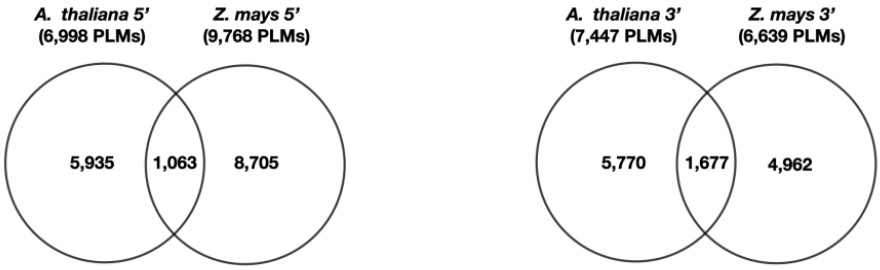


Figure 11:Diagramme de Venn représentant les PLM spécifiques et communs identifiés dans les régions proximales 5' (à gauche) et 3' (à droite) de Arabidopsis thaliana *et* Zea mays.

PLMdetect prend en entrée 2 fichiers, l'un contenant la liste de motifs à rechercher et l'autre les régions proximales dans lesquelles les motifs vont être recherchés. Pour le bon fonctionnement de la méthode, il est nécessaire que la position du TSS ou du TTS soit correctement renseignée. [8]
Cette méthode retourne une liste de motifs identifiés comme PLM avec leur score, reflétant la significativité de la prédiction ; leur position préférentielle par rapport au TSS ou au TTS ; la taille de leur fenêtre fonctionnelle ainsi que les positions de début et de fin de cette fenêtre fonctionnelle. C'est dans cette fenêtre que ces PLM sont significativement prédits comme préférentiellement localisés et pourraient être activement liés par un FT. Nous disposons également de la liste des gènes dans lesquels le PLM a été retrouvé.

2.2. Détection des PLM chez 20 espèces d'angiospermes

Avant mon arrivée au sein de l'IPS2, l'équipe a appliqué PLMdetect sur 20 espèces de plantes réparties dans 9 familles botaniques différentes :

- Les arabettes *Arabidopsis lyrata* (Arabette lyrée) et *Arabidopsis thaliana* (Arabette des dames) appartenant à la famille des Brassicaceae ;
- Les arbres fruitiers *Prunus persica* (Pêcher), *Malus domestica* (Pommier cultivé) et *Fragaria vesca* (Fraisier des bois) de la famille des Rosaceae ;
- *Phaseolus vulgaris* (Haricot commun), *Medicago truncatula* (Luzerne tronquée) et *Lupinus albus* (Lupin blanc) appartenant à la famille des légumineuses Fabaceae ;
- *Cucumis melo* (Melon) et *Cucurbita maxima* (Potiron) à la famille des cucurbitacées Cucurbitaceae ;
- Les céréales *Triticum aestivum* (Blé tendre), *Hordeum vulgare* (Orge commune), *Brachypodium distachyon* (Brachypode à 2 épis), *Oryza sativa* (Riz asiatique), *Zea mays* (Maïs) et *Sorghum bicolor* (Sorgo commun) appartenant à la famille des Poaceae ;
- *Populus trichocarpa* (Peuplier de l'Ouest) appartenant à la famille des Salicaceae ;
- *Vitis vinifera* (Vigne) de la famille des Vitaceae ;
- *Solanum lycopersicum* (Tomate) de la famille des Solanaceae ;
- *Helianthus annuus* (Tournesol) étant dans cette étude l'unique représentant de la famille des Asteraceae. [Annexe - Figure 1]

Les fichiers pris en entrée par PLMdetect sont un fichier de 87296 motifs, contenant l'ensemble des 4-mers à 8-mers possibles selon l'alphabet {A,T,G,C}et permettant la détection de PLM *de novo* ainsi qu'un fichier contenant, pour chaque plante, les régions proximales 5' et 3', définies par l'intervalle [-1000pb ; +500pb] du TSS ou du TTS.

Seules les régions proximales correctement annotées ont été gardées, elles ont été extraites :

- de Phytozome pour *A.lyrata*, *B.distachyon*, *F.vesca*, *H.annuus*, *H.vulgare*, *M.domestica*, *M.truncatula*, *O.sativa*, *P.vulgaris*, *P.trichocarpa*, *P.persica*, *S.lycopersicum*, *S.bicolor*, *T.aestivum* et *V.vinifera*;
- de TAIR10 pour *A.thaliana* ;
- de Ensembl pour *C.melo*;
- de Cucurbitgenomics pour *C.maxima* ;
- de Whitelupin pour *L.albus* ;
- de Maizegdb pour *Z.mays*.

Pour chaque plante analysée, je dispose de 2 fichiers contenant les motifs identifiés comme préférentiellement localisés sur l'intervalle [-500pb ;+500pb] des régions proximales 5' et 3'.

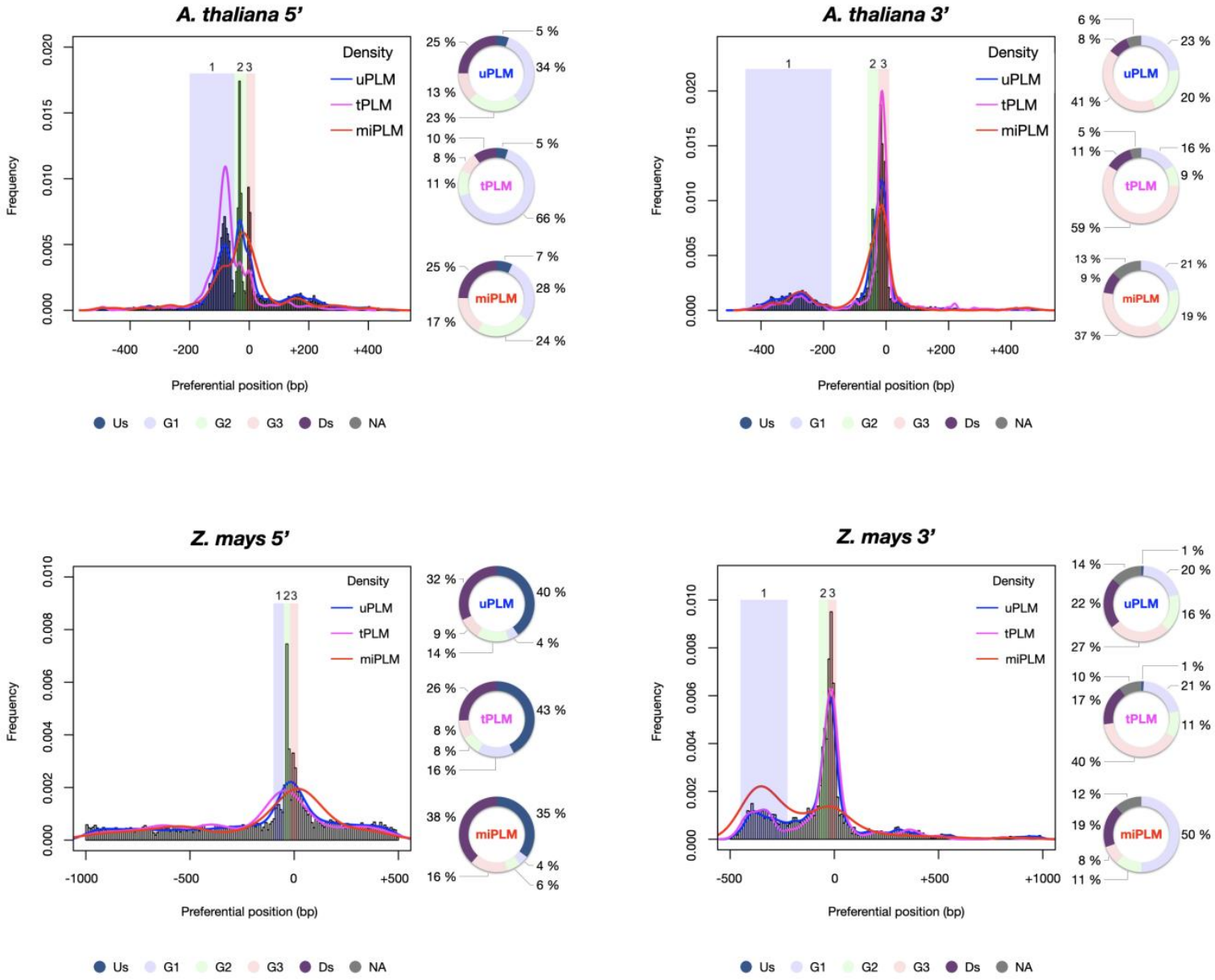


Figure 12: Histogrammes représentant la fréquence des PLM en fonction de leur position préférentielle par rapport au TSS pour les régions 5' proximales et par rapport au TTS pour les régions 3' proximale. Les groupes identifiés dans chaque région proximale de chaque espèce sont :
Pour les PLM en 5' chez A.thaliana (en haut à gauche) : Groupe 1: [-200;-50[; Groupe 2: [-50;-10[; Groupe 3 :[-10;+20[.
Pour les PLM en 3' chez A.thaliana *(en haut à droite): Groupe 1:]-450;-175]; Groupe 2:]-60;-25]; Groupe 3:]-25;+10].*
Pour les PLM en 5' chez Z. mays *(en bas à gauche): Groupe 1: [-100;-50[; Groupe 2: [-50;-20[; Groupe 3: [-20;+20[.*
Pour les PLM en 3' chez Z. mays *(en bas à droite): Groupe 1:]-450;-225]; Groupe 2:]-75;-30]; Groupe 3:]-30;+10 bp].*
Les PLM ont aussi été répartis en trois groupes : les tPLM qui sont des TFBS déjà connus, les miPLM qui sont vraisemblablement des sites de fixation pour miRNA et les uPLM qui sont les PLM n'appartenant pas aux deux groupes précédents.
Enfin, il est aussi indiqué si les PLM ainsi catégorisés sont retrouvés dans le groupe 1 (G1), dans le groupe 2 (G2), dans le groupe 3 (G3), dans une région en amont de ces 3 groupes (Us), dans une région en aval de ces 3 groupes (Ds) ou dans une région entre ces groupes (lorsque les groupes ne sont pas juxtaposés) (NA).

2.3. Comparaison des PLM avec des TFBS déterminés expérimentalement

L'algorithme de comparaison de motif TOMTOM a été utilisé pour comparer nos PLM avec une collection de TFBS déterminés expérimentalement (JASPAR PLANT 2022 [9]). Les paramètres suivants ont été choisis pour les comparaisons : distance de comparaison euclidienne , un seuil de q-value à 0.05 et l'option « complete scoring » désélectionnée (les PLM comparés étant de plus petite taille que les motifs contenus dans la base de données).

2.4. Langages de programmation

J'ai utilisé le langage de programmation R (4.0.3), adapté pour traiter rapidement un grand volume de données, réaliser des analyses statistiques et générer des graphiques afin de visualiser les données et les analyses effectuées.
Au sein de R, j'ai utilisé les librairies ggplot2 (3.3.5) pour générer les différentes distributions, dplyr (1.0.7) pour manipuler les données des différents tableaux et xlsx (0.6.5) pour importer et exporter les fichiers excel.

J'ai aussi utilisé du bash afin de modifier l'en tête des fichiers et remplacer la colonne « Functionnal Window » par les colonnes « fWSrt » et « fWStp » renseignant les positions de début et de fin de fenêtre fonctionnelle pour chaque PLM.
J'ai aussi utilisé bash afin de reprendre le document .csv contenant la liste des motifs à comparer et les séparer par un saut de ligne pour qu'ils puissent être pris en compte séparément par l'algorithme TOMTOM.

```
awk -F'"' '{print $4,"\n"}' ListeMotifPasInitTATA.csv > ListeMotif
```

3. Résultats obtenus

3.1. Contrôle qualité

J'ai d'abord commencé par filtrer les motifs identifiés comme PLM par la méthode mais n'ayant en réalité pas de localisation préférentielle par rapport au TSS ou TTS. Pour cela j'ai retiré les PLM dont la fenêtre fonctionnelle dépassait la région d'étude du motif, c'est-à-dire allant au-delà de +500pb du TSS ou du TTS. En effet ces motifs ne sont pas des PLM, ils sont simplement plus représentés dans la région d'étude que dans la région neutre [-1000pb ;-500pb]. Garder ces motifs mènerait à de la surinterprétation, c'est pourquoi ils doivent être filtrés. Comme dans l'étude genome-wide menée chez l'arabette des dames et le maïs, les PLM avec une fenêtre fonctionnelle supérieure à 150pb, ont été retirés de l'étude.

Une fois les différents filtres de contrôle qualité appliqués, la distribution des scores des PLM a été générée. Toutes les espèces, à l'exception de *S.bicolor* et *Z.mays*, présentaient dans les 2 régions proximales, 2 populations de PLM avec un score inférieur ou supérieur à 2 et seuls les PLM avec un score supérieur à 2 ont été gardés. [Figure 14]

La distribution des positions préférentielles des PLM pour chaque région de chaque espèce a ensuite été générée. Pour 18 des 20 espèces, toujours à l'exception du maïs et du sorgho, les pics des distributions sont observés autour du TSS ou du TTS. Pour les 2 autres espèces, les PLM sont localisés à +500pb du TSS et -500pb du TTS. Cela a permis de mettre en évidence qu'il y avait eu une erreur dans l'extraction des séquences proximales pour le maïs et le sorgho, de ce fait ces deux espèces ont été retirées de l'étude. [Figure 15]

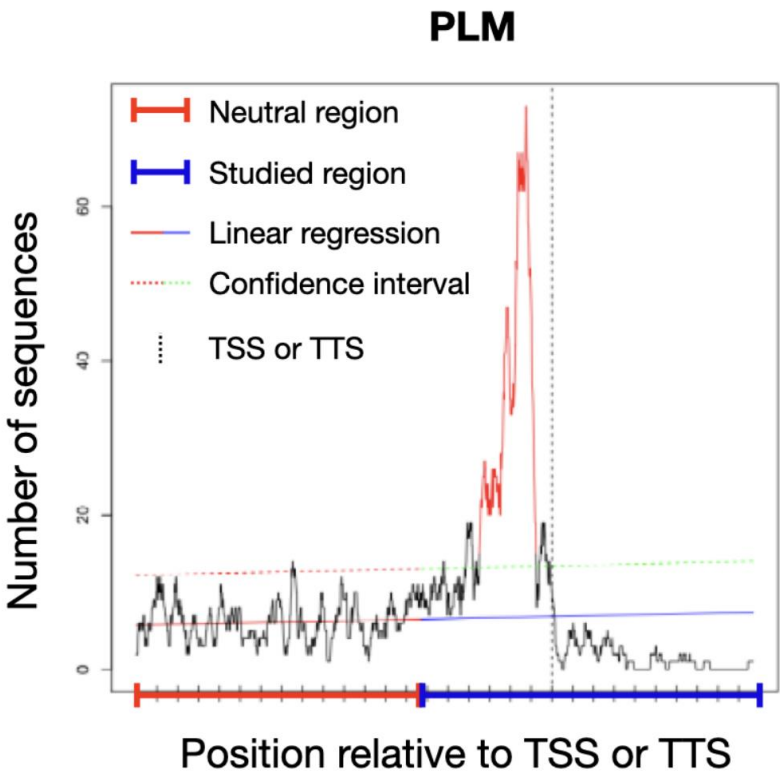


Figure 13: Position relative d'un motif par rapport au TSS ou au TTS en fonction du nombre de promoteurs dans lequel ce motif est retrouvé (à ladite position). Une régression linéaire est alors effectuée sur la région neutre [-1000pb ; -500pb] puis est étendue à la région d'étude [-500pb ; +500pb]. La position préférentielle associée au PLM est celle lue sur l'axe des abscisses là où le pic est maximal. Plus l'écart entre la ligne représentant la borne supérieure de l'intervalle de confiance et le pic est grand, plus le score du PLM est haut, signifiant ainsi que le motif présente des contraintes topologiques fortes.

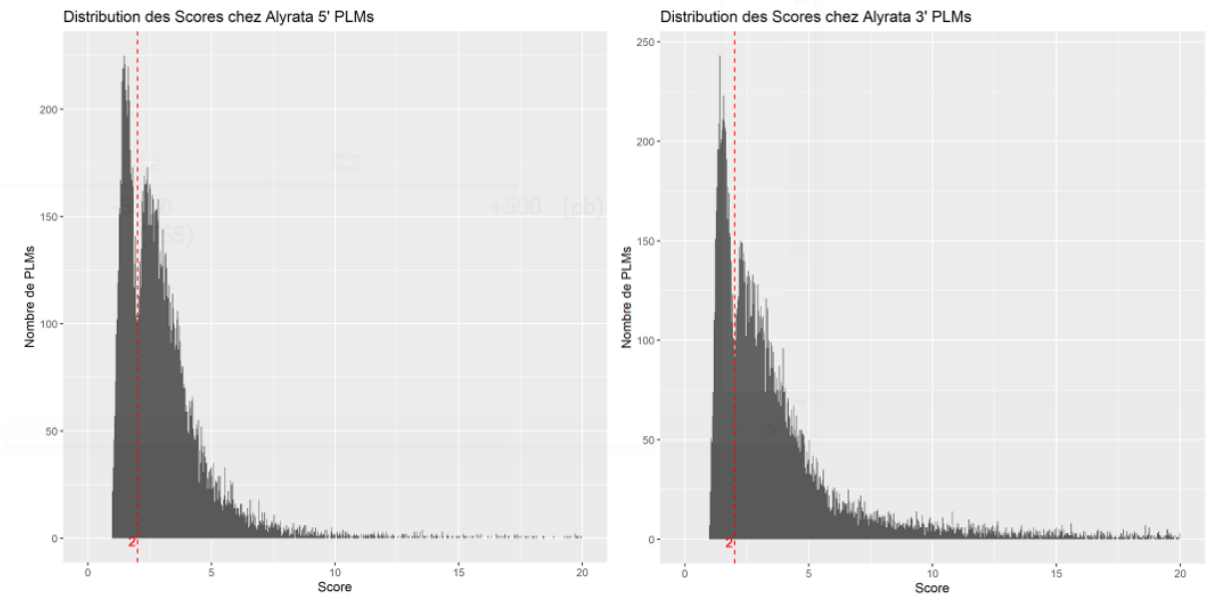


Figure 14 : Distribution des scores dans les régions proximales 5' (à gauche) et 3' (à droite) chez A.lyrata.

3.2. Analyse descriptive générale

Après filtrage, pour chacune des 18 espèces, nous avons détecté de 2725 à 7598 PLM dans les séquences 5' proximales et de 1964 à 9986 PLM dans les séquences 3' proximales [Figure 16 & Tableau 1**Error! Reference source not found.**].

Afin de vérifier si le nombre de PLM identifiés dans les séquences 5' et 3' proximales est dépendant ou non de l'espèce, un test du khi-deux a été effectué. J'ai obtenu une statistique de test χ^2=2350.3 et une p-value inférieure à 2.2e-16, ce qui permet de rejeter H0 au risque α=1% et conclure que le nombre de PLM en 5' et en 3' dépend de l'espèce.

Dans les séquences 5' proximales (et dans les séquences 3' proximales respectivement), de 1292 à 5508 (et de 1300 à 8594) des PLM identifiés étaient localisés sur ou en amont du TSS (ou TTS) représentant ainsi entre 46.5% et 78.1% (et 46% et 86.1%) des PLM détectés au sein de l'espèce [Figure 17, Tableau 2 & Tableau 3].
Afin de tester si le nombre de PLM en amont et en aval du TSS (ou du TTS) dépend de l'espèce, un test du khi-deux a été réalisé. Avec une statistique de test χ^2=4093.4 (et χ^2=4654.1) et une p-value inférieure à 2.2e-16 pour les 2 tests, on peut rejeter H0 au risque α=1% et conclure que le nombre de PLM en amont et en aval du TSS (et du TTS) est dépendante de l'espèce.

Nous avons ensuite étudié le nombre de PLM en fonction du nombre de promoteurs analysés pour chaque région proximale de chaque espèce. Nous avons observé que les points les plus à droites sur le graphique, correspondant à un nombre élevé de séquences analysées, ne sont pas les points correspondant à un grand nombre de PLM. Ainsi il ne semble pas y avoir de corrélation entre le nombre de séquences et le nombre de PLM. Le nuage de points ne permettant pas de déceler une relation explicite entre ces deux paramètres, il est possible d'affirmer que le nombre de PLM identifiés ne dépend pas du nombre de séquences analysées [Figure 18 & Tableau 4].

Par ailleurs, pour toutes les espèces la majorité des PLM identifiés sont des 8-mers pour les 2 régions proximales. En effet, ils représentent entre 48% et 74% des PLM identifiés en 5' et entre 42% et 73% des PLM identifiés en 3'. Le nombre de 4-mers est quant à lui très faible, il représente entre 0.4% et 2.8% des PLM identifiés en 5' et entre 0.6% et 3.7% de ceux identifiés en 3' [Figure 19, Tableau 5 & Tableau 6].

3.1. Organisation des PLM au sein des 18 espèces

Afin de comparer les distributions des positions préférentielles des PLM dans chaque espèce, elles ont été rapportées à une même échelle.

Dans la région 5' proximale, l'observation de la distribution préférentielle des PLM pour chaque espèce a permis de caractériser manuellement 3 types de distributions pour les 18 espèces :

- Les espèces présentant les mêmes patterns que ceux décrits dans l'étude genome-wide sur *A.thaliana* et *Z.mays*, c'est-à-dire avec un groupe de PLM localisé au niveau du TSS et 2 groupes en amont. Il s'agit de *A.lyrata*, *A.thaliana*, *M.domestica*, *M.truncatula*, *O.sativa*, *P.trichocarpa* et *P.persica* (pour *P.persica* le pic correspondant aux PLM localisés au niveau du TSS est plus faible que ceux correspondant aux PLM en amont).

- Les espèces avec 2 groupes de PLM, un localisé au niveau du TSS et l'autre en amont. Il s'agit de *B.distachyon*, *C.melo*, *H.annuus*, *H.vulgare*, *L.albus* et *T.aestivum*.

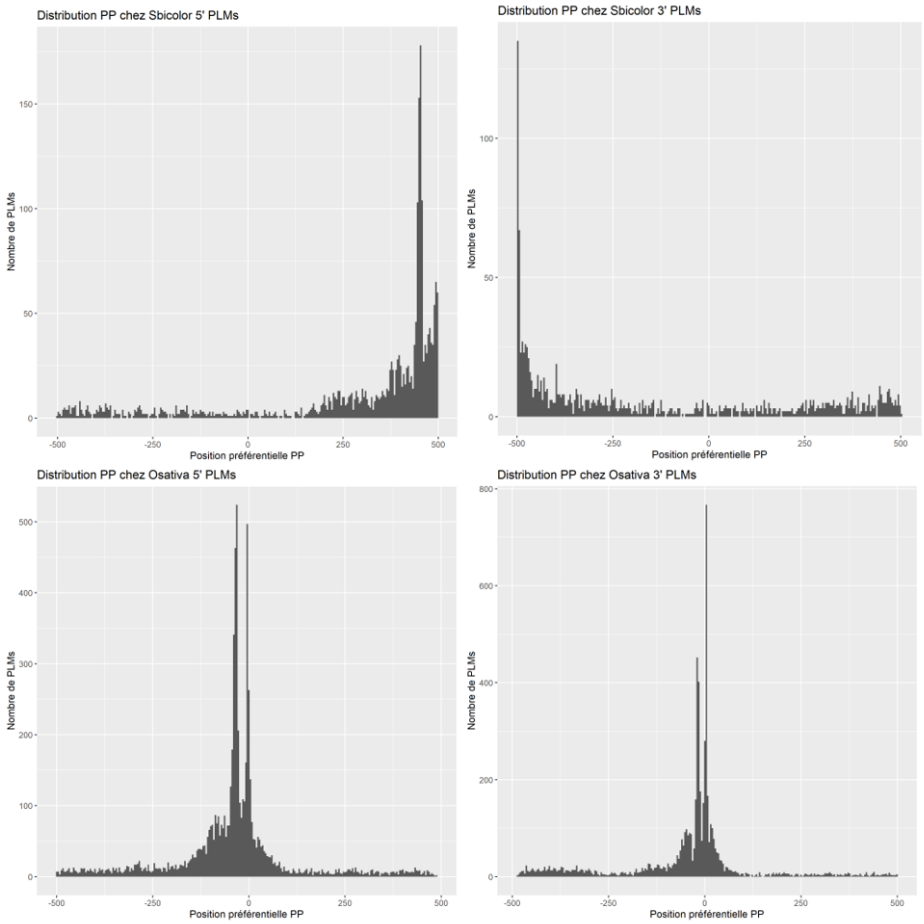


Figure 15: Distribution des PP chez S.bicolor *pour les régions 5' et 3' (en haut, à gauche et à droite respectivement) et des PP chez* O.sativa *pour les régions 5' et 3' (en bas, à gauche et à droite respectivement).*

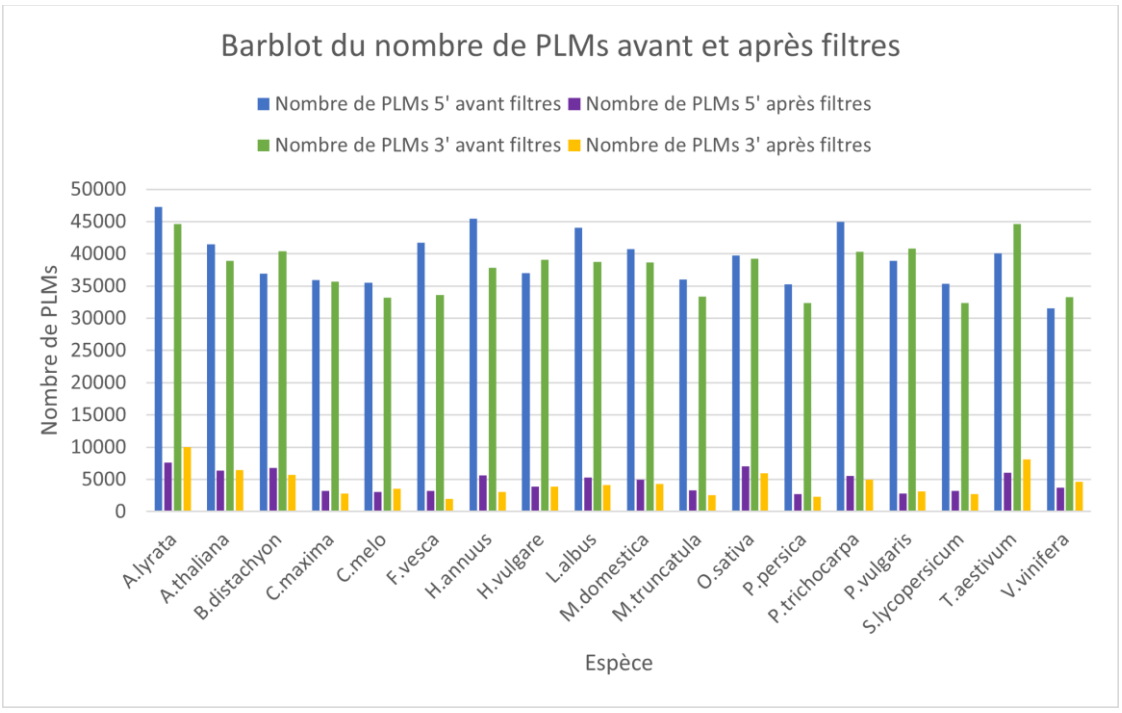


Figure 16: Barplot du nombre de PLM en 5' et en 3'avant (en bleu et vert) et après (en violet et jaune) filtrage des PLM.

- Les espèces où aucun groupe ne ressort : *C.maxima*, *F.vesca*, *P.vulgaris*, *S.lycopersicum* et *V.vinifera*. Chez *F.vesca* et *S.lycopersicum*, on observe néanmoins une majorité de PLM en amont du TSS. Pour *P.vulgaris*, on observe un groupe de PLM en amont du TSS ainsi qu'un autre groupe localisé sur [0pb ; +70 pb] du TSS.
  [Annexe - Figure 2]

Dans la région 3' proximale, l'observation de la distribution préférentielle des PLM pour chaque espèce a permis de caractériser manuellement 4 types de distributions :

- Les espèces présentant les mêmes patterns que ceux décrits dans l'étude genome-wide avec 3 groupes différents : un groupe de PLM localisés au niveau du TTS et 2 groupes en amont : *A.lyrata*, *A.thaliana*, *B.distachyon*, *L.albus*, *O.sativa* et *P.trichocarpa.*

- Les espèces avec 4 groupes de PLM : les 3 mêmes groupes que les espèces citées précédemment ainsi qu'un quatrième groupe en aval du TTS. Il s'agit de *C.melo*, *M.domestica* et *V.vinifera*.

- Les espèces avec 2 groupes de PLM, l'un sur le TTS et l'autre en aval : *C.maxima*, *H.annuus*, *M.truncatula*, *P.persica*, *P.vulgaris* et *T.aestivum*. Chez *M.truncatula* et *P.persica*, le groupe localisé au niveau du TTS est prolongé en amont.

- Les espèces où aucun groupe ne ressort : *F.vesca*, *H.vulgare* et *S.lycopersicum*. Chez *H.vulgare* les PLM sont néanmoins concentrés au niveau du TTS tandis qu'ils sont plutôt en aval du TTS chez *S.lycopersicum*.
  [Annexe - Figure 3**Error! Reference source not found.**]

Ainsi, des espèces phylogénétiquement éloignées comme *P.trichocarpa* et *O.sativa* semblent présenter des contraintes topologiques similaires autant en 5' qu'en 3'. Pour la majorité des plantes, on observe 2 groupes de PLM en amont du TSS ou du TTS et 1 groupe sur le TSS ou TTS.

3.2. Conservation des motifs entre les espèces

Nous nous sommes ensuite intéressés à la conservation des motifs au sein des espèces. En tout, 62467 motifs différents ont été identifiés. Parmi ces motifs, 43250 ont été identifiés en 5' et 33923 en 3'. Afin d'analyser cette conservation, j'ai construit une matrice pour les motifs en 5', avec 43250 motifs en ligne et 18 colonnes correspondant aux espèces de plantes ; et une autre matrice pour les motifs en 3' de 33923 lignes et 18 colonnes. Le coefficient [*i*,*j*] est égal à 1 lorsque le motif de la ligne *i* a été identifié chez l'espèce *j*, et 0 sinon. La somme en ligne indique le nombre d'espèces dans lequel chaque motif est retrouvé.

En 5', 54 motifs sont retrouvés au sein des 18 espèces dont 46 contenants le pattern « TATA », un résultat attendu puisqu'il s'agit d'un pattern spécifique aux boîtes TATA. Un total de 25423 motifs n'est retrouvé que dans une seule des 18 espèces (soit environ 59% des motifs identifiés dans les séquences 5' proximales) et 8167 motifs sont communs à au moins 3 espèces (18.9% des motifs).

Dans les séquences proximales 3', seul le motif TTCAT est retrouvé chez toutes les espèces. Cinq motifs sont retrouvés chez 17 des 18 espèces, il s'agit des motifs ATTCAT, ATTCATT, TCATT, TTCATT et TTTCATTT qui présentent tous le pattern « TCAT ». Un total de 21885 motifs est spécifique d'une seule espèce, cela représente environ 65% des motifs identifiés dans cette région, les motifs communs à au moins 3 espèces représentent 13.6% des motifs. [Figure 20]

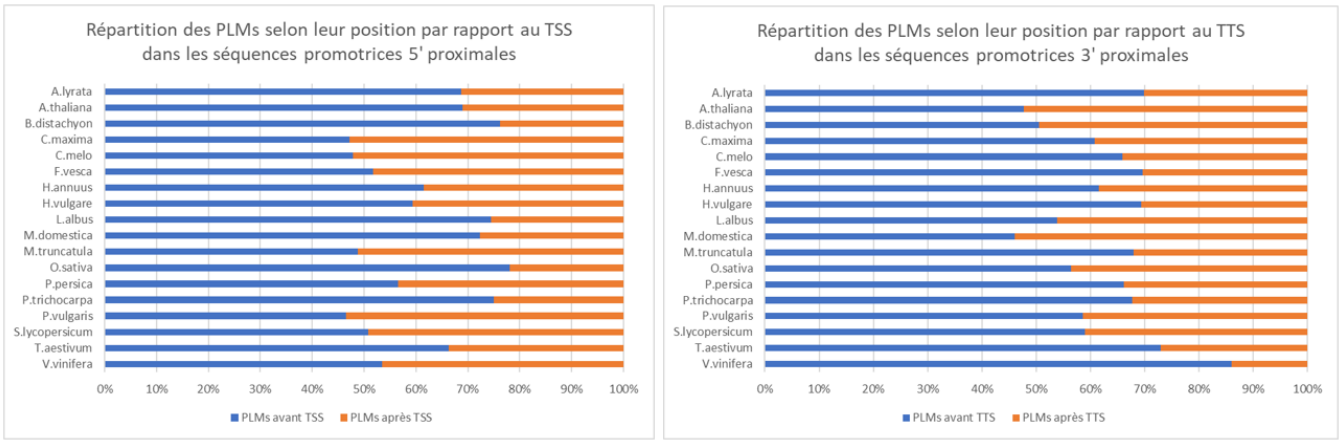


Figure 17:Répartition des PLM selon leur position par rapport au TSS (à gauche) ou au TTS (à droite) avec en bleu les PLM en amont et en orange les PLM en aval.

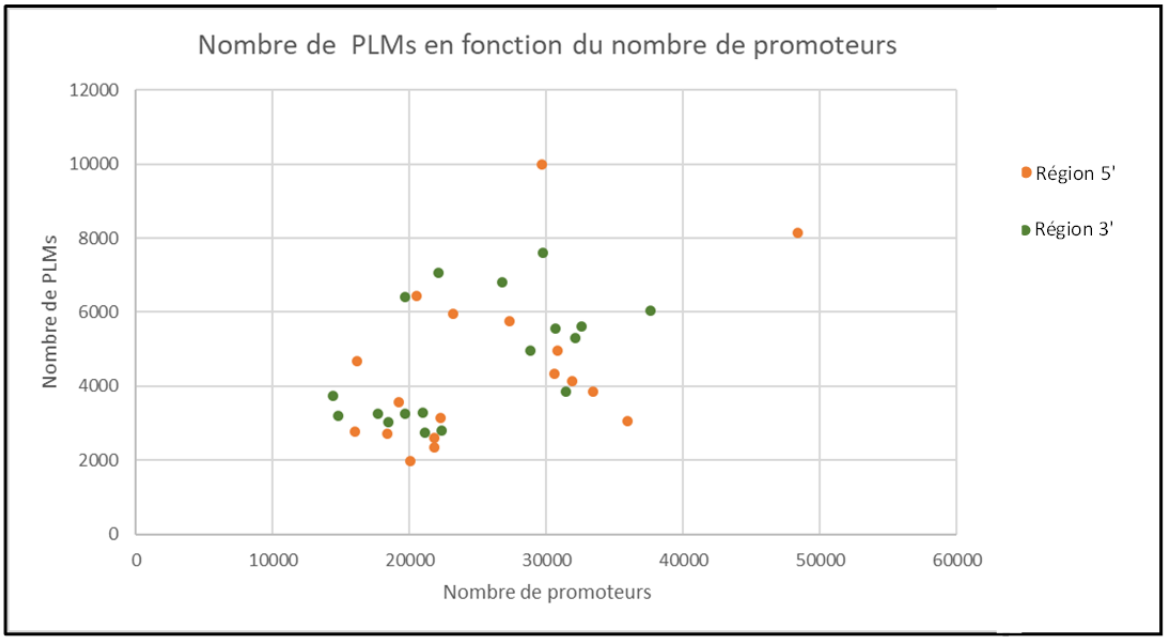


Figure 18: Scatterplot du nombre de PLM en fonction du nombre de promoteurs en 5' (en orange) et en 3' (en vert).

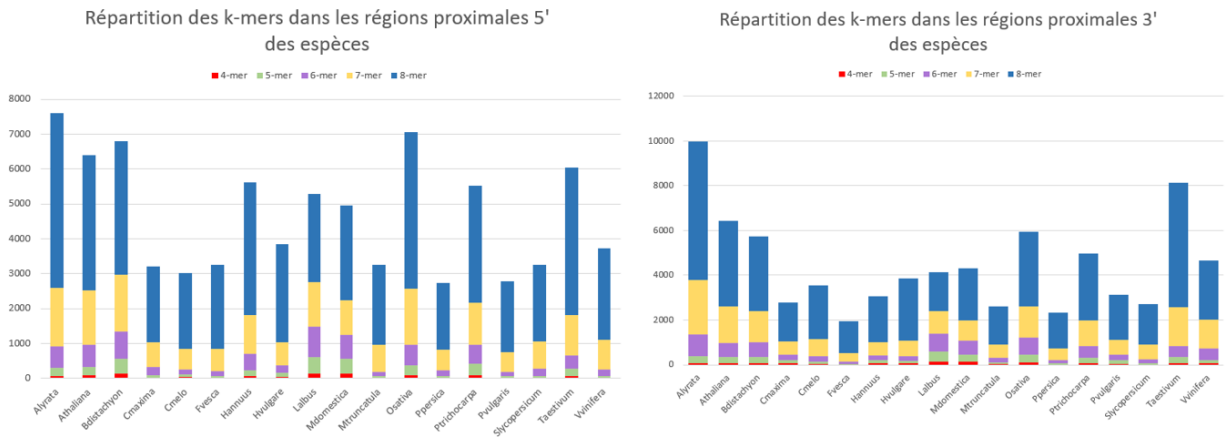


Figure 19: Barplot de la répartition des k-mers dans les régions proximales 5' (à gauche) et 3' (à droite) de chaque espèce avec en bleu les 8-mers, en jaune les 7-mers, en violet les 6-mers, en vert les 5-mers et en rouge les 4-mers.

Cette même étude a ensuite été effectuée au sein des différentes familles de plantes. Dans les séquences 5' proximales, la famille des Poaceae, qui ne compte plus que 4 espèces (puisque *Z.mays* et *S.bicolor* ont été retirées de l'étude), possède 17721 motifs différents, parmi ces motifs, 2901 (16.4%) sont retrouvés chez 2 espèces de la famille, 898 (5%) chez 3 des 4 espèces et 440 motifs (2.5%) sont communs aux 4 espèces de la famille. Dans la famille des Rosaceae, 9492 motifs sont identifiés, 851 motifs (9%) sont communs à 2 espèces et 285 (3%) sont communs aux 3 espèces de la famille. La famille des Fabaceae compte 9936 motifs différents, 897 (9%) de ces motifs sont communs à 2 des 3 espèces de la famille et seuls 248 (2.5%) sont communs aux 3 espèces. La famille des Cucurbitaceae possède 5802 motifs différents dont 426 (7.3%) communs aux 2 espèces. Enfin, la famille des Brassicaceae compte un nombre important de motifs : 11417, dont 2568 (22.5%) sont retrouvés chez les 2 espèces.

3.3. Analyse des PLM sur les régions 5' proximales

Les régions d'étude 5' proximales ont ensuite été séparées en 3 régions d'intérêt, l'une appelée R1 située en amont du promoteur central sur l'intervalle [-500pb ; -50pb[, la deuxième appelée R2 correspondant au promoteur central sur [-50pb ;0 pb] où 0 correspond au TSS et la dernière région, appelée R3 sur]0pb ;+500pb] [Figure 21].

Comme évoqué précédemment, 43250 motifs différents avaient été obtenus dans les régions 5'proximales. Ces motifs sont répartis de la manière suivante : 17133 motifs dans la région R1, 12564 dans R2 et 22770 en R3. En 5', le nombre de PLM (donc le motif associé à une position préférentielle) était de 84283 : 27267 PLM dans la région R1, 27135 dans R2 et 29882 dans R3. Ainsi, la région R2, de seulement 50pb contient autant de PLM que la région R1 de 400bases de plus. Un barplot représentant le nombre de PLM par région pour chaque espèce a ensuite été généré. On observe alors que, pour la plupart des espèces, une majorité de PLM se trouve dans R2 et que les espèces avec moins de 5000 PLM ont aussi une minorité de PLM en R2 [Figure 22].

Les distributions des positions préférentielles des PLM par région chez toutes les espèces ont été générées. En R1, les PLM sont majoritairement localisés sur la région [-150pb ; -50pb[. La distribution des PLM sur la région [-500pb ;-150pb[est homogène. Dans la région R2, on observe un pic de PLM à la position -50pb ainsi qu'un groupe de PLM sur la région [-38pb ; -29pb] correspondant à la région des boîtes TATA. La plupart des PLM de la région R2 sont localisés sur la région [-6pb ; 0pb] du TSS correspondant à la région des motifs initiateurs. Sur R3, la majorité des PLM sont localisés sur] 0pb ;+100pb] et dans cet intervalle, la plupart des PLM sont localisés à +1 base du TSS. Sur le reste de la région R3, la distribution des positions préférentielles des motifs est homogène. [Figure 23]

Je me suis ensuite intéressée aux distributions des positions préférentielles par région espèce par espèce. Sur R1 on observe toujours une accumulation de PLM de [-100pb ; -50pb[. Sur R2, on observe chez *A.lyrata*, *A.thaliana*, *B.distachyon*, *P.trichocarpa* et *O.sativa* 2 groupes de PLM, l'un sur l'intervalle [-39pb ; -31pb] et l'autre sur [-5pb ; 0pb]. Chez *H.annuus*, *H.vulgare*, *L.albus*, *M.domestica* et *T.aestivum* on observe un groupe de PLM aux alentours de -50pb et un second sur [-6pb ; 0pb]. Les espèces restantes présentent peu de PLM sur R2, ainsi le paysage des distributions de positions préférentielles observé est plat. Sur R3, on observe un groupe important de PLM localisés à 1pb du TSS chez *B.distachyon*, *H.vulgare*, *M.truncatula*, *O.sativa* et *T.aestivum.* Chez *C.maxima*, on observe une concentration plus importante de PLM sur [+90pb ; +100pb] du TSS que dans le reste de la région R3. C'est la seule espèce chez qui ce pic est retrouvé. Pour les autres espèces, la distribution des PLM est plutôt homogène sur l'ensemble de la région. [Annexe - Figure 4 & Annexe - Figure 5]

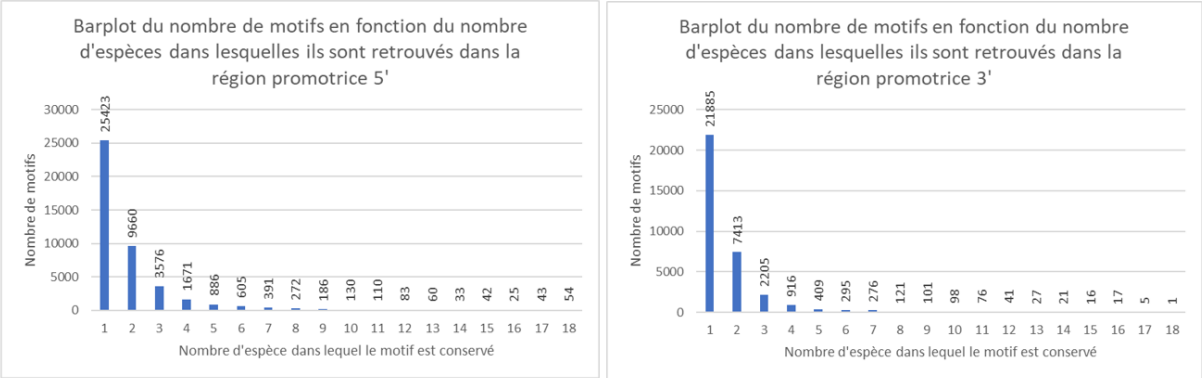


Figure 20:Barplot du nombre de motifs en fonction du nombre d'espèces dans lesquelles ils sont retrouvés en 5' (à gauche) et en 3' (à droite).

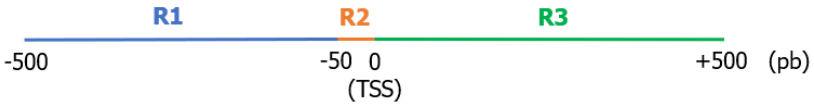


Figure 21: Découpage des régions 5' proximales en 3 régions, R1 en bleu, R2 en orange et R3 en vert.

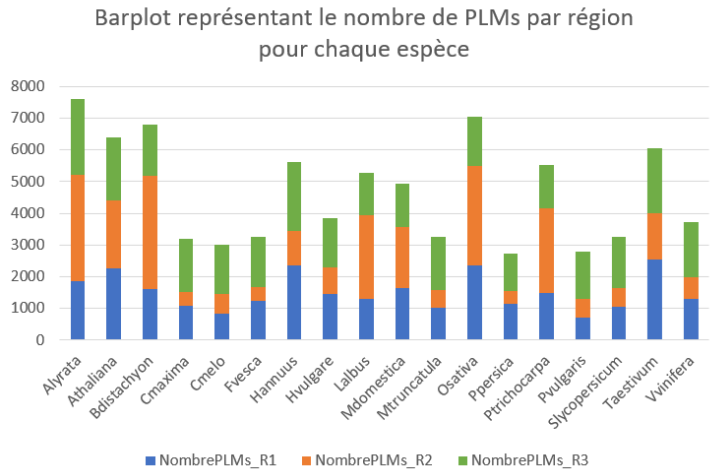


Figure 22: Barplot représentant le nombre de PLM par région pour chaque espèce (avec R1 en bleu, R2 en orange et R3 en vert.)

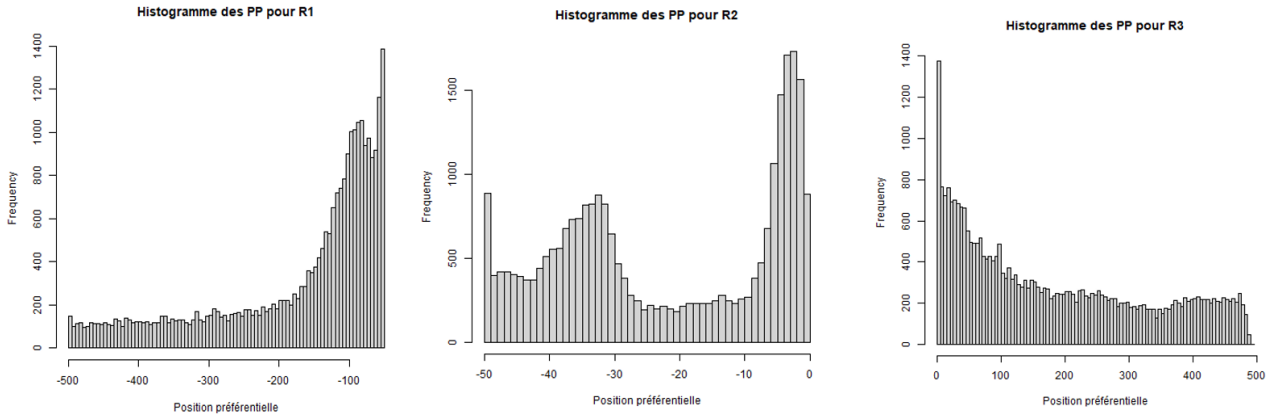


Figure 23: Histogrammes des PP des PLM sur R1 (à gauche), R2 (au milieu) et R3 (à droite).

Toutes ces distributions, pour chaque région et pour chaque espèce, ont alors été comparées 2 à 2 grâce à un test de Kolmogorov-Smirnov afin de voir si elles suivaient la même loi. Pour chacun de ces tests réalisés au seuil α =5%, la p-value était inférieure à 0.05 et on peut conclure que les distributions ne suivent pas la même loi.

Je me suis ensuite focalisée sur la région R2 dans laquelle se trouvent les boîtes TATA et les motifs initiateurs. L'objectif était de voir si les boîtes TATA sont conservées chez toutes les espèces aux mêmes positions et si des motifs initiateurs sont retrouvés chez toutes les espèces mais aussi de voir si de nouveaux motifs sont retrouvés dans cette région.

Ainsi, les boîtes TATA ont été étudiées chez les 18 espèces en recherchant les séquences consensus TATAWA (TATATA et TATAAA). La séquence TATATA est retrouvée chez les 18 espèces, de -30pb à -48pb La séquence TATAAA est retrouvée chez toutes les espèces sauf *M.truncatula*. Elle est conservée de -30pb à -49pb. Au sein d'une même espèce, la variabilité de la position préférentielle entre ces 2 séquences est comprise entre 0 et 4 bases. [Tableau 7]

Les boîtes TATA présentent donc, au sein d'une même espèce, de fortes contraintes topologiques avec une faible variabilité. Au sein des 18 espèces, les boîtes TATA seraient ainsi localisées dans l'intervalle [-49pb ; -30pb]. La position des TATAWA est donc conservée au sein d'une même espèce, on observe peu de variabilité, en revanche la position de ces séquences entre espèces peut varier

Ensuite, 37 autres hexamères identifiés comme boîtes TATA chez *A.thaliana* [10] ont été étudiés. Plus de 30 de ces motifs sont retrouvés chez *A.lyrata, A.thaliana, B.distachyon, O.sativa* et *P.trichocarpa*. Seuls 6, 7 et 11 de ces 6-mers sont retrouvés respectivement chez *M.truncatula, C.maxima* et *L.albus* [Figure 24]. Certains motifs comme CGCCTA, TCTCTA, CTCCTA, TCCCTA et CTTCTA sont retrouvés *chez A.thaliana* avec un suffixe T. Ces 7-mers se trouvent sur l'intervalle [-41pb ; -38pb], il s'agit donc bien de boîtes TATA.

Grâce aux positions préférentielles de cette liste de 37 hexamères ainsi que celles des 2 motifs TATAWA, il est possible d'établir que les boîtes TATA seraient retrouvées dans la région [-50pb ;-24pb] chez toutes les espèces. En recherchant sur R1 les hexamères de la liste étant au moins conservés chez 9 espèces sur R2 [Figure 25] , on peut redéfinir les intervalles des régions des boîtes TATA pour 6 espèces. Toutes espèces confondues, les boîtes TATA seraient ainsi retrouvées sur [-60pb ; -24pb]. Ainsi, bien que chez certaines espèces la région des boîtes TATA est légèrement différente de [-39pb ; -26pb] identifiée chez l'espèce modèle *A.thaliana*, elle reste tout de même autour de la région attendue [Tableau 8].

Les séquences TATAWA dégénérées d'une base sont appelées variants TATAboxΔ1. Toutes les séquences proches de TATAWA ne sont pas fonctionnelles mais *in vitro*, les variants ayant une seule substitution sont les plus observés. Cela représente 32 variants de la boîte TATA dont 11 avaient été identifiés comme communs entre l'arabette des dames et le riz [2]. Dans notre analyse, nous observons une grande variabilité du nombre de ces variants entre les espèces : alors que *B.distachyon* présente 24 de ces 32 variants, *M.truncatula* n'en présente qu'un seul. [Figure 26 & Tableau 9]

Les motifs initiateurs de séquence consensus YTCAY mais aussi YYANWYY, YANWYY et YYANWY ont ensuite été recherchés dans R2. Cela représentait alors 507 PLM dont les positions préférentielles variaient de -50pb à 0pb. Les motifs initiateurs ne pouvant se trouver à -50pb du TSS, ils ont été filtrés selon leur PP. La distribution des positions préférentielles de ces motifs a alors révélé qu'une majorité était située sur la région [-6pb ;0pb], ainsi seuls les motifs de cette région ont été gardés comme motifs initiateurs [Figure 27].

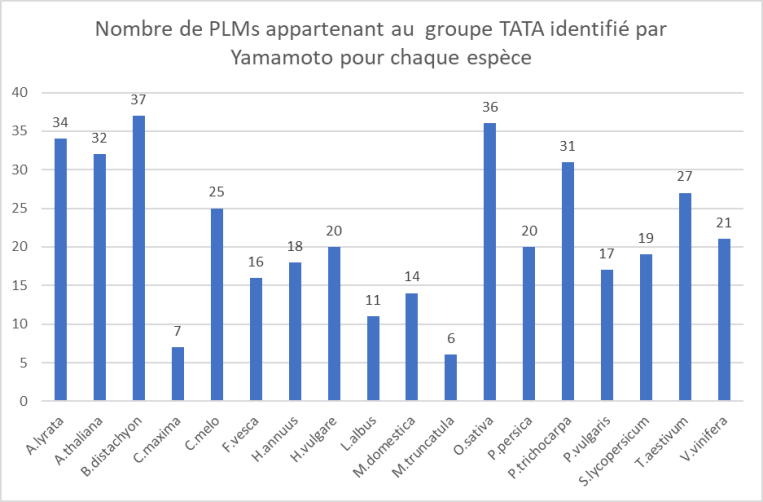


Figure 24:Barplot du nombre de PLM appartenant au groupe de boîtes TATA identifié par Yamamoto pour chaque espèce sur R2.

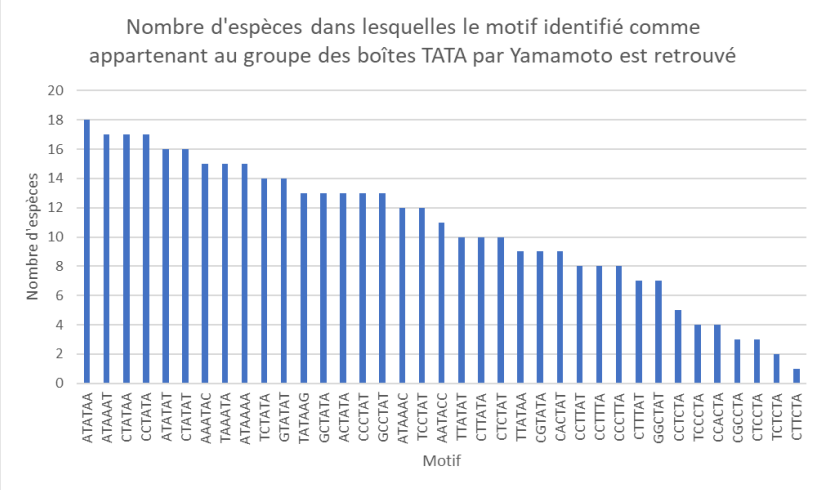


Figure 25 :Barplot du nombre d'espèces dans lesquelles chaque motif identifié comme boîte TATA par l'étude menée par Yamamoto et al est retrouvé sur R2.

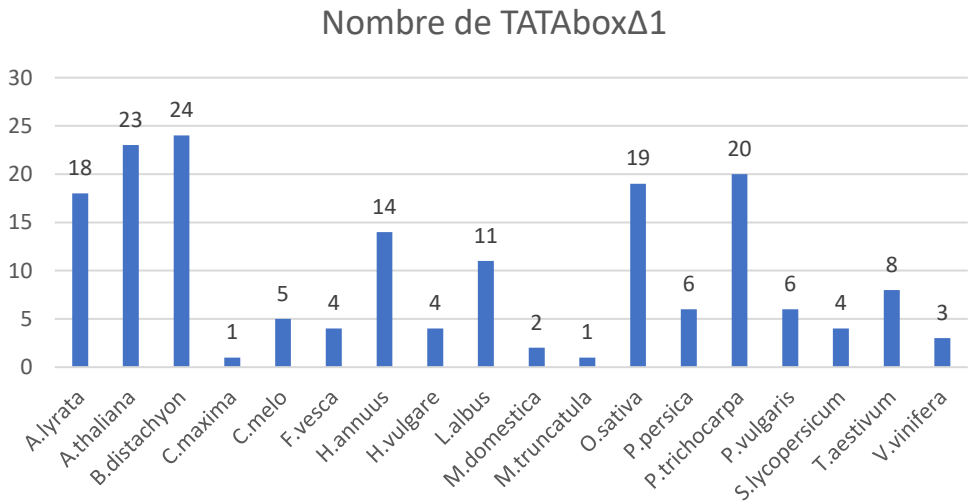


Figure 26: Barplot du nombre de variants Δ1 des boîtes TATA dans chacune des espèces.

Cela a mené à l'identification de 310 PLM ensuite filtrés selon la taille de leur fenêtre fonctionnelle, devant être inférieure ou égale à 15pb. Ainsi, nous avons obtenu 250 PLM comprenant 117 motifs différents. Parmi ces motifs, 2 sont communs à 6 espèces, 8 à 5 espèces, 6 à 4 espèces, 20 à 3 espèces et 33 à 2 espèces. Parmi les motifs initiateurs conservés chez un tiers des espèces on retrouve CCAAAT qui est retrouvé de -2pb à -5pb et le motif TCATTC retrouvé à -2pb et -3pb. Le nombre de motifs initiateurs trouvé dans chaque espèce varie beaucoup, avec ces paramètres, assez stringents, aucun motif initiateur n'a été identifié chez *C.maxima*, *F.vesca*, *H.vulgare*, *M.truncatula*, *P.persica*, *P.vulgaris*, *S.lycopersicum* et *V.vinifera* et seulement 1 chez *T.aestivum*. L'unique PLM initiateur identifié chez *T.aestivum* est CAGTTT, il est aussi retrouvé chez *M.domestica* et *L.albus* à -2pb du TSS. Le nombre de motifs initiateurs identifiés chez *L.albus* est quant à lui plutôt élevé puisqu'on en retrouve 79. [Figure 28 & Tableau 10]

Enfin, les dernières analyses réalisées ont visé à identifier l'ensemble des motifs PLM de la région R2 ne correspondant ni aux boîtes TATA (séquences TATAWA, liste des 37 hexamères identifiés comme appartenant au groupe des boîtes TATA ainsi que les variants TATAboxΔ1) ni aux motifs initiateurs. Pour cela, en plus des motifs explicités dans les paragraphes précédents, les motifs contenant la séquence « TATA » ou « TAAA » et localisés sur l'intervalle [-50pb ; -24pb] ont aussi été retiré de la liste (-24pb étant la PP comprise dans les intervalles des régions des boîtes TATA établis précédemment la plus proche du TSS). Un total de 5755 motifs est gardé. Afin d'identifier de potentiels évènements de régulation associés à ces motifs, je les ai comparés, grâce à l'algorithme de comparaison de motifs TOMTOM, aux TFBS répertoriés dans la base de données JASPAR PLANT 2022. Cette comparaison a révélé que 562 motifs sont similaires à des TFBS et que ceux-ci peuvent être ciblés par 502 FT différents. Ces FT sont répartis dans 33 familles différentes. On trouve en majorité des FT impliqués dans les clusters de tryptophane dont les MYB et MYB-related mais aussi des facteurs hélice-boucle-hélice ainsi que des AP2/EREBP et des glissières à leucine. Toutes ces familles de FT sont impliquées dans le développement de la plante ou les réponses au stress biotique et environnemental [Figure 29 & Tableau 11].

4. Discussion – Perspectives

4.1. PLMdetect

La méthode PLMdetect repose essentiellement sur l'annotation des positions du TSS ou du TTS. Ainsi, plus une espèce est bien annotée (comme l'espèce modèle *A.thaliana* ou l'espèce très étudiée *O.sativa*) et plus on a de l'information sur les motifs présentant des contraintes topologiques chez cette espèce. A l'inverse, chez les espèces moins bien annotées comme la fraise *F.vesca* pour laquelle la qualité d'assemblage du génome est faible avec des séquences intergéniques manquantes, le nombre de PLM identifié est fortement diminué. Pour le blé *T.aestivum*, possédant plus de 100 000 gènes, seules 35% des séquences proximales ont été analysées en 5' et 45% en 3' ; de même pour *M.truncatula* qui, bien que possédant moitié moins de gènes, ne voit que 41% et 43% de ses séquences proximales respectivement 5' et 3' considérées. [Figure 30]

Ainsi, puisque lorsque le génome d'une espèce est mal annoté, le nombre de PLM identifié, particulièrement au niveau du promoteur central, est significativement diminué, une perspective possible serait alors d'utiliser l'approche PLMdetect pour évaluer la qualité des assemblages et l'annotation de génomes de plantes.

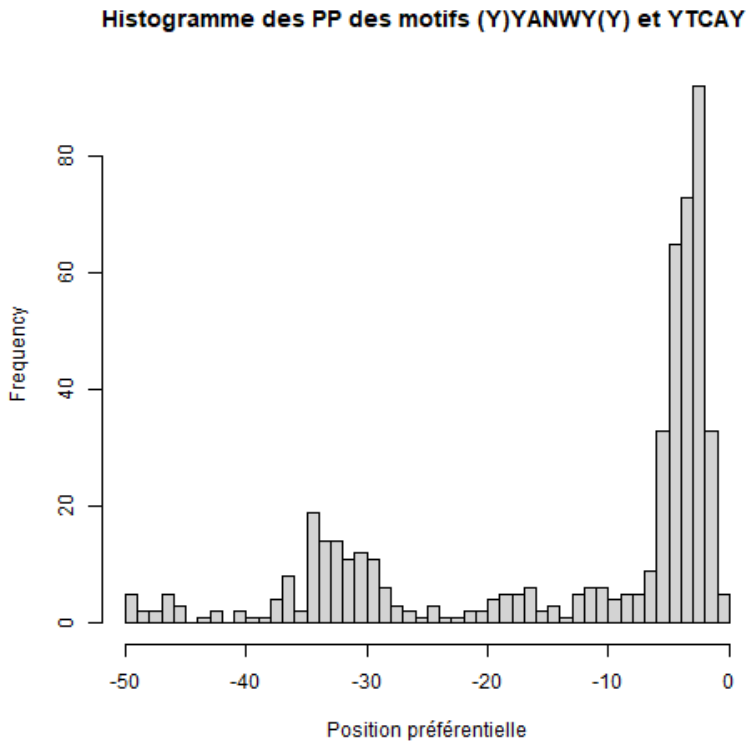


Figure 27:Histogramme des positions préférentielles des motifs YYANWYY, YYANWY, YANWYY et YTCAY sur R2.

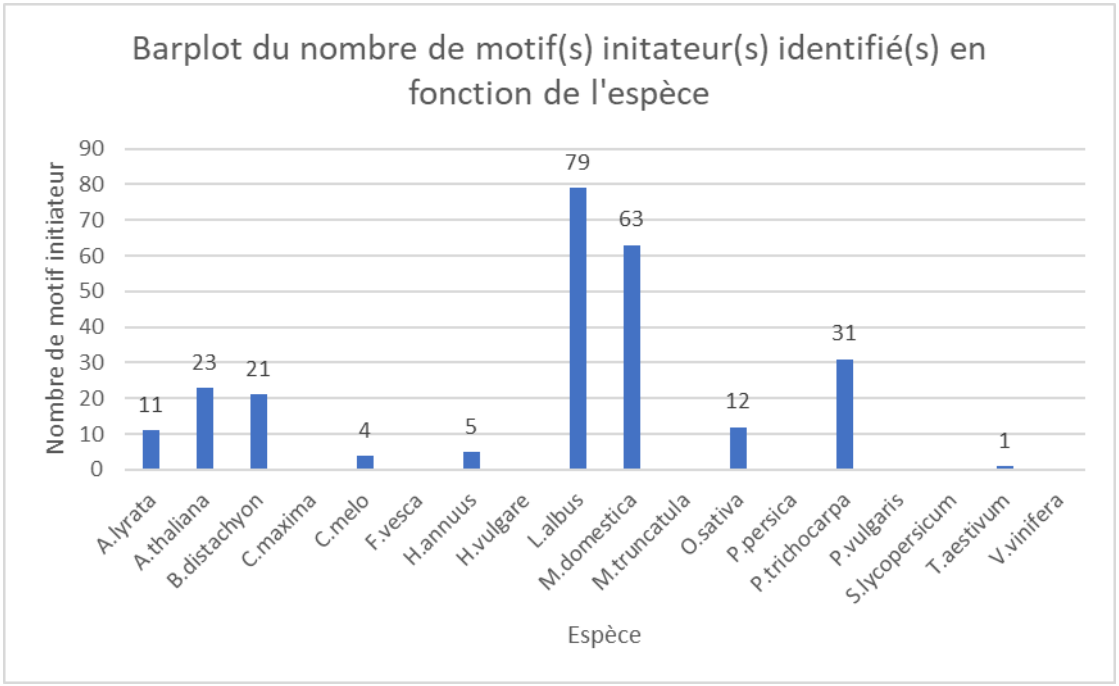


Figure 28: Barplot du nombre de motif initiateur identifié chez chaque espèce.

4.2. Résultats obtenus

Une différence notable entre l'analyse comparative que j'ai effectuée et celle menée dans l'étude genome-wide sur *A.thaliana* et *Z.mays* est que, dans cette dernière, les motifs avaient été reverse-transcrits. Par exemple les motifs ATCG et CGAT étaient considérés comme un seul PLM, là où, dans mon analyse, les PLM tiennent compte du sens transcriptionnel. Chez certains motifs, comme les boîtes TATA ou les motifs initiateurs, le sens transcriptionnel est très important. Ainsi, comparer les PLM identifiés chez *A.thaliana* dans ces 2 études pourrait mettre en évidence les motifs orientés et vérifier que les boîtes TATA et motifs initiateurs sont effectivement orientés.

Par ailleurs, le test de Kolmogorov Smirnov effectué pour comparer les différentes distributions obtenues n'est peut-être pas le plus adapté puisqu'il ne tient pas compte des ex aequo et donc ne comptabilise pas la surreprésentation de certaines positions préférentielles. Aussi, puisque le même test est réalisé plusieurs fois, il faudrait appliquer la méthode de Bonferroni afin de corriger le seuil de significativité pour contrôler les faux positifs.

Enfin, parmi les motifs qui n'appartenaient ni aux motifs initiateurs ni aux motifs répertoriés en tant que boîtes TATA dans mon analyse, et n'ayant pas de correspondance avec les TFBS répertoriés dans JASPAR 2022 on retrouve, chez les 18 espèces, le 7-mer AAATACC, le 8-mer AAATACCC commun à 16 des 18 espèces ainsi que l'hexamère AAATAC conservé chez 15 espèces. Chez toutes les espèces ces PLM sont préférentiellement localisés dans la région des boîtes TATA identifiée pour chaque espèce. Ainsi, l'hexamère AAATAC pourrait lui aussi être une séquence reconnue par les TBP.

4.3. Perspectives

Dans les deux mois de stage restants, je vais continuer à analyser les données obtenues en commençant par effectuer une analyse d'enrichissement des termes d'annotations sur les motifs ni TATA ni initiateurs les plus conservés entre les 18 espèces. Cette analyse serait réalisée en utilisant des tests statistiques afin d'identifier, dans la liste des termes d'annotations GO annotant les gènes associés aux motifs, les termes les plus enrichis.

L'analyse d'enrichissement est d'autant plus importante qu'elle a permis, dans l'étude menée sur l'arabette et le maïs, d'inférer une fonction à une famille de 65 gènes (HOM04M002476 sur PLAZA) peu annotée chez le maïs, ces gènes étant ciblés par des FT des familles AP2/ERF, Myb-related et WRKY impliqués dans la régulation transcriptionnelle ainsi que la réponse aux stimulus mais aussi d'autres FT impliqués dans la réponse aux brassinostéroïdes, des hormones stéroïdiennes essentielles à la croissance et au développement des plantes ainsi que des FT impliqués dans la méthylation, la réponse aux différents niveaux de nutriments ainsi que le développement primaire des racines.

Par la suite, les comparaisons pourront être appliquées aux régions R1 et R3 dans le but de révéler d'autres PLM fortement conservés au sein des angiospermes.

Enfin, il serait aussi possible d'effectuer le même travail mais sur les séquences 3' proximales, assez peu étudiées et investiguer le potentiel rôle du pattern « TCAT » retrouvé au sein des 18 espèces de l'étude.

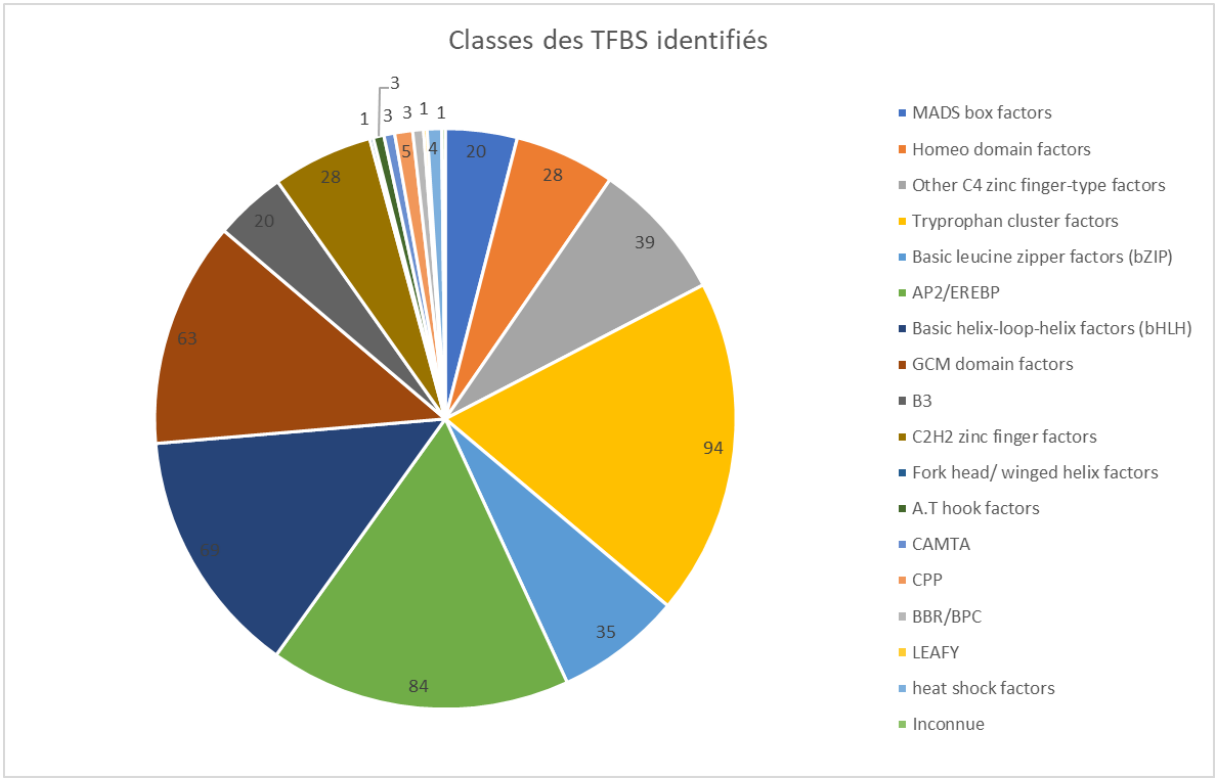


Figure 29: Camembert représentant les classes des FT fixant des TFBS identifiés comme similaires aux PLM n'appartenant ni aux boîtes TATA ni aux motifs initiateurs sur R2.

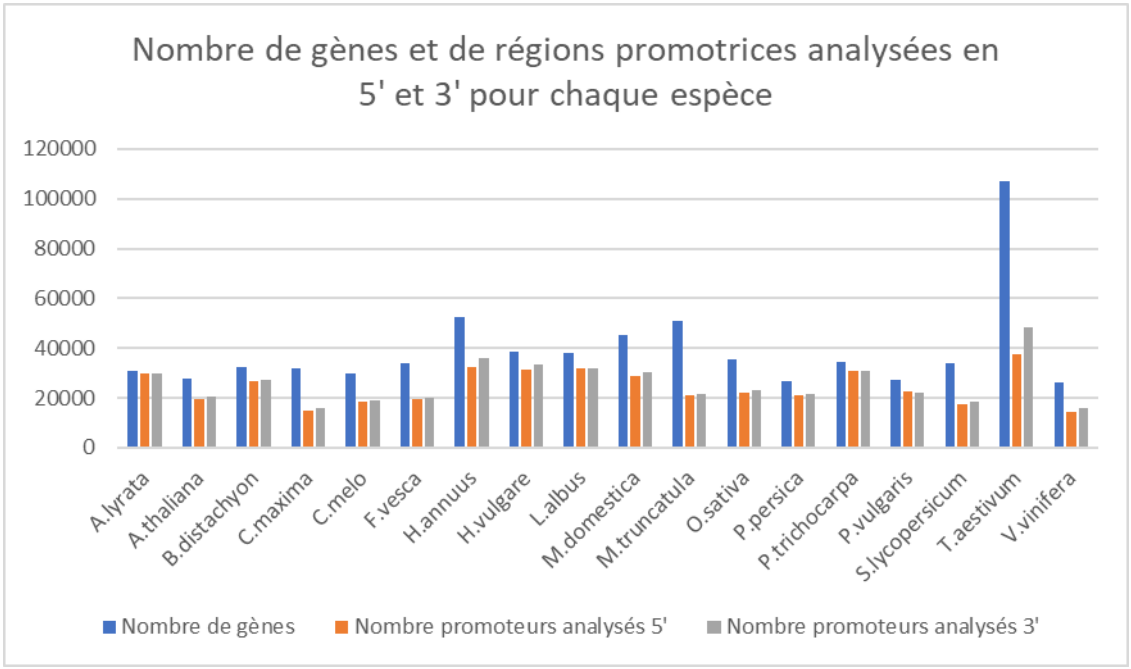


Figure 30: Barplot du nombre de gènes (en bleu) et de séquences proximales analysées en 5' (en orange) et 3'(en gris) pour chaque espèce.

5. Conclusion

Pour conclure, au cours de ces deux premiers mois de stage, j'ai pu mettre en avant que chacune des plantes de l'étude possédait des motifs présentant des contraintes topologiques fortes. Ces PLM sont principalement localisés autour du TSS ou du TTS. Au niveau des séquences 5' proximales et plus particulièrement du promoteur central, on retrouve 2 régions avec une forte accumulation de PLM, la région [-52pb ; -27pb] dans laquelle on trouve les boîtes TATA pour l'ensemble des 18 espèces ainsi que la région [-6pb ; 0pb] dans laquelle on trouve les motifs initiateurs. J'ai ainsi pu caractériser les régions dans lesquelles se trouvaient les boîtes TATA pour chaque espèce et identifier un nouvel hexamère potentiellement reconnu par les TBP. J'ai aussi pu mettre en avant des PLM dans ces régions similaires à des TFBS et d'autres PLM non caractérisés mais qui représentent des candidats intéressants à explorer.

Ce stage m'a permis de continuer mon apprentissage du langage de programmation R et de voir de nouveaux outils bio-informatiques tel que l'algorithme de comparaison de motifs TOMTOM. J'ai aussi pu découvrir l'existence des méthodes de détection de motif *in silico* que ce soit à travers la méthode PLMdetect ou mes recherches sur les différents algorithmes existants.

J'ai aussi pu en apprendre plus sur les plantes ainsi que sur la génomique et la transcriptomique grâce aux réunions scientifiques auxquelles j'ai assisté ainsi qu'aux journées plateformes pendant lesquelles des chercheurs présentent leur travail, organisées au sein de l'IPS2.

Bibliographie

[1] Bernard, V., Brunaud, V. & Lecharny, A. TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics* 11, 166 (2010). https://doi.org/10.1186/1471-2164-11-166

[2] Virginie Bernard. Relations entre l'organisation des sites de fixation des facteurs de transcription, la fonction des gènes et l'expression des gènes chez Arabidopsis thaliana: vers une annotation des sites de fixation. Biologie végétale. Université d'Evry-Val d'Essonne, 2009. Français. ⟨tel-00444896⟩

[3] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, Keji Zhao, Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data, *Nucleic Acids Research*, Volume 36, Issue 16, September 2008, Pages 5221–5231, https://doi.org/10.1093/nar/gkn488

[4] Gary D. Stormo, DNA binding sites: representation and discovery, *Bioinformatics*, Volume 16, Issue 1, January 2000, Pages 16–23, https://doi.org/10.1093/bioinformatics/16.1.16

[5] Chadi Saad. Caractérisation des erreurs de séquençage non aléatoires, application aux mosaïques et tumeurs hétérogènes. Bio-informatique [q-bio.QM]. Université de Lille Nord de France, 2018. Français. ⟨tel-01936291⟩

[6] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Research*, Volume 37, Issue suppl_2, 1 July 2009, Pages W202–W208, https://doi.org/10.1093/nar/gkp335

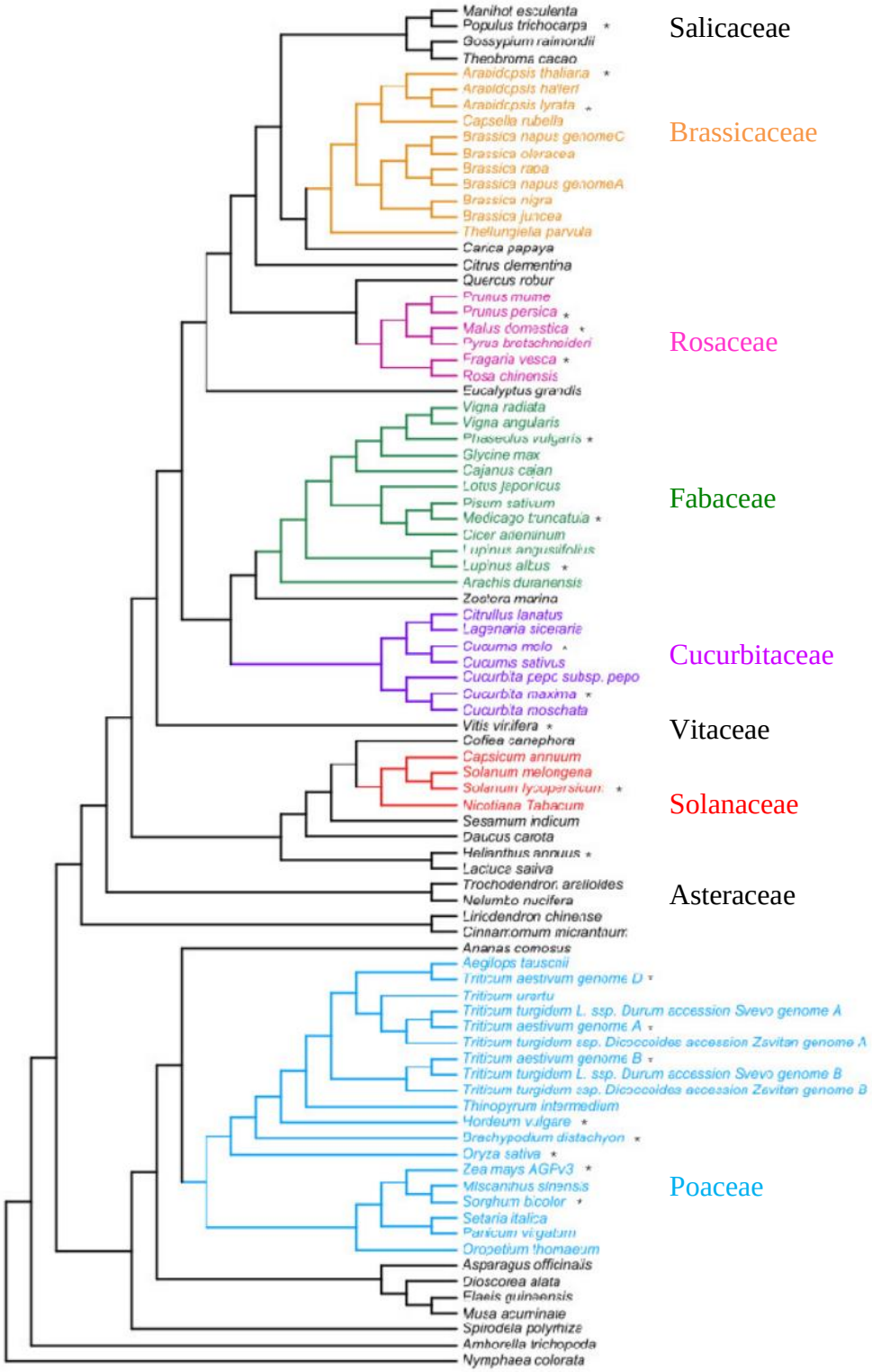
[7] Rozière J, Guichard C, Brunaud V, Martin M-L, Coursol S (2022). A comprehensive map of preferentially located motifs reveal novel proximal *cis*-regulatory elements in plants, *preprint on biorxiv*, https://doi.org/10.1101/2022.01.17.476590

[8] Bernard, V., Lecharny, A., & Brunaud, V. (2010). Improved detection of motifs with preferential location in promoters. *Genome*, *53*(9), 739–752, https://doi.org/10.1139/g10-042

[9] Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, Fornes O, Leung TY, Aguirre A, Hammal F, Schmelter D, Baranasic D, Ballester B, Sandelin A, Lenhard B, Vandepoele K, Wasserman WW, Parcy F, and Mathelier A JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles *Nucleic Acids Res.* 2022 Jan 7;50(D1):D165-D173.; doi: 10.1093/nar/gkab1113

[10] Yamamoto YY, Ichida H, Matsui M, et al. Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics*. 2007;8:67. Published 2007 Mar 8. doi:10.1186/1471-2164-8-67

Annexe



Annexe - Figure 1: Arbre phylogénétique des 20 espèces de plantes étudiées, représentées par un astérisque. Les différentes familles auxquelles elles appartiennent sont représentées par des couleurs différentes et sont répertoriées à droite.

Espèce	Nombre de PLMs 5'		Nombre de PLMs 3'	
	avant filtres	après filtres	avant filtres	après filtres
A.lyrata	47268	7598	44602	9986
A.thaliana	41466	6387	38914	6424
B.distachyon	36901	6789	40436	5732
C.maxima	35901	3202	35635	2770
C.melo	35501	3024	33186	3558
F.vesca	41687	3244	33586	1964
H.annuus	45421	5605	37818	3056
H.vulgare	37017	3851	39053	3845
L.albus	44039	5288	38733	4134
M.domestica	40724	4944	38671	4323
M.truncatula	35997	3263	33325	2594
O.sativa	39708	7053	39253	5948
P.persica	35296	2725	32333	2328
P.trichocarpa	44973	5532	40342	4962
P.vulgaris	38870	2778	40825	3125
S.lycopersicum	35361	3244	32365	2711
T.aestivum	40029	6041	44638	8118
V.vinifera	31561	3715	33249	4662

Tableau 1:Nombre de PLM identifié en 5' et en 3' avant et après filtres

Espèce	Nombre de PLM avant TSS	% de PLM en amont du TSS	Nombre de PLM après TSS	% de PLM en aval du TSS	Nombre de PLM en 5'
V.vinifera	1988	53,5	1727	46,5	3715
T.aestivum	4005	66,3	2036	33,7	6041
S.lycopersicum	1648	50,8	1596	49,2	3244
P.vulgaris	1292	46,5	1486	53,5	2778
P.trichocarpa	4152	75,1	1380	24,9	5532
P.persica	1542	56,6	1183	43,4	2725
O.sativa	5508	78,1	1545	21,9	7053
M.truncatula	1592	48,8	1671	51,2	3263
M.domestica	3574	72,3	1370	27,7	4944
L.albus	3936	74,4	1352	25,6	5288
H.vulgare	2286	59,4	1565	40,6	3851
H.annuus	3443	61,4	2162	38,6	5605
F.vesca	1680	51,8	1564	48,2	3244
C.melo	1446	47,8	1578	52,2	3024
C.maxima	1510	47,2	1692	52,8	3202
B.distachyon	5172	76,2	1617	23,8	6789
A.thaliana	4407	69	1980	31	6387
A.lyrata	5221	68,7	2377	31,3	7598

Tableau 2: Nombre et pourcentage de PLM en amont et en aval du TSS pour chaque espèce.

Espèce	Nombre de PLM avant TTS	% de PLM en amont du TTS	Nombre de PLM après TTS	% de PLM en aval du TTS	Nombre de PLM en 3'
V.vinifera	8594	86,1	1392	13,9	9986
T.aestivum	4689	73	1735	27	6424
S.lycopersicum	3378	58,9	2354	41,1	5732
P.vulgaris	1624	58,6	1146	41,4	2770
P.trichocarpa	2410	67,7	1148	32,3	3558
P.persica	1300	66,2	664	33,8	1964
O.sativa	1725	56,4	1331	43,6	3056
M.truncatula	2610	67,9	1235	32,1	3845
M.domestica	1903	46	2231	54	4134
L.albus	2330	53,9	1993	46,1	4323
H.vulgare	1799	69,4	795	30,6	2594
H.annuus	3663	61,6	2285	38,4	5948
F.vesca	1621	69,6	707	30,4	2328
C.melo	3269	65,9	1693	34,1	4962
C.maxima	1899	60,8	1226	39,2	3125
B.distachyon	1370	50,5	1341	49,5	2711
A.thaliana	3868	47,6	4250	52,4	8118
A.lyrata	3258	69,9	1404	30,1	4662

Tableau 3:Nombre et pourcentage de PLM en amont et en aval du TTS pour chaque espèce.

Espèce	Nombre de PLM en 5'	Nombre de PLM en 3'	Nombre de séquences en 5'	Nombre de séquences en 3'
A.lyrata	7598	9986	29792	29726
A.thaliana	6387	6424	19736	20573
B.distachyon	6789	5732	26787	27315
C.maxima	3202	2770	14780	16054
C.melo	3024	3558	18463	19270
F.vesca	3244	1964	19710	20117
H.annuus	5605	3056	32581	35930
H.vulgare	3851	3845	31486	33407
L.albus	5288	4134	32122	31914
M.domestica	4944	4323	28834	30583
M.truncatula	3263	2594	21022	21849
O.sativa	7053	5948	22122	23182
P.persica	2725	2328	21129	21817
P.trichocarpa	5532	4962	30661	30846
P.vulgaris	2778	3125	22367	22281
S.lycopersicum	3244	2711	17683	18408
T.aestivum	6041	8118	37640	48429
V.vinifera	3715	4662	14412	16155

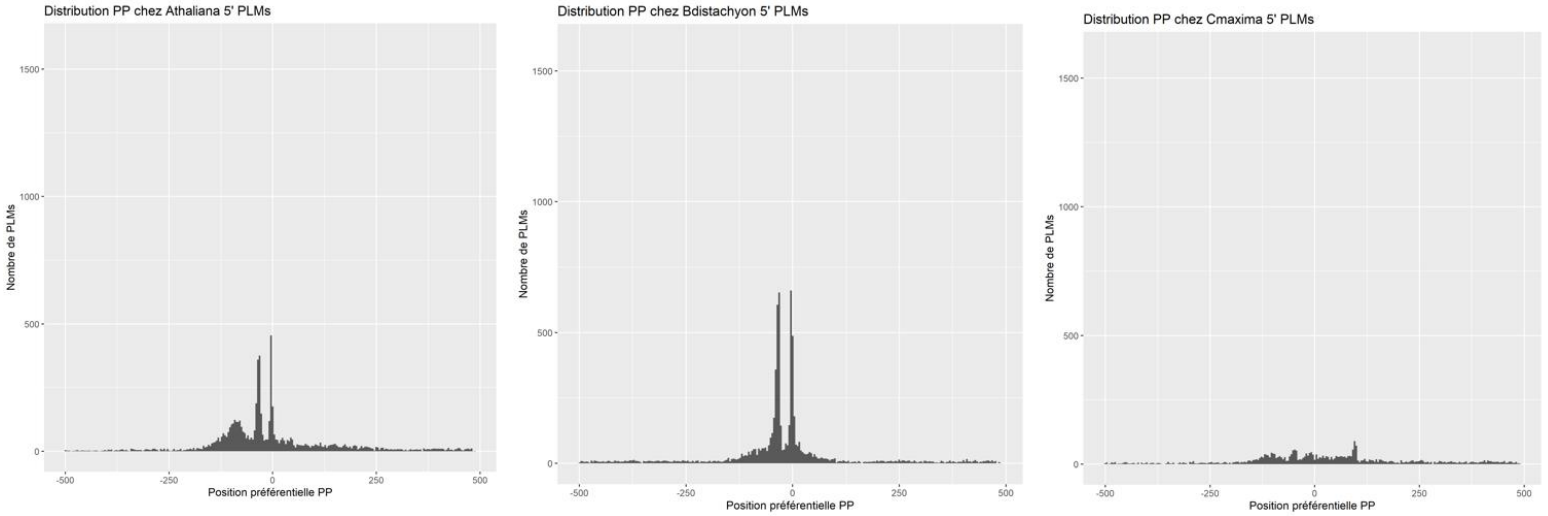
Tableau 4:Nombre de PLM et de séquences proximales analysées en 5' et en 3' pour chaque espèce.

Espèce	Nombre de 4-mer	% 4-mer	Nombre de 5-mer	% 5-mer	Nombre de 6-mer	% 6-mer	Nombre de 7-mer	% 7-mer	Nombre de 8-mer	% 8-mer
A.lyrata	78	1,0	216	2,8	616	8,1	1683	22,2	5005	65,9
A.thaliana	90	1,4	248	3,9	620	9,7	1566	24,5	3863	60,5
B.distachyon	139	2,0	425	6,3	771	11,4	1635	24,1	3819	56,3
C.maxima	26	0,8	73	2,3	221	6,9	706	22,0	2176	68,0
C.melo	32	1,1	78	2,6	146	4,8	598	19,8	2170	71,8
F.vesca	17	0,5	46	1,4	140	4,3	633	19,5	2408	74,2
H.annuus	61	1,1	169	3,0	468	8,3	1118	19,9	3789	67,6
H.vulgare	43	1,1	116	3,0	222	5,8	664	17,2	2806	72,9
L.albus	147	2,8	452	8,5	891	16,8	1267	24,0	2531	47,9
M.domestica	129	2,6	430	8,7	693	14,0	991	20,0	2701	54,6
M.truncatula	15	0,5	40	1,2	132	4,0	773	23,7	2303	70,6
O.sativa	100	1,4	281	4,0	593	8,4	1585	22,5	4494	63,7
P.persica	19	0,7	49	1,8	173	6,3	589	21,6	1895	69,5
P.trichocarpa	94	1,7	331	6,0	542	9,8	1201	21,7	3364	60,8
P.vulgaris	20	0,7	37	1,3	139	5,0	551	19,8	2031	73,1
S.lycopersicum	25	0,8	53	1,6	200	6,2	783	24,1	2183	67,3
T.aestivum	72	1,2	202	3,3	388	6,4	1159	19,2	4220	69,9
V.vinifera	15	0,4	50	1,3	197	5,3	843	22,7	2610	70,3

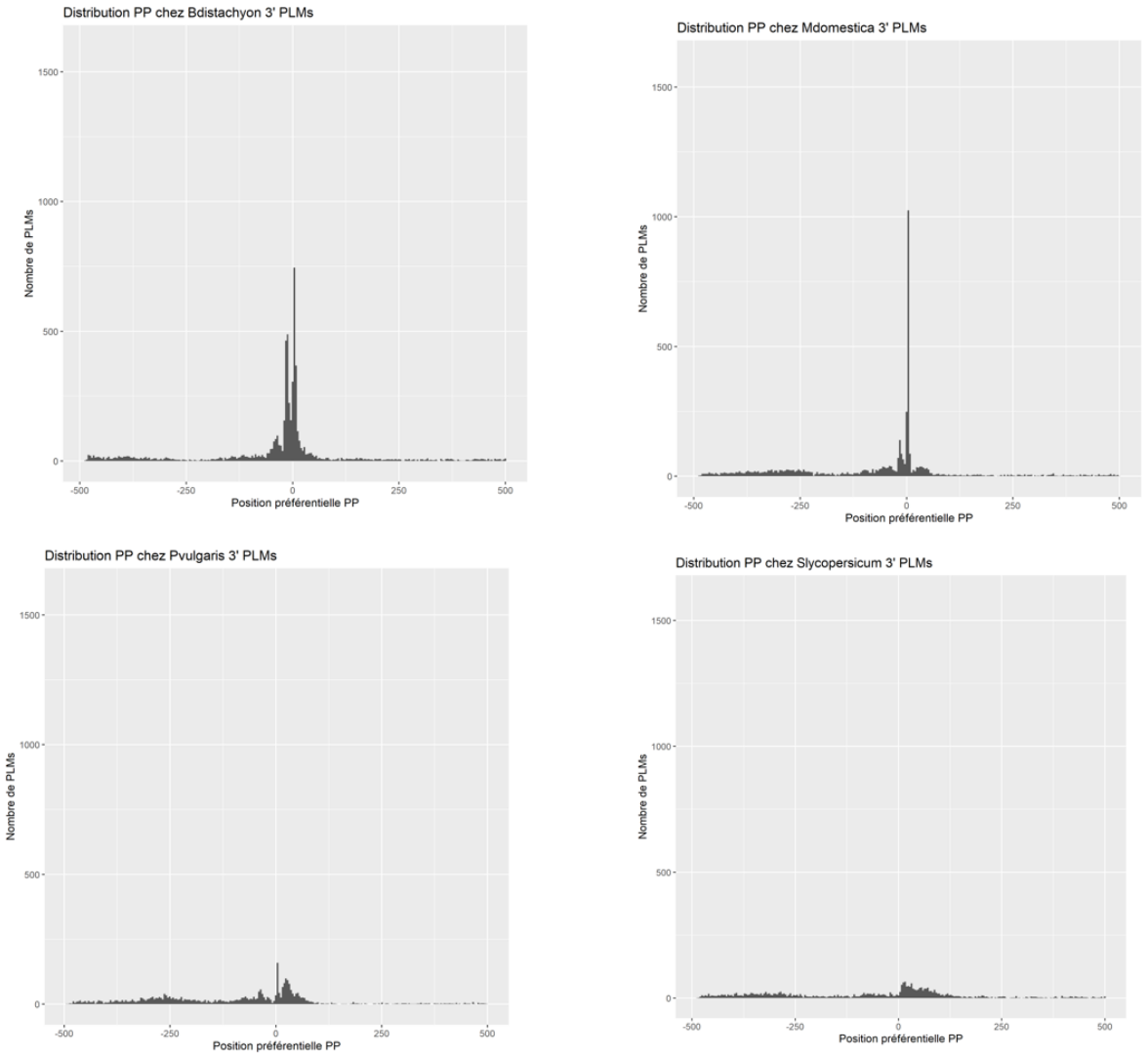
Tableau 5:Nombre et pourcentage de k-mers en 5' pour chaque espèce.

Espèce	Nombre de 4-mer	% 4-mer	Nombre de 5-mer	% 5-mer	Nombre de 6-mer	% 6-mer	Nombre de 7-mer	% 7-mer	Nombre de 8-mer	% 8-mer
A.lyrata	90	0,9	314	3,1	959	9,6	2441	24,4	6182	61,9
A.thaliana	84	1,3	261	4,1	647	10,1	1609	25,0	3823	59,5
B.distachyon	85	1,5	264	4,6	676	11,8	1389	24,2	3318	57,9
C.maxima	68	2,5	139	5,0	253	9,1	592	21,4	1718	62,0
C.melo	47	1,3	105	3,0	239	6,7	751	21,1	2416	67,9
F.vesca	11	0,6	23	1,2	109	5,5	392	20,0	1429	72,8
H.annuus	68	2,2	131	4,3	224	7,3	572	18,7	2061	67,4
H.vulgare	62	1,6	122	3,2	209	5,4	687	17,9	2765	71,9
L.albus	151	3,7	435	10,5	801	19,4	1001	24,2	1746	42,2
M.domestica	129	3,0	336	7,8	615	14,2	911	21,1	2332	53,9
M.truncatula	36	1,4	74	2,9	190	7,3	623	24,0	1671	64,4
O.sativa	121	2,0	347	5,8	759	12,8	1369	23,0	3352	56,4
P.persica	16	0,7	53	2,3	149	6,4	501	21,5	1609	69,1
P.trichocarpa	87	1,8	236	4,8	526	10,6	1142	23,0	2971	59,9
P.vulgaris	54	1,7	142	4,5	250	8,0	670	21,4	2009	64,3
S.lycopersicum	20	0,7	48	1,8	176	6,5	661	24,4	1806	66,6
T.aestivum	91	1,1	246	3,0	485	6,0	1747	21,5	5549	68,4
V.vinifera	61	1,3	167	3,6	519	11,1	1268	27,2	2647	56,8

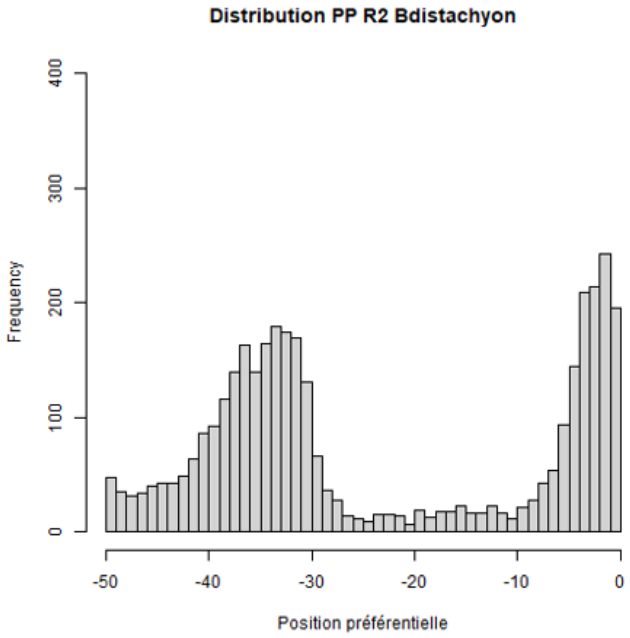
Tableau 6:Nombre et pourcentage de k-mers en 3' pour chaque espèce.

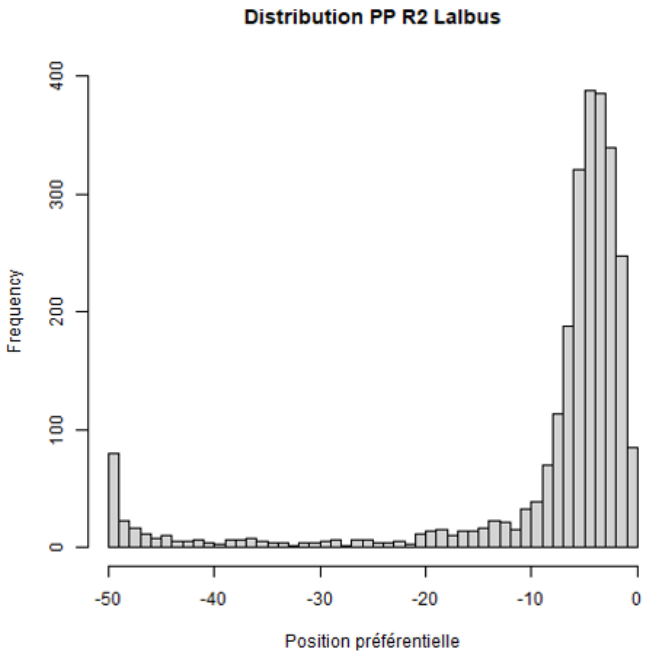


Annexe - Figure 2: Exemples des différents types de distribution des PP obtenus en 5', chez A.lyrata *(à gauche) avec un groupe sur le TSS et 2 en amont,* B.distachyon *(au milieu) avec un groupe sur le TSS et un autre en amont et* C.maxima *(à droite) où aucun groupe ne ressort.*

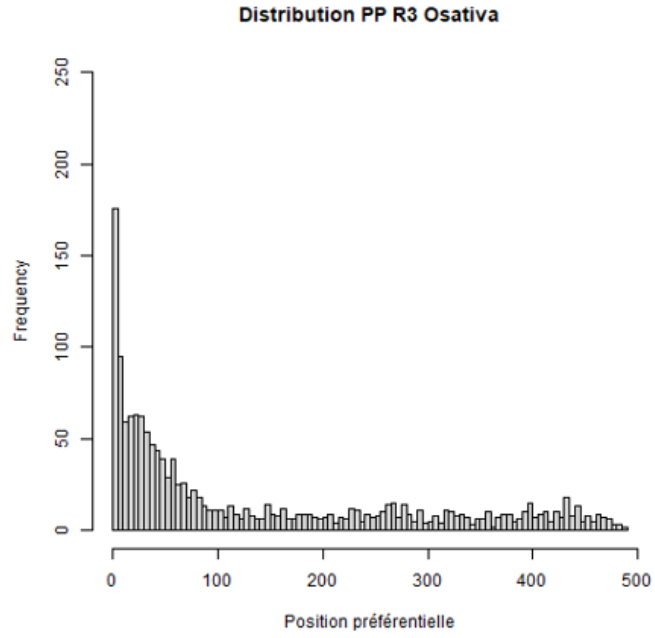


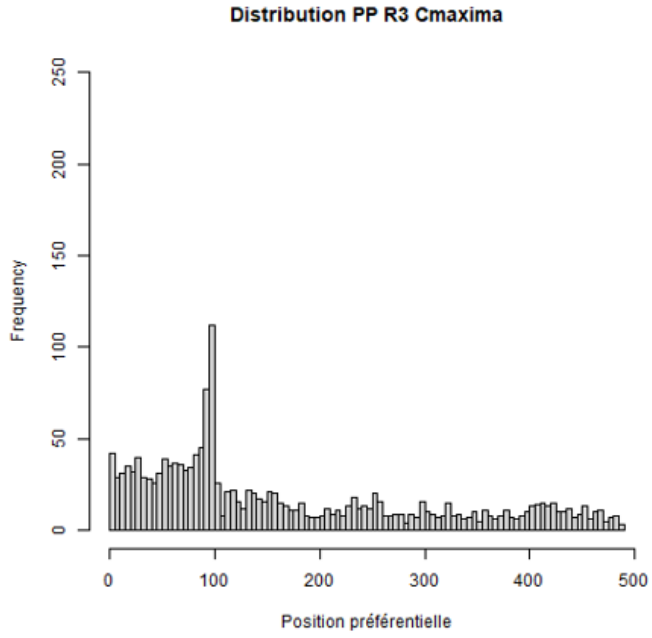
Annexe - Figure 3: Exemples des différents types de distribution des PP obtenus en 3', chez B.distachyon *(en haut à gauche) avec 3 groupes de PLM : un sur le TTS et 2 en amont,* M.domestica *(en haut à droite) avec un 4ème groupe en aval du TTS ,* P.vulgaris *(en bas à gauche) avec un groupe sur le TTS et un autre en aval et* S.lycopersicum *(en bas à droite) où le paysage de PLM est plat bien qu'ils soient plutôt concentrés en aval du TTS.*





*Annexe - Figure 4: Histogrammes représentant les différents types de distribution des PP obtenus sur R2 :*B.distachyon *(à gauche) représentant les espèces pour lesquelles on observe 2 groupes de PLM : un sur [-39pb ; -31pb] et l'autre sur [-5pb ; 0pb] et* L.albus *représentant les espèces chez lesquelles on retrouve un groupe de PLM aux alentours de -50pb et un autre sur [-6pb ; 0pb].*





Annexe - Figure 5: Histogrammes représentant les différents types de distribution des PP obtenus sur R3: O.sativa *représentant les espèces pour lesquelles on observe une forte concentration de PLM en aval du TTS (sur [+1pb ; +50pb) et* C.maxima *seule espèce chez laquelle on observe un groupe de PLM localisé sur [+90pb ; +100pb].*

Espèce	Position TATAAA (en pb)	Position TATATA (en pb)	Variabilité (en pb
A.lyrata	-32	-34	2
A.thaliana	-31	-33	2
B.distachyon	-31	-32	1
C.maxima	-45	-45	0
C.melo	-31	-32	1
F.vesca	-32	-32	0
H.annuus	-32	-33	1
H.vulgare	-43	-44	1
L.albus	-47	-48	1
M.domestica	-49	-45	4
M.truncatula		-32	
O.sativa	-30	-32	2
P.persica	-44	-45	1
P.trichocarpa	-31	-32	1
P.vulgaris	-44	-45	1
S.lycopersicum	-31	-30	1
T.aestivum	-44	-43	1
V.vinifera	-33	-32	1

Tableau 7: Position des TATAWA chez les 18 espèces et leur variabilité.

Espèce	Nombre de PLM appartenant au groupe des boîtes TATA identifié par Yamamoto et al sur R2	Intervalle des PP (en pb) sur R2	Intervalle des PP (en pb) sur R1 et R2 après recherche des PLM sur R1
A.lyrata	34	[-48;-24]	[-48;-24]
A.thaliana	32	[-41;-29]	[-41;-29]
B.distachyon	37	[-38;-28]	[-38;-28]
C.maxima	7	[-48;-43]	[-48;-43]
C.melo	25	[-37;-27]	[-59;-27]
F.vesca	16	[-50;-30]	[-52;-30]
H.annuus	18	[-50;-32]	[-55;-32]
H.vulgare	20	[-50;-35]	[-50;-35]
L.albus	11	[-50;-47]	[-60;-47]
M.domestica	14	[-50;-45]	[-56;-45]
M.truncatula	6	[-49;-29]	[-49;-29]
O.sativa	36	[-38;-27]	[-38;-27]
P.persica	20	[-50;-42]	[-57;-42]
P.trichocarpa	31	[-37;-28]	[-37;-28]
P.vulgaris	17	[-50;-39]	[-50;-39]
S.lycopersicum	19	[-37;-30]	[-37;-30]
T.aestivum	27	[-50;-40]	[-50;-40]
V.vinifera	21	[-50;-29]	[-50;-29]

Tableau 8 :Nombre de PLM appartenant au groupe des boîtes TATA identifié par Yamamoto et al et intervalles sur lesquels ils sont retrouvés sur R2 et sur R1 & R2.

Espèce	Nombre de TATAboxΔ1	Intervalle des PP (en pb)
A.lyrata	18	[-34;-25]
A.thaliana	23	[-35;-29]
B.distachyon	24	[-34;-30]
C.maxima	1	-48
C.melo	5	[-35;-30]
F.vesca	4	[-45;-30]
H.annuus	14	[-50;-37]
H.vulgare	4	[-48;-37]
L.albus	11	[-50;-2]
M.domestica	2	[-48;-6]
M.truncatula	1	[-48]
O.sativa	19	[-35;-28]
P.persica	6	[-50;-46]
P.trichocarpa	20	[-35;-28]
P.vulgaris	6	[-49;-44]
S.lycopersicum	4	[-35;-30]
T.aestivum	8	[-49;-39]
V.vinifera	3	[-37;-31]

Tableau 9:Nombre de variants TATAboxΔ1 identifié dans chaque espèce et régions sur lesquelles ils sont retrouvés.

Espèce	Nombre de motif(s) initateur(s) identifié
A.lyrata	11
A.thaliana	23
B.distachyon	21
C.maxima	0
C.melo	4
F.vesca	0
H.annuus	5
H.vulgare	0
L.albus	79
M.domestica	63
M.truncatula	0
O.sativa	12
P.persica	0
P.trichocarpa	31
P.vulgaris	0
S.lycopersicum	0
T.aestivum	1
V.vinifera	0

Tableau 10 : Nombre de motifs initiateurs identifiés chez chaque espèce.

Classe	Famille	Nombre de FT
MADS box factors	M alpha	1
	MIKC	19
Homeo domain factors	HD-ZIP	22
	PLINC	5
	LBD	1
Other C4 zinc finger-type factors	DOF	29
	C4-GATA-related	10
Tryprophan cluster factors	Myb	34
	GARP_ARR-B	7
	Myb-related	28
	Trihelix	9
	GARP_G2-like	16
Basic leucine zipper factors (bZIP)	Group S	9
	Group G	5
	Group D	10
	Group H	2
	Group A	7
	Group K	1
	Group I	1
	Group B	1
AP2/EREBP	ERF/DREB	79
	AP2	5
Basic helix-loop-helix factors (bHLH)	BES/BZR	5
	TCP	29
	Inconnue	35
GCM domain factors	WRKY-like_FRS/FRF	2
	WRKY	44
	NAC	17
B3	ABI3	3
	RAV	2
	ARF	15
C2H2 zinc finger factors	SBP	14
	Inconnue	14
Fork head/ winged helix factors	E2F	1
A.T hook factors	Inconnue	3
CAMTA	Inconnue	3
CPP	Inconnue	5
BBR/BPC	Inconnue	3
LEAFY	Inconnue	1
heat shock factors	Inconnue	4
Inconnue	Inconnue	1

Tableau 11:Classe, famille et nombre de FT se liant aux TFBS présentant des similarités avec les PLM ni boîtes TATA ni motif initiateur.

Résumé

Afin de répondre au stress de leur environnement, les plantes ont besoin de s'adapter. Cette adaptation passe notamment par la régulation transcriptionnelle, dans laquelle des facteurs de transcription FT viennent se lier à des séquences d'ADN spécifiques situées à proximité relative des gènes pour en activer ou non la transcription. Ces séquences d'ADN appelés motifs cis-régulateurs sont retrouvés à des positions spécifiques du site d'initiation ou de terminaison de la transcription. La méthode PLMdetect permet ainsi d'identifier ces motifs caractérisés par leur position préférentielle ainsi que leur fenêtre fonctionnelle, c'est-à-dire la région dans laquelle ils peuvent être activement liés par des FT. Afin d'étudier la conservation de ces motifs préférentiellement localisés au cours de l'évolution, une analyse comparative a été menée sur les régions proximales 5' et 3' d'une vingtaine d'espèces de plantes, appartenant à 9 familles botaniques différentes. Cette analyse a alors permis de caractériser les régions des boîtes TATA pour chaque angiosperme ainsi que d'identifier des motifs initiateurs. Elle a aussi mis en avant des PLM dans le promoteur central, similaires à des TFBS et d'autres PLM non caractérisés mais qui représentent des candidats intéressants à explorer.

Summary

In order to respond to the stress of their environment, plants need to adapt. This adaptation requires transcriptional regulation, in which transcription factors bind to specific DNA sequences located in relative proximity to the genes to activate or not the transcription. These DNA sequences called cis-regulatory elements are found at specific positions in the site of initiation or termination of transcription. The PLMdetect method makes it possible to identify these patterns characterized by their preferential position as well as their functional window, the region in which they can be actively linked by FT. In order to study the conservation of these preferably localized motifs during evolution, a comparative analysis was conducted on the 5' and 3' proximal regions of about twenty plant species belonging to nine different families. This analysis allowed the characterization of the TATA boxes regions for each angiosperm and the identification of initiating motives. She also highlighted PLM in the central promoter, similar to TFBS and other PLMs not characterized but that represent interesting candidates to explore.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Caractérisation des séquences *cis*-régulatrices dans les régions proximales des gènes chez les plantes

Mots clés : Séquences *cis*-régulatrices ; Transcription ; Motifs Préférentiellement Localisés ; Régions proximales des gènes ; Biologie computationnelle ; Plante

Résumé : La transcription des gènes constitue un processus essentiel dans la réponse adaptative des plantes aux contraintes environnementales qu'elles subissent. Ce processus est finement régulé par de nombreux acteurs moléculaires agissant en *cis* ou en *trans*. Les séquences *cis*-régulatrices correspondent à de courtes portions de l'ADN capables de moduler l'expression de gènes cibles. Elles sont présentes en forte densité dans la région entourant le site d'initiation de la transcription (région 5'-proximale), ainsi que dans celle entourant le site de terminaison de la transcription (région 3'-proximale). Bien que de nombreux travaux expérimentaux et computationnels aient permis de progresser quant à notre connaissance des séquences *cis*-régulatrices présentes dans ces régions proximales, la caractérisation de ces séquences reste encore lacunaire. Dans ce contexte, cette thèse synthétise les travaux réalisés pour mieux comprendre la structure et la fonction des séquences *cis*-régulatrices présentes dans les régions proximales des gènes chez les plantes grâce à la détection de courtes séquences d'ADN préférentiellement localisées (PLM) dans ces régions proximales.

Dans un premier temps, j'ai identifié l'ensemble des PLM *de novo* présents chez *Arabidopsis thaliana* et *Zea mays*, deux espèces végétales dont le génome diffère en termes de contenu et d'architecture. Cette analyse a permis de révéler trois types de PLM dans les régions proximales des deux espèces végétales étudiées : (1) des sites de fixation de facteurs de transcription, (2) des séquences présentant des homologies avec des microARN et (3) des séquences *cis*-régulatrices putatives qui constituent 79% des PLM identifiés et dont une partie est supportée par des données expérimentales d'accessibilité de la chromatine. Ce premier axe de recherche a aussi permis de ré-étayer l'importance de la région 3'-proximale des gènes dans le contrôle de l'expression des gènes et l'intérêt qu'il y a à poursuivre sa caractérisation dans un futur proche.

Dans un deuxième temps, j'ai étendu l'analyse conduite chez *A. thaliana* et *Z. mays* à 18 autres espèces de plantes à fleurs pour déterminer dans quelle mesure les PLM sont conservés. Une base de données, nommée Plant-PLMview, a aussi été développée pour les 20 espèces étudiées dans le but de mettre la méthode de détection des PLM (PLMdetect) à la disposition de toute la communauté scientifique. Cette base de données vise à accélérer la caractérisation des régions proximales des gènes chez les plantes. Pour cela, elle offre la possibilité d'utiliser simplement la méthode PLMdetect et d'interpréter les résultats aisément en visualisant des modules de PLM impliqués dans la co-régulation d'un groupe de gènes d'intérêt.

Enfin, la dernière partie de cette thèse s'est focalisée sur l'implication des PLM dans la réponse globale aux stress chez *A. thaliana*. Pour cela, j'ai utilisé une ressource génomique originale consistant en un réseau de plusieurs milliers de gènes regroupés au sein de plusieurs dizaines de clusters de co-expression identifiés pour constituer la réponse transcriptionnelle commune aux stress. Pour chaque cluster de co-expression, j'ai identifié les PLM *de novo* et enrichis, ce qui a permis d'identifier plus de 200 facteurs de transcription potentiellement impliqués dans la régulation de ce réseau. Ces travaux ont également révélé des séquences *cis*-régulatrices putatives non caractérisées dans les bases de données. Pour les valider, j'ai développé une approche *in silico* pour identifier les expériences transcriptomiques et les gènes sur lesquels entreprendre les expériences humides de validation. À la suite de ce travail, différentes constructions moléculaires ont été produites afin d'initier des validations expérimentales.

Title : Characterization of gene-proximal *cis*-regulatory sequences in plants

Keywords : *Cis*-regulatory sequences ; Transcription ; Preferentially Located Motifs ; Gene-proximal regions ; Computational biology ; Plant

Abstract : Gene transcription is an essential process in the adaptation of plants to environmental conditions. This process is finely regulated by numerous molecular players acting in *cis* or trans. *Cis*-regulatory sequences are short DNA segments that can modulate the expression of target genes. They are found in high density in the region around the transcription initiation site (5'-proximal region) and in the region around the transcription termination site (3'-proximal region). Although numerous experimental and computational studies have increased our knowledge of the *cis*-regulatory sequences in these proximal regions, the characterization of these sequences remains incomplete. In this context, this thesis summarizes the work that has been done to better understand the structure and function of *cis*-regulatory sequences in the proximal regions of plant genes by detecting short preferentially located DNA sequences (PLM) in these proximal regions.

First, I identified the set of *de novo* PLMs in *Arabidopsis thaliana* and *Zea mays*, two species whose genomes differ in content and architecture. This analysis revealed three types of PLMs in the proximal regions of the two species studied : (1) transcription factor-binding sites, (2) sequences with homologies to microRNA, and (3) putative *cis*-regulatory sequences, accounting for 79% of the identified PLMs, part of which is supported by experimental chromatin accessibility data. This initial line of research has also allowed me to recover the importance of the 3'-proximal region of genes in controlling gene expression and the interest to pursue its characterization in the near future.

In a second step, I extended the analysis performed in *A. thaliana* and *Z. mays* to 18 other flowering plant species to determine the extent to which genomic PLMs are conserved. A database called Plant-PLMview was also developed for the 20 species studied to make the PLM detection method (PLMdetect) available to the entire scientific community. This database aims to accelerate the characterization of proximal regions of genes in plants. To this end, it provides the ability to easily apply the PLMdetect method and easily interpret the results by visualizing PLM modules involved in the coregulation of a group of genes of interest.

Finally, the last part of this work addressed the involvement of PLMs in the global stress response in *A. thaliana*. For this purpose, I used an original genomic resource consisting of a network of several thousand genes grouped into several dozen coexpression clusters identified as sharing a common transcriptional response to stress. For each coexpression cluster, I identified *de novo* and enriched PLMs, leading to the identification of over 200 transcription factors potentially involved in the regulation of this network. This work also uncovered putative *cis*-regulatory sequences that are not characterized in the databases. To validate these, I developed an *in silico* approach to identify transcriptomic experiments and genes on which to perform wet validation experiments. As a result of this work, several molecular constructs were produced to initiate experimental validations.